

Contents

I	Calculus	1
1	Limit Theory	2
1.1	Function	2
2	Differential Calculus	3
3	Integral Calculus	4
II	Matrix Theory	5
4	Matrix Norms	6
4.1	Matrix Norms Induced by Vector Norms	6
5	Matrix Decompositions	7
5.1	Spectral Decomposition	7
5.2	Singular Value Decomposition	8
5.2.1	Relationship to Matrix Norm	9
III	Real Analysis	10
6	Measure Theory	11
6.1	Semi-algebras, Algebras and Sigma-algebras	11
6.2	Measure	13
7	Lebesgue Integration	14
7.1	Properties of the Integral	14
7.2	Product Measures	15
IV	Functional Analysis	16
V	Probability Theory	17
8	Random Variables	18
8.1	Probability Space	18
8.2	Random Variables	18

8.3	Distributions	19
8.3.1	Definition of Distributions	19
8.3.2	Properties of Distributions	19
8.3.3	Families of Distributions	20
8.4	Expected Value	20
8.5	Independence	20
8.5.1	Definition of Independence	20
8.5.2	Sufficient Conditions for Independence	21
8.5.3	Independence, Distribution, and Expectation	21
8.5.4	Sums of Independent Random Variables	22
8.6	Moments	22
8.7	Characteristic Functions	22
8.7.1	Definition of Characteristic Functions	22
8.7.2	Properties of Characteristic Functions	23
8.7.3	The Inversion Formula	23
8.7.4	Convergence in Distribution	23
8.7.5	Moments and Derivatives	24
9	Convergence of Random Variables	25
9.1	Convergence in Mean	25
9.2	Convergence in Probability	25
9.3	Convergence in Uninform	25
9.4	Convergence in Distribution	26
9.5	Almost Sure Convergence	29
9.6	Asymptotic Notation for Random Variables	30
10	Law of Large Numbers	31
10.1	Weak Law of Large Numbers	31
10.2	Strong Law of Large Numbers	32
10.2.1	Borel-Cantelli Lemmas	32
10.2.2	Strong Law of Large Numbers	33
10.3	Uniform Law of Large Numbers	33
11	Central Limit Theorems	35
11.1	Classic Central Limit Theorem	35
11.1.1	The De Moivre-Laplace Theorem	35
11.1.2	Classic Central Limit Theorem	37
11.2	Central Limit Theorem for independent non-identical Random Variables	37
11.3	Central Limit Theorem for dependent Random Variables	38
12	The Delta Methods	39
13	Exercises for Probability Theory and Examples	40
13.1	Measure Theory	40
13.2	Laws of Large Numbers	40
13.3	Central Limit Theorems	40

VI Stochastic Process	43
14 Martingales	44
14.1 Conditional Expectation	44
14.2 Martingales	44
14.3 Doob's Inequality	45
14.4 Uniform Integrability	46
14.5 Optional Stopping Theorems	46
15 Markov Chains	47
15.1 Markov Chain	47
15.2 Markov Properties	48
15.3 Recurrence and Transience	49
15.4 Stationary Measures	50
15.5 Asymptotic Behavior	50
15.6 Ergodic Theorems	50
16 Brownian Motion	52
16.1 Markov Properties	53
16.2 Martingales	53
16.3 Sample Paths	54
16.4 Itô Stochastic Calculus	56
17 Exercises for Probability Theory and Examples	58
17.1 Martingales	58
17.2 Markov Chains	58
17.3 Ergodic Theorems	58
17.4 Brownian Motion	58
17.5 Applications to Random Walk	58
17.6 Multidimensional Brownian Motion	58
VII Statistics Inference	59
18 Introduction	60
18.1 Populations and Samples	60
18.2 Statistics	60
18.2.1 Sufficient Statistics	60
18.2.2 Complete Statistics	60
18.3 Estimators	61
19 Maximum Likelihood Estimator	63
19.1 Consistency of MLE	63
19.2 Asymptotic Normality of MLE	65
19.3 Efficiency of MLE	65
20 Minimum-Variance Unbiased Estimator	66

21 Bayes Estimator	69
21.1 Single-Prior Bayes	70
21.2 Hierarchical Bayes	71
21.3 Empirical Bayes	72
21.4 Bayes Prediction	72
22 Hypothesis Testing	73
 VIII Nonparametric Statistics	 74
23 Statistical Functionals	75
23.1 Empirical Distribution Function	75
24 Kernel Methods	76
24.1 Positive Definite Kernels	76
24.1.1 Construction of the Reproducing Kernel Hilbert Space	76
24.1.2 Examples of Kernels	78
 IX Convex Optimization	 80
25 Convex Sets	81
25.1 Affine and Convex Sets	81
25.1.1 Affine Sets	81
25.1.2 Convex Sets	81
25.1.3 Cones	81
25.2 Some Important Examples	82
25.3 Generalized Inequalities	82
25.3.1 Definition of Generalized Inequalities	82
25.3.2 Properties of Generalized Inequalities	83
26 Convex Optimization Problems	84
26.1 Generalized Inequality Constraints	84
26.1.1 Conic Form Problems	84
26.1.2 Semidefinite Programming	84
26.2 Vector Optimization	84
27 Unconstrained Minimization	85
27.1 Definition of Unconstrained Minimization	85
27.2 General Descent Method	87
27.3 Gradient Descent Method	87
27.4 Steepest Descent Method	87
27.5 Newton's Method	87
28 Exercises for Convex Optimization	88
28.1 Convex Sets	88

X	Regression Analysis	89
29	Modified Likelihood	90
29.1	Marginal Likelihood	90
29.2	Conditional Likelihood	90
29.3	Profile Likelihood	91
29.4	Quasi Likelihood	91
30	Generalized Linear Model	92
30.1	Introduction	92
30.2	Binary Data	95
30.3	Polytomous Data	96
30.4	Count Data	97
31	Survival Analysis	98
31.1	General Formulation	98
31.2	Estimation of Survival Function	99
31.3	Proportional Hazards Model	100
XI	Machine Learning	101
32	Support Vector Machine	102
33	Linear Discriminant Analysis	104
34	K-Nearest Neighbor	105
35	Decision Tree	106
XII	Random Matrix Theory	107
36	Sample Covariance Matrices	108
36.1	Eigenvalues and Singular Values	108
36.2	Laguerre Orthogonal Ensemble	109
36.3	Marčenko-Pastur Theorem	113
36.3.1	Preliminary	113
36.3.2	Marčenko-Pastur Theorem	117
36.4	Limits of Extreme Eigenvalues	120
XIII	Applications	121
37	Missingness Data	122
37.1	The Problem of Missing Data	122
37.1.1	Missingness Mechanisms	123
37.1.2	Commonly Used Methods for Missing Data	123
37.2	Likelihood-Based Inference with Missing Data	124
37.2.1	Ignorable Missingness Mechanism	124
37.2.2	Expectation-Maximization Algorithm	126

37.3 Missing Not At Random Models	131
37.3.1 Normal Models for MNAR Missing Data	131
38 Treatment-effects Analysis	132
38.1 Evaluations	132
38.1.1 Average Treatment Effect	132
38.1.2 Mann-Whitney Statistic	132
38.1.3 Distribution-type Index	132

Part I

Calculus

Chapter 1

Limit Theory

1.1 Function

Definition 1.1.1 (Mapping)

Let $X : \Omega_1 \rightarrow \Omega_2$ 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

Definition 1.1.2 (Closed Function)

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be closed, if for each $\alpha \in \mathbb{R}$, the sublevel set

$$\{\mathbf{x} \in \text{dom } f \mid f(\mathbf{x}) \leq \alpha\} \tag{1.1}$$

is closed.

Chapter 2

Differential Calculus

Chapter 3

Integral Calculus

Part II

Matrix Theory

Chapter 4

Matrix Norms

4.1 Matrix Norms Induced by Vector Norms

Chapter 5

Matrix Decompositions

5.1 Spectral Decomposition

Definition 5.1.1 (Eigenvectors and Eigenvalues)

A (non-zero) vector \mathbf{v} of dimension n is an **eigenvector** of a square $n \times n$ matrix \mathbf{A} , if

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \quad (5.1)$$

where λ is a scalar, termed the **eigenvalue** corresponding to \mathbf{v} .

Definition 5.1.2 (Spectral Decomposition)

For any $n \times n$ matrix with n linearly independent eigenvectors $\mathbf{q}_i, i = 1, \dots, n$. Then \mathbf{A} can be factorized as

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$$

where \mathbf{Q} is the square $n \times n$ matrix whose i -th column is the eigenvector \mathbf{q}_i of \mathbf{A} , and $\mathbf{\Lambda}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, $\mathbf{\Lambda} = \lambda_i$. This factorization is called eigendecomposition or sometimes spectral decomposition.

Example (Real Symmetric Matrices). As a special case, for every $n \times n$ real symmetric matrix, the eigenvalues are real and the eigenvectors can be chosen real and orthonormal. Thus a real symmetric matrix \mathbf{A} can be decomposed as

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}' \quad (5.2)$$

where \mathbf{Q} is an orthogonal matrix whose columns are eigenvectors of \mathbf{A} , and $\mathbf{\Lambda}$ is a diagonal matrix whose entries are the eigenvalues of \mathbf{A} .

5.2 Singular Value Decomposition

Definition 5.2.1 (Singular Value Decomposition)

For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we have

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' \quad (5.3)$$

where

- $\mathbf{U} \in \mathbb{R}^{m \times m}$ is an orthogonal matrix whose columns are the eigenvectors of $\mathbf{A}\mathbf{A}'$
- $\mathbf{V} \in \mathbb{R}^{n \times n}$ is an orthogonal matrix whose columns are the eigenvectors of $\mathbf{A}'\mathbf{A}$
- $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is an all zero matrix except for the first r diagonal elements

$$\sigma_i = \Sigma_{ii}, \quad i = 1, 2, \dots, r$$

which is called singular values, that are the square roots of the eigenvalues of $\mathbf{A}'\mathbf{A}$ and of $\mathbf{A}\mathbf{A}'$ (these two matrices have the same eigenvalues)

Remark. We assume above that the singular values are sorted in descending order and the eigenvectors are sorted according to descending order of their eigenvalues.

Proof. Without loss of generality, we assume $m \geq n$. Since for the case $n > m$, can then be established by transposing the SVD of \mathbf{A}' ,

$$\mathbf{A} = (\mathbf{A}')' = (\mathbf{U}'\mathbf{\Sigma}\mathbf{V})' = \mathbf{V}'(\mathbf{U}'\mathbf{\Sigma})' = \mathbf{V}'\mathbf{\Sigma}\mathbf{U}$$

For $m \geq n$, suppose $\text{rank}(\mathbf{A}) = r$, and then $\text{rank}(\mathbf{A}'\mathbf{A}) = r$ and the spectral decomposition of $\mathbf{A}'\mathbf{A}$ be

$$\mathbf{A}'\mathbf{A}\mathbf{V} = \mathbf{V} \text{diag}(\sigma_1^2, \dots, \sigma_r^2, 0, \dots, 0)$$

where σ_i^2 are the eigenvalues of $\mathbf{A}'\mathbf{A}$ and the columns of \mathbf{V} , denoted $\mathbf{v}^{(i)}$, are the corresponding orthonormal eigenvectors.

Let

$$\mathbf{u}^{(i)} = \frac{\mathbf{A}\mathbf{v}^{(i)}}{\sigma_i}$$

then

$$\begin{aligned} \mathbf{A}'\mathbf{u}^{(i)} &= \frac{\mathbf{A}'\mathbf{A}\mathbf{v}^{(i)}}{\sigma_i} = \sigma_i \mathbf{v}^{(i)} \Rightarrow \\ \mathbf{A}\mathbf{A}'\mathbf{u}^{(i)} &= \sigma_i \mathbf{A}\mathbf{v}^{(i)} = \sigma_i^2 \mathbf{u}^{(i)} \end{aligned}$$

implying that $\mathbf{u}^{(i)}$ are eigenvectors of $\mathbf{A}\mathbf{A}'$ corresponding to eigenvalues σ_i^2 .

Since the eigenvectors $\mathbf{v}^{(i)}$ are orthonormal, then so are the eigenvectors $\mathbf{u}^{(i)}$

$$\left(\mathbf{u}^{(i)}\right)' \mathbf{u}^{(j)} = \frac{\left(\mathbf{v}^{(i)}\right)' \mathbf{A}'\mathbf{A}\mathbf{v}^{(j)}}{\sigma_i^2} = \left(\mathbf{v}^{(i)}\right)' \mathbf{v}^{(j)} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

We have thus far a matrix \mathbf{V} whose columns are eigenvectors of $\mathbf{A}'\mathbf{A}$ with eigenvalues σ_i^2 , and a matrix \mathbf{U} whose columns are r eigenvectors of $\mathbf{A}\mathbf{A}'$ corresponding to eigenvalues σ_i^2 .

We augment the eigenvectors $\mathbf{u}^{(i)}, i = 1, \dots, r$ with orthonormal vectors $\mathbf{u}^{(i)}, i = r+1, \dots, m$ that span $\text{null}(\mathbf{A}\mathbf{A}')$, and together $\mathbf{u}^{(i)}, i = 1, \dots, m$ are a full orthonormal set of eigenvectors of $\mathbf{A}\mathbf{A}'$ with eigenvalues σ_i^2 (with $\sigma_i = 0$ for $i > r$).

Since

$$[\mathbf{U}'\mathbf{A}\mathbf{V}]_{ij} = (\mathbf{u}^{(i)})' \mathbf{A} \mathbf{v}^{(j)} = \begin{cases} \sigma_j (\mathbf{u}^{(i)})' \mathbf{u}^{(j)} & i \leq r \\ 0 & i > r \end{cases}$$

we get

$$\mathbf{U}'\mathbf{A}\mathbf{V} = \mathbf{\Sigma}$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \text{diag}(\sigma_1, \dots, \sigma_r) & & \\ & \mathbf{0} & \\ & & \end{pmatrix}, \quad \sigma_i = 0 \text{ for } r < i \leq n$$

Consequently, we get the desired decompositions

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$$

□

5.2.1 Relationship to Matrix Norm

Theorem 5.2.1

For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$,

$$\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A}) \quad (5.4)$$

Proof. For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the SVD implies that,

$$\|\mathbf{A}\|_2 = \sup_{\mathbf{x} \neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\mathbf{x} \neq 0} \frac{\|\mathbf{U}\mathbf{\Sigma}\mathbf{V}'\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$$

Since \mathbf{U} is unitary, that is,

$$\|\mathbf{U}\mathbf{x}\|_2^2 = \mathbf{x}'\mathbf{U}'\mathbf{U}\mathbf{x} = \|\mathbf{x}\|_2^2, \quad \forall \mathbf{x} \in \mathbb{R}^m$$

thus,

$$= \sup_{\mathbf{x} \neq 0} \frac{\|\mathbf{\Sigma}\mathbf{V}'\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$$

Let $\mathbf{y} = \mathbf{V}'\mathbf{x}$, and since \mathbf{V} is unitary, we have

$$\|\mathbf{y}\|_2 = \|\mathbf{V}'\mathbf{x}\|_2 = \|\mathbf{x}\|_2 = 1$$

thus,

$$= \sup_{\mathbf{y} \neq 0} \frac{\|\mathbf{\Sigma}\mathbf{y}\|_2}{\|\mathbf{V}\mathbf{y}\|_2} = \sup_{\mathbf{y} \neq 0} \frac{\left(\sum_{i=1}^r \sigma_i^2 |y_i|^2\right)^{\frac{1}{2}}}{\left(\sum_{i=1}^r |y_i|^2\right)^{\frac{1}{2}}} \leq \sigma_{\max}(\mathbf{A})$$

which takes "=", if $\mathbf{y} = (1, 0, \dots, 0)'$.

□

Theorem 5.2.2

For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, suppose $\text{rank}(\mathbf{A}) = n$, then

$$\min_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \sigma_n(\mathbf{A}) \quad (5.5)$$

Proof. The proof process is analogous to the above theorem.

□

Remark. If $\text{rank}(\mathbf{A}) < n$, then there is an \mathbf{x} such that the minimum is zero.

Part III

Real Analysis

Chapter 6

Measure Theory

6.1 Semi-algebras, Algebras and Sigma-algebras

Definition 6.1.1 (Semi-algebra)

A nonempty class \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.,

$$A^C = \sum_{i=1}^n A_i, \text{ where } A_i \in \mathcal{S}, A_i \cap A_j = \emptyset, i \neq j.$$

Definition 6.1.2 (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 6.1.3 (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 (Special σ -algebra). 1. **Trivial 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 (Lebesgue) 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 to some σ -algebra (mostly consider the smallest one containing \mathcal{S}).

Lemma 6.1.1

If \mathcal{S} is a semi-algebra, then

$$\overline{\mathcal{S}} = \{\text{finite disjoint unions of sets in } \mathcal{S}\}$$

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 $A_i, B_j \in \mathcal{S}$.

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

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

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 6.1.1

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 (Borel σ -algebras generated from semi-algebras). 1.

6.2 Measure

Definition 6.2.1 (Measure)

Measure is a nonnegative countably additive set function, that is, a function $\mu : \mathcal{A} \rightarrow \mathbf{R}$ 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 6.2.2 (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**.

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

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

1. **monotonicity** if $A \subset B$, then $\mu(A) \leq \mu(B)$.
2. **subadditivity** if $A \subset \cup_{m=1}^{\infty} A_m$, then $\mu(A) \leq \sum_{m=1}^{\infty} \mu(A_m)$.
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 7

Lebesgue Integration

7.1 Properties of the Integral

Theorem 7.1.1 (Jensen's Inequality)

Let $(\Omega, \mathcal{A}, \mu)$ 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:

$$\varphi\left(\int_{\Omega} f d\mu\right) \leq \int_{\Omega} \varphi(f) d\mu. \quad (7.1)$$

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

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

Then, we got

$$\int_{\Omega} \varphi(f) d\mu \geq \int_{\Omega} af + b d\mu = a \int_{\Omega} f d\mu + b = ax_0 + b = \varphi\left(\int_{\Omega} f d\mu\right).$$

□

Theorem 7.1.2 (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 ,

$$\int_{\Omega} |f \cdot g| d\mu \leq \|f\|_p \|g\|_q. \quad (7.2)$$

Proof.

□

Theorem 7.1.3 (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 ,

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p. \quad (7.3)$$

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

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

Therefore,

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

By Hölder's Inequality (7.1.2),

$$\begin{aligned} \|f + g\|_p^p &= \int |f + g|^p d\mu \\ &= \int |f + g| \cdot |f + g|^{p-1} d\mu \\ &\leq \int (|f| + |g|) |f + g|^{p-1} d\mu \\ &= \int |f| |f + g|^{p-1} d\mu + \int |g| |f + g|^{p-1} d\mu \\ &\leq \left(\left(\int |f|^p d\mu \right)^{\frac{1}{p}} + \left(\int |g|^p d\mu \right)^{\frac{1}{p}} \right) \left(\int |f + g|^{(p-1)(\frac{p}{p-1})} d\mu \right)^{1-\frac{1}{p}} \\ &= (\|f\|_p + \|g\|_p) \frac{\|f + g\|_p^p}{\|f + g\|_p} \end{aligned}$$

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

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

When $p = \infty$,

a

□

Theorem 7.1.4 (Bounded Convergence Theorem)

Theorem 7.1.5 (Fatou's Lemma)

Theorem 7.1.6 (Monotone Convergence Theorem)

7.2 Product Measures

Theorem 7.2.1 (Fubini's Theorem)

Part IV

Functional Analysis

Part V

Probability Theory

Chapter 8

Random Variables

8.1 Probability Space

Definition 8.1.1 (Probability Space)

A probability space is a triple (Ω, \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 $P : \mathcal{F} \rightarrow [0, 1]$: a function on \mathcal{F} which is a measure function.

8.2 Random Variables

Definition 8.2.1 (Random Variable)

A random variable is a measurable function $X : \Omega \rightarrow S$ from a set of possible outcomes (Ω, \mathcal{F}) to a measurable space (S, \mathcal{S}) , that is,

$$X^{-1}(B) \equiv \{\omega : X(\omega) \in B\} \in \mathcal{F} \quad \forall B \in \mathcal{S}. \quad (8.1)$$

Typically, $(S, \mathcal{S}) = (R^d, \mathcal{R}^d)$ ($d > 1$).

How to prove that functions are measurable?

Theorem 8.2.1

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.

8.3 Distributions

8.3.1 Definition of Distributions

Definition 8.3.1 (Distribution)

A distribution of random variable X is a probability function $P : \mathcal{R} \rightarrow \mathbb{R}$ by setting

$$\mu(A) = P(X \in A) = P(X^{-1}(A)), \quad \text{for } A \in \mathcal{R}. \quad (8.2)$$

Definition 8.3.2 (Distribution Function)

The distribution of a random variable X is usually described by giving its **distribution function**,

$$F(x) = P(X \leq x). \quad (8.3)$$

Definition 8.3.3 (Density Function)

If the distribution function $F(x) = P(X \leq x)$ has the form

$$F(x) = \int_{-\infty}^x f(y) dy,$$

that X has density function f .

8.3.2 Properties of Distributions

Theorem 8.3.1 (Properties of Distribution Function)

Any distribution function F has the following properties,

1. F is nondecreasing.
2. $\lim_{x \rightarrow \infty} F(x) = 1, \lim_{x \rightarrow -\infty} F(x) = 0$.
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 8.3.2

If F satisfies (1), (2), and (3) in Theorem 8.3.1, then it is the distribution function of some random variable.

Proof.

□

Theorem 8.3.3

A distribution function has at most countably many discontinuities

Proof.

□

8.3.3 Families of Distributions

8.4 Expected Value

Definition 8.4.1 (Expectation)

Theorem 8.4.1 (Bounded Convergence theorem)

Theorem 8.4.2 (Fatou's Lemma)

If $X_n \geq 0$, then

$$\liminf_{n \rightarrow \infty} EX_n \geq E\left(\liminf_{n \rightarrow \infty} X_n\right). \quad (8.4)$$

Theorem 8.4.3 (Monotone Convergence theorem)

If $0 \leq X_n \uparrow X$, then

$$EX_n \uparrow EX. \quad (8.5)$$

Theorem 8.4.4 (Dominated Convergence theorem)

If $X_n \rightarrow X$ a.s., $|X_n| \leq Y$ for all n , and $EY < \infty$, then

$$EX_n \rightarrow EX. \quad (8.6)$$

8.5 Independence

8.5.1 Definition of Independence

Definition 8.5.1 (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 $C, D \in \mathcal{R}$

$$P(X \in C, Y \in D) = P(X \in C)P(Y \in D). \quad (8.7)$$

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 8.5.1

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 8.5.2

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 8.5.2

An infinite collection of objects (σ -fields, random variables, or sets) is said to be independent if every finite subcollection is,

1. σ -fields $\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_n$ are independent if whenever $A_i \in \mathcal{F}_i$ for $i = 1, \dots, n$, we have

$$P(\cap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i). \quad (8.8)$$

2. Random variables X_1, \dots, X_n are independent if whenever $B_i \in \mathcal{R}$ for $i = 1, \dots, n$ we have

$$P(\cap_{i=1}^n \{X_i \in B_i\}) = \prod_{i=1}^n P(X_i \in B_i). \quad (8.9)$$

3. Sets A_1, \dots, A_n are independent if whenever $I \subset \{1, \dots, n\}$ we have

$$P(\cap_{i \in I} A_i) = \prod_{i \in I} P(A_i). \quad (8.10)$$

8.5.2 Sufficient Conditions for Independence**8.5.3 Independence, Distribution, and Expectation****Theorem 8.5.3**

Suppose X_1, \dots, X_n are independent random variables and X_i has distribution μ_i , then (X_1, \dots, X_n) has distribution $\mu_1 \times \dots \times \mu_n$.

Theorem 8.5.4

If X_1, \dots, X_n are independent and have

1. $X_i \geq 0$ for all i , or
2. $E|X_i| < \infty$ for all i .

then

$$E\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n EX_i \quad (8.11)$$

8.5.4 Sums of Independent Random Variables

Theorem 8.5.5 (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

$$P(X + Y \leq z) = \int F(z - y) dG(y). \quad (8.12)$$

2. If X and Y are independent, X with density f and Y with distribution function G , then $X + Y$ has density

$$h(x) = \int f(x - y) dG(y). \quad (8.13)$$

Suppose Y has density g , the last formula can be written as

$$h(x) = \int f(x - y) g(y) dy. \quad (8.14)$$

3. If X and Y are independent, integral-valued random variables, then

$$P(X + Y = n) = \sum_m P(X = m) P(Y = n - m). \quad (8.15)$$

8.6 Moments

Lemma 8.6.1

If $Y > 0$ and $p > 0$, then

$$E(Y^p) = \int_0^\infty p y^{p-1} P(Y > y) dy. \quad (8.16)$$

8.7 Characteristic Functions

8.7.1 Definition of Characteristic Functions

Definition 8.7.1 (Characteristic Function)

If X is a random variable, we define its characteristic function (ch.f) by

$$\varphi(t) = E(e^{itX}) = E(\cos tX) + iE(\sin tX). \quad (8.17)$$

Remark. Euler Equation.

8.7.2 Properties of Characteristic Functions

Theorem 8.7.1 (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.

□

8.7.3 The Inversion Formula

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

Theorem 8.7.2 (The Inversion Formula)

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

$$\lim_{T \rightarrow \infty} (2\pi)^{-1} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt = \mu(a, b) + \frac{1}{2}\mu(\{a, b\}) \quad (8.18)$$

Proof.

□

Theorem 8.7.3

If $\int |\varphi(t)| dt < \infty$, then μ has bounded continuous density

$$f(y) = \frac{1}{2\pi} \int e^{-ity} \varphi(t) dt. \quad (8.19)$$

Proof.

□

8.7.4 Convergence in Distribution

Theorem 8.7.4 (Lèvy's Continuity Theorem)

Let $\mu_n, 1 \leq n \leq \infty$ be probability measures with ch.f. φ_n .

1. If $\mu_n \xrightarrow{d} \mu_\infty$, then $\varphi_n(t) \rightarrow \varphi_\infty(t)$ 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.

□

8.7.5 Moments and Derivatives

Theorem 8.7.5

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

$$\varphi^{(n)}(t) = \int (ix)^n e^{itx} \mu(dx). \quad (8.20)$$

Theorem 8.7.6

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

$$\varphi(t) = 1 + itEX - t^2 E(X^2)/2 + o(t^2). \quad (8.21)$$

Theorem 8.7.7

If $\limsup_{h \downarrow 0} \{\varphi(h) - 2\varphi(0) + \varphi(-h)\}/h^2 > -\infty$, then

$$E|X|^2 < \infty. \quad (8.22)$$

Chapter 9

Convergence of Random Variables

9.1 Convergence in Mean

Definition 9.1.1 (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. $\lim_{n \rightarrow \infty} E(|X_n - X|^r) = 0$.

Convergence in the r-th mean is denoted by

$$X_n \xrightarrow{L^r} X. \quad (9.1)$$

9.2 Convergence in Probability

Definition 9.2.1 (Convergence in Probability)

A sequence $\{X_n\}$ of real-valued random variables **converges in probability** towards the random variable X , if

$$\forall \varepsilon > 0, \quad \lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) = 0. \quad (9.2)$$

Convergence in probability is denoted by

$$X_n \xrightarrow{P} X. \quad (9.3)$$

9.3 Convergence in Uninform

Definition 9.3.1 (Convergence in Uninform)

9.4 Convergence in Distribution

Definition 9.4.1 (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

$$\lim_{n \rightarrow \infty} F_n(x) = F(x), \quad (9.4)$$

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

$$X_n \xrightarrow{d} X, \text{ or } X_n \Rightarrow X. \quad (9.5)$$

Remark.

- 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 9.4.1

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

$$Y_n \xrightarrow{a.s.} Y_\infty. \quad (9.6)$$

Theorem 9.4.1 (Portmanteau Lemma)

$\{X_n\}$ converges in distribution to X , if and only if any of the following statements are true,

- $P(X_n \leq x) \rightarrow P(X \leq x)$, for all continuity points of the distribution of X .
- $Ef(X_n) \rightarrow Ef(X)$, for all bounded, continuous (Lipschitz) functions f .
- $\liminf_{n \rightarrow \infty} P(X_n \in G) \geq P(X_\infty \in G)$, for all open sets G .
- $\limsup_{n \rightarrow \infty} P(X_n \in K) \leq P(X_\infty \in K)$, for all closed sets K .
- $\lim_{n \rightarrow \infty} P(X_n \in A) = P(X_\infty \in A)$, for all Borel sets A with $P(X_\infty \in \partial A) = 0$.

Proof.

□

Theorem 9.4.2 (Continuous Mapping Theorem)

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

$$\begin{aligned} X_n &\xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X), \\ X_n &\xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X), \\ X_n &\xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X). \end{aligned} \quad (9.7)$$

If in addition g is bounded, then

$$Eg(X_n) \rightarrow Eg(X). \quad (9.8)$$

Proof.

□

Theorem 9.4.3

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

$$X_n \xrightarrow{d} X, \quad (9.9)$$

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

$$X_n \xrightarrow{p} c. \quad (9.10)$$

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

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

similarly, 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,

$$\begin{aligned} P(X \leq x + \varepsilon) &\leq P(X_n \leq x) + P(|X_n - X| > \varepsilon), \\ P(X_n \leq x) &\leq P(X \leq x - \varepsilon) + P(|X_n - X| > \varepsilon). \end{aligned}$$

So, we got

$$\begin{aligned} P(X \leq x + \varepsilon) - P(|X_n - X| > \varepsilon) &\leq P(X_n \leq x) \\ &\leq P(X \leq x - \varepsilon) + P(|X_n - X| > \varepsilon) \end{aligned}$$

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

$$\begin{aligned} P(X \leq x - \varepsilon) &\leq \lim_{n \rightarrow \infty} P(X_n \leq x) \leq P(X \leq x + \varepsilon) \\ &\Rightarrow F(x - \varepsilon) \leq \lim_{n \rightarrow \infty} F_n(x) \leq F(x + \varepsilon). \end{aligned}$$

By the property of distribution (Theorem 8.3.1), as $\varepsilon \rightarrow 0$, then

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

which means,

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

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

$$\begin{aligned} \lim_{n \rightarrow \infty} P(X_n \leq c + \varepsilon) &= 1 \Rightarrow \lim_{n \rightarrow \infty} P(X_n > c + \varepsilon) = 0 \\ \lim_{n \rightarrow \infty} P(X_n \leq c - \varepsilon) &= 0. \end{aligned}$$

Therefore,

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

which means

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

□

Theorem 9.4.4 (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.

□

Remark. 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 9.4.5 (Cramér-Wold Theorem)

Theorem 9.4.6 (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 $\lim_{k \rightarrow \infty} F_{n(k)}(y) = F(y)$ at all continuity points y of F .

Theorem 9.4.7

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

$$\limsup_{n \rightarrow \infty} 1 - F_n(M_\epsilon) + F_n(-M_\epsilon) \leq \epsilon. \quad (9.11)$$

9.5 Almost Sure Convergence

Definition 9.5.1 (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

$$P\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1. \quad (9.12)$$

Almost sure convergence is denoted by

$$X_n \xrightarrow{a.s.} X. \quad (9.13)$$

Remark.

Theorem 9.5.1

If $X_n \xrightarrow{a.s.} X$, then

$$X_n \xrightarrow{p} X. \quad (9.14)$$

Proof.

□

Theorem 9.5.2

$X_n \xrightarrow{p} X$ if and only if for all subsequence $X_{n(m)}$ exists a further subsequence $X_{n(m_k)}$, such that

$$X_{n(m_k)} \xrightarrow{a.s.} X. \quad (9.15)$$

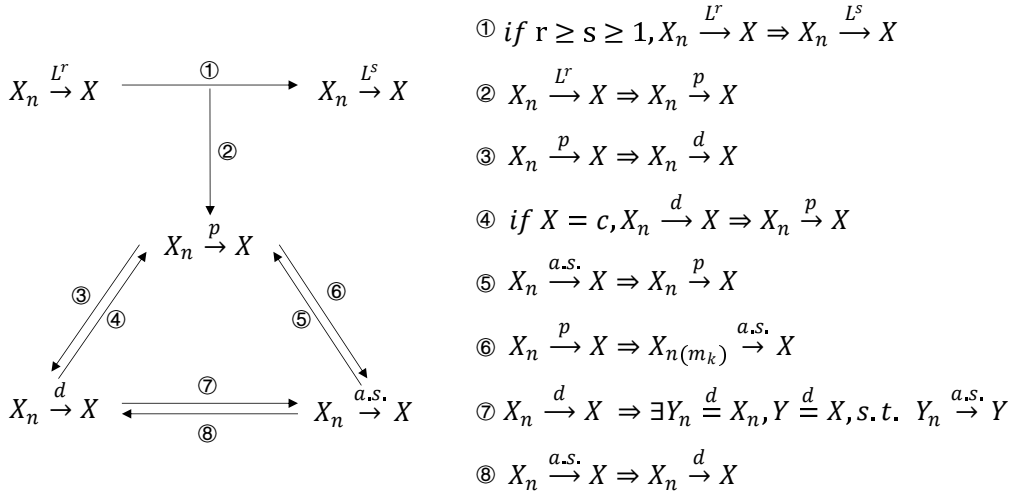


Figure 9.1: Relations of Convergence of Random Variables

9.6 Asymptotic Notation for Random Variables

Definition 9.6.1

A sequence $\{A_n\}$ of real-valued random variables is of smaller order in probability than a sequence $\{B_n\}$, if

$$\frac{A_n}{B_n} \xrightarrow{p} 0. \quad (9.16)$$

Smaller order in probability is denoted by

$$A_n = o_p(B_n). \quad (9.17)$$

Particularly,

$$A_n = o_p(1) \iff A_n \xrightarrow{p} 0. \quad (9.18)$$

Definition 9.6.2

A sequence $\{A_n\}$ of real-valued random variables is of smaller order than or equal to a sequence $\{B_n\}$ in probability, if

$$\forall \varepsilon > 0 \exists M_\varepsilon, \quad \lim_{n \rightarrow \infty} P(|A_n| \leq M_\varepsilon |B_n|) \geq 1 - \varepsilon. \quad (9.19)$$

Smaller order than or equal to in probability is denoted by

$$A_n = O_p(B_n). \quad (9.20)$$

Definition 9.6.3

A sequence $\{A_n\}$ of real-valued random variables is of the same order as a sequence $\{B_n\}$ in probability, if

$$\forall \varepsilon > 0 \exists m_\varepsilon < M_\varepsilon, \quad \lim_{n \rightarrow \infty} P\left(m_\varepsilon < \frac{|A_n|}{|B_n|} < M_\varepsilon\right) \geq 1 - \varepsilon. \quad (9.21)$$

Same order in probability is denoted by

$$A_n \asymp_p B_n. \quad (9.22)$$

Chapter 10

Law of Large Numbers

10.1 Weak Law of Large Numbers

Lemma 10.1.1

If $p > 0$ and $E|Z_n|^p \rightarrow 0$, then

$$Z_n \xrightarrow{p} 0. \quad (10.1)$$

Proof.

□

Theorem 10.1.1 (Weak Law of Large Numbers with Finite Variances)

Let X_1, X_2, \dots be i.i.d. random variables with $EX_i = \mu$ and $\text{Var}(X_i) \leq C < \infty$. Suppose $S_n = X_1 + X_2 + \dots + X_n$, then

$$S_n/n \xrightarrow{L^2} \mu, \quad S_n/n \xrightarrow{p} \mu. \quad (10.2)$$

Proof.

□

Theorem 10.1.2 (Weak Law of Large Numbers without i.i.d.)

Let X_1, X_2, \dots be random variables, Suppose $S_n = X_1 + X_2 + \dots + X_n$, $\mu_n = ES_n$, $\sigma_n^2 = \text{Var}(S_n)$, if $\sigma_n^2/b_n^2 \rightarrow 0$, then

$$\frac{S_n - \mu_n}{b_n} \xrightarrow{p} 0. \quad (10.3)$$

Proof.

□

Theorem 10.1.3 (Weak Law of Large Numbers for Triangular Arrays)

For each n , let $X_{n,m}$, $1 \leq m \leq n$, be independent random variables. Suppose $b_n > 0$ with $b_n \rightarrow \infty$, $\bar{X}_{n,m} = X_{n,m}I_{(|X_{n,m}| \leq b_n)}$, if

1. $\sum_{m=1}^n P(|X_{n,m}| > b_n) \rightarrow 0$, and
2. $b_n^{-2} \sum_{m=1}^n E\bar{X}_{n,m}^2 \rightarrow 0$.

Suppose $S_n = X_{n,1} + \dots + X_{n,n}$ and $a_n = \sum_{m=1}^n E\bar{X}_{n,m}$, then

$$\frac{S_n - a_n}{b_n} \xrightarrow{p} 0. \quad (10.4)$$

Proof. □

Theorem 10.1.4 (Weak Law of Large Numbers by Feller)

Let X_1, X_2, \dots be i.i.d. random variables with

$$\lim_{x \rightarrow 0} xP(|X_i| > x) = 0. \quad (10.5)$$

Suppose $S_n = X_1 + X_2 + \dots + X_n$, $\mu_n = E(X_1 I_{(|X_1| < n)})$, then

$$S_n/n - \mu_n \xrightarrow{p} 0. \quad (10.6)$$

Proof. □

Theorem 10.1.5 (Weak Law of Large Numbers)

Let X_1, X_2, \dots be i.i.d. random variables with $E|X_i| < \infty$. Suppose $S_n = X_1 + X_2 + \dots + X_n$, $\mu = EX_i$, then

$$S_n/n \xrightarrow{p} \mu. \quad (10.7)$$

Proof. □

Remark. $E|X_i| = \infty$

10.2 Strong Law of Large Numbers

10.2.1 Borel-Cantelli Lemmas

Lemma 10.2.1 (Borel-Cantelli Lemma)

If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then

$$P(A_n \text{ i.o.}) = 0. \quad (10.8)$$

Lemma 10.2.2 (The Second Borel-Cantelli Lemma)

If $\{A_n\}$ are independent with $\sum_{n=1}^{\infty} P(A_n) = \infty$, then,

$$P(A_n \text{ i.o.}) = 1. \quad (10.9)$$

Corollary 10.2.1

Suppose $\{A_n\}$ are independent with $P(A_n) < 1, \forall n$. If $P(\cup_{n=1}^{\infty} A_n) = 1$ then

$$\sum_{n=1}^{\infty} P(A_n) = \infty, \quad (10.10)$$

and hence $P(A_n \text{ i.o.}) = 1$

Proof.

□

10.2.2 Strong Law of Large Numbers

Theorem 10.2.1 (Strong Law of Large Numbers)

Let X_1, X_2, \dots be i.i.d. random variables with $E|X_i| < \infty$. Suppose $S_n = X_1 + X_2 + \dots + X_n$, $\mu = EX_i$, then

$$S_n/n \xrightarrow{a.s.} \mu. \quad (10.11)$$

10.3 Uniform Law of Large Numbers

Theorem 10.3.1 (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. $|g(X_i, \theta)| \leq G(X_i)$.
4. $EG(X_i) < \infty$.

Then

$$\sup_{\theta \in \Theta} \left| n^{-1} \sum_{i=1}^n g(X_i, \theta) - Eg(X_i, \theta) \right| \xrightarrow{p} 0. \quad (10.12)$$

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

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

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

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

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

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

$$\begin{aligned}
& \sup_{\theta \in \Theta} \left[n^{-1} \sum_{i=1}^n g(X_i, \theta) - Eg(X_i, \theta) \right] \\
&= \max_k \sup_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))} \left[n^{-1} \sum_{i=1}^n g(X_i, \theta) - Eg(X_i, \theta) \right] \\
&\leq \max_k \left[n^{-1} \sum_{i=1}^n \sup_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))} g(X_i, \theta) - E \inf_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))} g(X_i, \theta) \right]
\end{aligned}$$

Since

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

by the Weak Law of Large Numbers (Theorem 10.1.5),

$$\begin{aligned}
&= o_p(1) + \max_k \left[E \sup_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))} g(X_i, \theta) - E \inf_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))} g(X_i, \theta) \right] \\
&= o_p(1) + \max_k E \Delta_{\delta_\varepsilon(\theta_k)}(X_i, \theta_k) \\
&\leq o_p(1) + \varepsilon
\end{aligned}$$

By analogous argument,

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

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

Chapter 11

Central Limit Theorems

11.1 Classic Central Limit Theorem

11.1.1 The De Moivre-Laplace Theorem

Lemma 11.1.1 (Stirling's Formula)

$$n! \sim \sqrt{2\pi n} n^{n+\frac{1}{2}} e^{-n} \text{ as } n \rightarrow \infty. \quad (11.1)$$

Proof.

□

Lemma 11.1.2

If $c_j \rightarrow 0$, $a_j \rightarrow \infty$ and $a_j c_j \rightarrow \lambda$, then

$$(1 + c_j)^{a_j} \rightarrow e^\lambda. \quad (11.2)$$

Proof.

□

Theorem 11.1.1 (The De Moivre-Laplace Theorem)

Let X_1, X_2, \dots be i.i.d. with $P(X_1 = 1) = P(X_1 = -1) = 1/2$ and let $S_n = X_1 + \dots + X_n$. If $a < b$, then as $m \rightarrow \infty$

$$P(a \leq S_m/\sqrt{m} \leq b) \rightarrow \int_a^b (2\pi)^{-1/2} e^{-x^2/2} dx. \quad (11.3)$$

Proof. If n and k are integers

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

By lemma 11.1.1, we have

$$\begin{aligned} \binom{2n}{n+k} &= \frac{(2n)!}{(n+k)!(n-k)!} \\ &\sim \frac{(2n)^{2n}}{(n+k)^{n+k}(n-k)^{n-k}} \cdot \frac{(2\pi(2n))^{1/2}}{(2\pi(n+k))^{1/2}(2\pi(n-k))^{1/2}} \end{aligned}$$

Hence,

$$\begin{aligned}
 P(S_{2n} = 2k) &= \binom{2n}{n+k} 2^{-2n} \\
 &\sim \left(1 + \frac{k}{n}\right)^{-n-k} \cdot \left(1 - \frac{k}{n}\right)^{-n+k} \\
 &\quad \cdot (\pi n)^{-1/2} \cdot \left(1 + \frac{k}{n}\right)^{-1/2} \cdot \left(1 - \frac{k}{n}\right)^{-1/2} \\
 &= \left(1 - \frac{k^2}{n^2}\right)^{-n} \cdot \left(1 + \frac{k}{n}\right)^{-k} \cdot \left(1 - \frac{k}{n}\right)^k \\
 &\quad \cdot (\pi n)^{-1/2} \cdot \left(1 + \frac{k}{n}\right)^{-1/2} \cdot \left(1 - \frac{k}{n}\right)^{-1/2}
 \end{aligned}$$

Let $2k = x\sqrt{2n}$, i.e., $k = x\sqrt{\frac{n}{2}}$. By lemma 11.1.2, we have

$$\begin{aligned}
 \left(1 - \frac{k^2}{n^2}\right)^{-n} &= (1 - x^2/2n)^{-n} \rightarrow e^{x^2/2} \\
 \left(1 + \frac{k}{n}\right)^{-k} &= (1 + x/\sqrt{2n})^{-x\sqrt{n/2}} \rightarrow e^{-x^2/2} \\
 \left(1 - \frac{k}{n}\right)^k &= (1 - x/\sqrt{2n})^{x\sqrt{n/2}} \rightarrow e^{-x^2/2}
 \end{aligned}$$

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

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

Putting things together, we have

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

Therefore,

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

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 $2\mathbb{Z}/\sqrt{2n} = \{2z/\sqrt{2n} : z \in \mathbb{Z}\}$. As $n \rightarrow \infty$, 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 $m \rightarrow \infty$,

$$P(a \leq S_m/\sqrt{m} \leq b) \rightarrow \int_a^b (2\pi)^{-1/2} e^{-x^2/2} dx.$$

□

11.1.2 Classic Central Limit Theorem

Theorem 11.1.2 (Classic Central Limit Theorem (i.i.d.))

Let X_1, X_2, \dots be i.i.d. with $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2 < \infty$. Let $S_n = X_1 + X_2 + \dots + X_n$, then

$$\frac{S_n - n\mu}{\sigma n^{\frac{1}{2}}} \xrightarrow{d} \chi, \quad (11.4)$$

where χ has the standard normal distribution.

Proof.

□

Theorem 11.1.3 (The Linderberg-Feller Central Limit Theorem)

For each n , let $X_{n,m}$, $1 \leq m \leq n$, be independent random variables with $EX_{n,m} = 0$. If

1. $\sum_{m=1}^n EX_{n,m}^2 \rightarrow \sigma^2 > 0$.
2. $\forall \epsilon > 0, \lim_{n \rightarrow \infty} \sum_{m=1}^n E(|X_{n,m}|^2; |X_{n,m}| > \epsilon) = 0$

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

Theorem 11.1.4 (Berry-Esseen Theorem)

Let X_1, X_2, \dots, X_n be i.i.d. with distribution F , which has a mean μ and a finite third moment σ^3 , then there exists a constant C (independent of F),

$$|G_n(x) - \Phi(x)| \leq \frac{C}{\sqrt{n}} \frac{E|X_1 - \mu|^3}{\sigma^3}, \quad \forall x. \quad (11.5)$$

Corollary 11.1.1

Under the assumptions of Theorem 51 ,

$$G_n(x) \rightarrow \Phi(x) \text{ as } n \rightarrow \infty$$

for any sequence F_n with mean ξ_n and variance σ_n^2 for which

$$\frac{E_n |X_1 - \xi_n|^3}{\sigma_n^3} = o(\sqrt{n})$$

and thus in particular if (72) is bounded. Here E_n denotes the expectation under F_n .

11.2 Central Limit Theorem for independent non-identical Random Variables

Theorem 11.2.1 (The Liapounov Central Limit Theorem)

11.3 Central Limit Theorem for dependent Random Variables

Chapter 12

The Delta Methods

Theorem 12.0.1 (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 (9.4.2),

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

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 (9.4.4),

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

□

Chapter 13

Exercises for Probability Theory and Examples

13.1 Measure Theory

Exercise. 1. Show that if $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$ 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.

Proof. 1. **Complement:** Suppose $A \in \cup_i \mathcal{F}_i$, since $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$, 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 $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$, 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 $A_i = [1, 2 - \frac{1}{i}] \in \mathcal{F}_i$,

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

□

13.2 Laws of Large Numbers

13.3 Central Limit Theorems

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

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

Proof. Let $Y_n \stackrel{d}{=} X_n, 1 \leq n \leq \infty$ with $Y_n \xrightarrow{a.s.} Y_\infty$ (Lemma 9.4.1). Since $g \geq 0$ be continuous, $g(Y_n) \xrightarrow{a.s.} g(Y_\infty)$ and $g(Y_n) \geq 0$ (Theorem 9.4.2), and the Fatou's Lemma (8.4.2) implies,

$$\begin{aligned} \liminf_{n \rightarrow \infty} Eg(X_n) &= \liminf_{n \rightarrow \infty} Eg(Y_n) \geq E\left(\liminf_{n \rightarrow \infty} g(Y_n)\right) \\ &= Eg(Y_\infty) = Eg(X_\infty). \end{aligned}$$

□

Exercise. Suppose g, h are continuous with $g(x) > 0$, and $|h(x)|/g(x) \rightarrow 0$ as $|x| \rightarrow \infty$. If $F_n \xrightarrow{d} F$ and $\int g(x)dF_n(x) \leq C < \infty$, then

$$\int h(x)dF_n(x) \rightarrow \int h(x)dF(x).$$

Proof.

$$\begin{aligned} \left| \int h(x)dF_n(x) - \int h(x)dF(x) \right| &= \left| \int_{x \in [-M, M]} h(x)dF_n(x) + \int_{x \notin [-M, M]} h(x)dF_n(x) \right. \\ &\quad \left. - \int_{x \in [-M, M]} h(x)dF(x) - \int_{x \notin [-M, M]} h(x)dF(x) \right| \\ &\leq \left| \int_{x \in [-M, M]} h(x)dF_n(x) - \int_{x \in [-M, M]} h(x)dF(x) \right| \\ &\quad + \left| \int_{x \notin [-M, M]} h(x)dF_n(x) - \int_{x \notin [-M, M]} h(x)dF(x) \right|. \end{aligned}$$

Let $X_n, 1 \leq n < \infty$, with distribution F_n , so that $X_n \xrightarrow{a.s.} X$ (Lemma 9.4.1).

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

By Continuity Mapping Theorem (9.4.2), $\lim_{n \rightarrow \infty} |E(h(X_n) - h(X)) I_{x \in [-M, M]}| = 0$. Since

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

and by Exercise 13.3

$$Eg(X) \leq \liminf_{n \rightarrow \infty} Eg(X_n) = \liminf_{n \rightarrow \infty} \int g(x)dF_n(x) \leq C < \infty,$$

$$\begin{aligned} \left| \int_{x \notin [-M, M]} h(x)dF_n(x) - \int_{x \notin [-M, M]} h(x)dF(x) \right| &= |E(h(X_n) - h(X)) I_{x \notin [-M, M]}| \\ &\leq 2E \max(h(X_n), h(X)) I_{x \notin [-M, M]} \leq 2C \sup_{x \notin [-M, M]} \frac{h(x)}{g(x)}. \end{aligned}$$

Hence, let $M \rightarrow \infty$,

$$\lim_{n \rightarrow \infty} \left| \int h(x)dF_n(x) - \int h(x)dF(x) \right| \leq 2C \sup_{x \notin [-M, M]} \frac{h(x)}{g(x)} \rightarrow 0,$$

which means,

$$\int h(x) dF_n(x) \rightarrow \int h(x) dF(x).$$

□

Exercise. Let X_1, X_2, \dots 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. Show that if $|X_i| \leq M$ and $\sum_n \text{Var}(X_n) = \infty$, then

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

Exercise. 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 VI

Stochastic Process

Chapter 14

Martingales

14.1 Conditional Expectation

Definition 14.1.1 (Conditional Expectation)

Example. 1. If $X \in \mathcal{F}$, then

$$E(X | \mathcal{F}) = X.$$

2. If X is independent of \mathcal{F} , then

$$E(X | \mathcal{F}) = E(X).$$

3. If $\Omega_1, \Omega_2, \dots$ is a finite or infinite partition of Ω into disjoint sets, each of which has positive probability, and let $\mathcal{F} = \sigma(\Omega_1, \Omega_2, \dots)$, then

$$E(X | \mathcal{F}) = \frac{E(X; \Omega_i)}{P(\Omega_i)} \quad \text{on } \Omega_i.$$

Property.

14.2 Martingales

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

Definition 14.2.1 (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., $E|X_n| < \infty$
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.,

$$E(X_{n+1} | \mathcal{F}_n) = X_n, \quad \forall n \tag{14.1}$$

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

Example (Linear Martingale).

Example (Quadratic Martingale).

Example (Exponential Martingale).

Example (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 $\mathcal{F}_n = \sigma(\xi_1, \dots, \xi_n)$ 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,

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

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

Remark. 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:

$$P\left(\max_{1 \leq m \leq n} |X_m| \geq x\right) \leq x^{-2} \text{var}(X_n) \quad (14.2)$$

Theorem 14.2.1 (Orthogonality of Martingale Increments)

Theorem 14.2.2 (Conditional Variance Formula)

Definition 14.2.2 (Predictable Sequence)

Definition 14.2.3 (Stopping Time)

Theorem 14.2.3 (Martingale Convergence Theorem)

14.3 Doob's Inequality

Theorem 14.3.1 (Doob's Decomposition)

Theorem 14.3.2 (Doob's Inequality)

Theorem 14.3.3 (L^p Maximum Inequality)

14.4 Uniform Integrability

14.5 Optional Stopping Theorems

Chapter 15

Markov Chains

15.1 Markov Chain

Definition 15.1.1 (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, \dots, i_{n-1}, i , and j

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = j \mid X_n = i) \quad (15.1)$$

and the transition probability is

$$p(i, j) = P(X_{n+1} = j \mid X_n = i) \quad (15.2)$$

Example (Random Walk). Suppose $X_n = X_0 + \xi_1 + \dots + \xi_n$, 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 ,

$$\begin{aligned} & P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) \\ &= P(X_n + \xi_{n+1} = j \mid X_n = i) = P(\xi_{n+1} = j - i) = \mu(\{j - i\}) \end{aligned}$$

□

Definition 15.1.2 (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

$$Z_{n+1} = \begin{cases} \xi_1^{n+1} + \dots + \xi_{Z_n}^{n+1} & Z_n > 0 \\ 0 & Z_n = 0 \end{cases} \quad (15.3)$$

Z_n is called a Branching process.

Remark. 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 (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,

$$\begin{aligned} & P(Z_{n+1} = j \mid Z_n = i, Z_{n-1} = i_{n-1}, \dots, Z_0 = i_0) \\ &= P\left(\sum_{k=1}^{Z_n} \xi_k^{n+1} = j \mid Z_n = i\right) = P\left(\sum_{k=1}^i \xi_k = j\right) \end{aligned}$$

□

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

Definition 15.1.3 (Transition Probability)

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

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

Definition 15.1.4 (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

$$P(X_{n+1} \in B \mid \mathcal{F}_n) = p(X_n, B) \quad (15.4)$$

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

$$P(X_j \in B_j, 0 \leq j \leq n) = \int_{B_0} \mu(dx_0) \int_{B_1} p(x_0, dx_1) \cdots \int_{B_n} p(x_{n-1}, dx_n) F \quad (15.5)$$

15.2 Markov Properties

Definition 15.2.1 (Shift Operator)

Theorem 15.2.1 (Markov Property)

Corollary 15.2.1 (Chapman-Kolmogorov Equation)

Theorem 15.2.2 (Strong Markov Property)

15.3 Recurrence and Transience

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

$$T_y^k = \inf \{n > T_y^{k-1} : X_n = y\} \quad (15.6)$$

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

$$\rho_{xy} = P_x(T_y < \infty) \quad (15.7)$$

and we have

$$P_x(T_y^k < \infty) = \rho_{xy}\rho_{yy}^{k-1} \quad (15.8)$$

Proof.

□

Let

$$N(y) = \sum_{n=1}^{\infty} 1_{(X_n=y)} \quad (15.9)$$

be the number of visits to y at positive times.

Definition 15.3.1 (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 15.3.2

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.

$$\begin{aligned} E_x N(y) &= \sum_{k=1}^{\infty} P_x(N(y) \geq k) = \sum_{k=1}^{\infty} P_x(T_y^k < \infty) \\ &= \sum_{k=1}^{\infty} \rho_{xy}\rho_{yy}^{k-1} = \frac{\rho_{xy}}{1 - \rho_{yy}} < \infty \end{aligned}$$

□

Definition 15.3.3 (Closed State Set)

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

$$x \in C, \rho_{xy} > 0 \Rightarrow y \in C. \quad (15.10)$$

Definition 15.3.4 (Irreducible State Set)

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

$$x, y \in D \Rightarrow \rho_{xy} > 0. \quad (15.11)$$

Theorem 15.3.1

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 15.3.2

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 $w \in C_x$. \square

Example (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. \square

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

1. x is transient, if

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

2. x is recurrent, if

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

15.4 Stationary Measures

15.5 Asymptotic Behavior

15.6 Ergodic Theorems

Definition 15.6.1 (Stationary Sequence)**Theorem 15.6.1 (Ergodic Theorem)**

Example.

Chapter 16

Brownian Motion

Definition 16.0.1 (Brownian Motion (1))

A real-valued stochastic process $B(t), t \geq 0$ is said to be Brownian motion, if

1. for any $0 = t_0 \leq t_1 \leq \dots \leq t_n$ the increments

$$B(t_1) - B(t_0), \dots, B(t_n) - B(t_{n-1})$$

are independent

2. for any $s, t \geq 0$ and Borel sets $A \in \mathbb{R}$,

$$P(B(s+t) - B(s) \in A) = \int_A (2\pi t)^{-1/2} \exp(-x^2/2t) \, dx \quad (16.1)$$

3. the sample paths $t \rightarrow B(t)$ are a.s. continuous

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

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

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

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

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

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

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

$$\begin{aligned} EB_s B_t &= E[B_s \cdot ((B_t - B_s) + B_s)] \\ &= E[B_s \cdot (B_t - B_s)] + EB_s^2 \end{aligned}$$

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

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

Thus

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

And

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

□

A second equivalent definition of Brownian motion are as followed,

Definition 16.0.2 (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 $t \rightarrow B(t)$ are a.s. continuous

^aGaussian process, i.e., all its finite dimensional distributions are multivariate normal.

16.1 Markov Properties

16.2 Martingales

Example (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

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

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

□

Example (Exponential Martingale). Suppose B_t is a Brownian motion, then

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

is a martingale.

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

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

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

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

□

Theorem 16.2.1 (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

16.3 Sample Paths

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

$$\Delta_i^n B = B(t_{i+1}^n) - B(t_i^n) \quad (16.2)$$

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

Theorem 16.3.1

$$\lim_{n \rightarrow \infty} \sum_{i=0}^{n-1} (\Delta_i^n B)^2 = T \quad \text{in } L^2 \quad (16.3)$$

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

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

it follows that

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

□

Definition 16.3.1 (Variation)

The variation of a function $f : [0, T] \rightarrow \mathbb{R}$ is defined to be

$$\limsup_{\Delta t \rightarrow 0} \sum_{i=0}^{n-1} |f(t_{i+1}) - f(t_i)| \quad (16.4)$$

where $t = (t_0, t_1, \dots, t_n)$ is a partition of $[0, T]$, i.e. $0 = t_0 < t_1 < \dots < t_n = T$, and where

$$\Delta t = \max_{i=0, \dots, n-1} |t_{i+1} - t_i| \quad (16.5)$$

Theorem 16.3.2

The variation of the paths of $B(t)$ is infinite a.s..

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

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

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

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

By Theorem 16.3.1, we have

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

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

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

Since

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

hence,

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

while

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

□

16.4 Itô Stochastic Calculus

Definition 16.4.1 (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

$$I_T(f) = I(1_{[0,T)}f) \quad (16.6)$$

which can be denoted by

$$\int_0^T f(t) dB(t) \quad (16.7)$$

Property. The Itô Stochastic Integral has the following properties:

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

$$\int_0^t (\alpha f(r) + \beta g(r)) dB(r) = \alpha \int_0^t f(r) dB(r) + \beta \int_0^t g(r) dB(r) \quad (16.8)$$

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

$$E \left(\left| \int_0^t f(r) dB(r) \right|^2 \right) = E \left(\int_0^t |f(r)|^2 dr \right) \quad (16.9)$$

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

$$E \left(\int_0^t f(r) dB(r) \mid \mathcal{F}_s \right) = \int_0^s f(r) dB(r) \quad (16.10)$$

Proof.

□

Definition 16.4.2 (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

$$\xi(T) = \xi(0) + \int_0^T a(t) dt + \int_0^T b(t) dB(t) \quad \text{a.s.} \quad (16.11)$$

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

$$\int_0^T |a(t)| dt < \infty \quad \text{a.s.} \quad (16.12)$$

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

$$d\xi(t) = a(t) dt + b(t) dB(t) \quad (16.13)$$

Remark. The class of all adapted processes $a(t)$ satisfying 16.12 for some $T > 0$ will be denoted by \mathcal{L}_T^1 .

Theorem 16.4.1 (Itô Formula)

Suppose $F(t, x)$ 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

$$\begin{aligned} dF(t, \xi(t)) = & \left(F'_t(t, \xi(t)) + F'_x(t, \xi(t))a(t) + \frac{1}{2}F''_{xx}(t, \xi(t))b(t)^2 \right) dt \\ & + F'_x(t, \xi(t))b(t) dB(t) \end{aligned} \quad (16.14)$$

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

$$dF(t, B(t)) = \left(F'_t(t, B(t)) + \frac{1}{2}F''_{xx}(t, B(t)) \right) dt + F'_x(t, B(t)) dB(t) \quad (16.15)$$

Example (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.

Example.

Chapter 17

Exercises for Probability Theory and Examples

17.1 Martingales

17.2 Markov Chains

17.3 Ergodic Theorems

17.4 Brownian Motion

17.5 Applications to Random Walk

17.6 Multidimensional Brownian Motion

Part VII

Statistics Inference

Chapter 18

Introduction

18.1 Populations and Samples

18.2 Statistics

18.2.1 Sufficient Statistics

Definition 18.2.1 (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 18.2.1 (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.

□

18.2.2 Complete Statistics

Definition 18.2.2 (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 .

18.3 Estimators

Definition 18.3.1 (Estimator)

An estimator is a real-valued function defined over the sample space, that is

$$\delta : \mathbf{X} \rightarrow \mathbb{R}. \quad (18.1)$$

It is used to estimate an estimand, θ , a real-valued function of the parameter.

Unbiasedness

Definition 18.3.2 (Unbiasedness)

An estimator $\hat{\theta}$ of θ is unbiased if

$$E\hat{\theta} = \theta, \quad \forall \theta \in \Theta. \quad (18.2)$$

Remark. • Unbiased estimators of θ may not exist.

Example (Nonexistence of Unbiased Estimator).

Consistency

Definition 18.3.3 (Consistency)

An estimator $\hat{\theta}_n$ of θ is consistent if

$$\lim_{n \rightarrow \infty} P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = 0, \quad \forall \varepsilon > 0, \quad (18.3)$$

that is,

$$\hat{\theta}_n \xrightarrow{p} \theta. \quad (18.4)$$

Example (Consistency of Sample Moments).

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

Asymptotic Normality

Definition 18.3.4 (Asymptotic Normality)

An estimator $\hat{\theta}_n$ of θ is asymptotic normality if

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma_\theta^2). \quad (18.5)$$

Efficiency

Definition 18.3.5 (Efficiency)

Robustness

Definition 18.3.6 (Robustness)

Chapter 19

Maximum Likelihood Estimator

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

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.,

$$p(\mathbf{X}_n; \theta) = \prod_{i=1}^n p(x_i; \theta). \quad (19.1)$$

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

$$\hat{\theta}(\mathbf{X}_n) = \max_{\theta \in \Theta} L(\theta; \mathbf{X}_n). \quad (19.2)$$

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

$$\frac{l(\theta; \mathbf{X}_n)}{n} = \frac{\log L(\theta; \mathbf{X}_n)}{n} = \frac{1}{n} \sum_{i=1}^n \log p(X_i; \theta) = \frac{1}{n} \sum_{i=1}^n l(\theta; X_i). \quad (19.3)$$

Define

$$\begin{aligned} Q(\theta; \mathbf{X}_n) &:= \frac{1}{n} \sum_{i=1}^n l(\theta; X_i), \\ \hat{\theta}(\mathbf{X}_n) &:= \max_{\theta \in \Theta} Q(\theta; \mathbf{X}_n). \end{aligned} \quad (19.4)$$

19.1 Consistency of MLE

By the Weak Law of Large Numbers (Theorem 10.1.5), we can get,

$$\frac{1}{n} \sum_{i=1}^n l(\theta; X_i) \xrightarrow{p} E[l(\theta; X)]. \quad (19.5)$$

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 19.1.1

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 19.1.1 (Consistency of MLE)

Suppose that $Q(\theta; \mathbf{X}_n)$ 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. $Q(\theta; \mathbf{X}_n)$ converges uniformly in probability to $Q_0(\theta)$.

then

$$\hat{\theta}(\mathbf{X}_n) \xrightarrow{P} \theta_0. \quad (19.6)$$

Proof. $\forall \epsilon > 0$, 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 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 $A_n = \left\{ \sup_{\theta \in \Theta \cap \Theta(\epsilon)^C} |Q(\theta; \mathbf{X}_n) - Q_0(\theta)| < \frac{\delta}{2} \right\}$, then

$$\begin{aligned} A_n &\Rightarrow Q(\theta; \mathbf{X}_n) < Q_0(\theta) + \frac{\delta}{2} \\ &\leq Q_0(\theta^*) + \frac{\delta}{2} \\ &= Q_0(\theta_0) - \frac{\delta}{2} \end{aligned}$$

2. For $\theta \in \Theta(\epsilon)$. Let $B_n = \left\{ \sup_{\theta \in \Theta(\epsilon)} |Q(\theta; \mathbf{X}_n) - Q_0(\theta)| < \frac{\delta}{2} \right\}$, then

$$B_n \Rightarrow Q(\theta; \mathbf{X}_n) > Q_0(\theta) - \frac{\delta}{2}, \forall \theta \in \Theta(\epsilon)$$

By Assumption 1,

$$Q(\theta_0; \mathbf{X}_n) > Q_0(\theta_0) - \frac{\delta}{2}$$

If both A_n and B_n hold, then

$$\hat{\theta} \in \Theta(\epsilon).$$

By Assumption 4, we can concluded that $P(A_n \cap B_n) \rightarrow 1$, so

$$P(\hat{\theta} \in \Theta(\epsilon)) \rightarrow 1,$$

which means,

$$\hat{\theta}(\mathbf{X}_n) \xrightarrow{P} \theta_0.$$

□

19.2 Asymptotic Normality of MLE

19.3 Efficiency of MLE

Chapter 20

Minimum-Variance Unbiased Estimator

Definition 20.0.1 (UMVU Estimators)

An unbiased estimator $\delta(\mathbf{X})$ of $g(\theta)$ is the uniform minimum variance unbiased (UMVU) estimator of $g(\theta)$ if

$$\text{Var}_\theta \delta(\mathbf{X}) \leq \text{Var}_\theta \delta'(\mathbf{X}), \quad \forall \theta \in \Theta, \quad (20.1)$$

where $\delta'(\mathbf{X})$ is any other unbiased estimator of $g(\theta)$.

Remark. 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

$$E_\theta \delta(T(\mathbf{X})) = g(\theta), \quad \forall \theta \in \Theta, \quad (20.2)$$

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

$$E[\delta(\mathbf{X}) \mid T(\mathbf{X})]. \quad (20.3)$$

Example (Estimating Polynomials of a Normal Variance). Let X_1, \dots, X_n be distributed with joint density

$$\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp \left[-\frac{1}{2\sigma^2} \sum (x_i - \xi)^2 \right]. \quad (20.4)$$

Discussing the UMVU estimators of ξ^r , σ^r , ξ/σ .

Proof. 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}/σ .

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

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

therefore,

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

Hence,

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

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 $s_x^2 = \sum (x_i - \bar{X})^2$.

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,

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

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. Let X_1, \dots, 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 .

Proof. 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

$$\begin{aligned}
 E[X_{(1)}] &= 2\theta_2 E[Y_{(1)}] + (\theta_1 - \theta_2) \\
 &= 2\theta_2 \int_0^1 ny(1-y)^{n-1}dy + (\theta_1 - \theta_2) \\
 &= \theta_1 - \frac{3n+1}{n+1}\theta_2 \\
 E[X_{(n)}] &= 2\theta_2 E[Y_{(n)}] + (\theta_1 - \theta_2) \\
 &= 2\theta_2 \int_0^1 ny^n dy + (\theta_1 - \theta_2) \\
 &= \theta_1 + \frac{n-1}{n+1}\theta_2
 \end{aligned}$$

Thus,

$$\begin{aligned}
 \theta_1 &= E \left[\frac{n-1}{4n} X_{(1)} + \frac{3n+1}{4n} X_{(n)} \right], \\
 \theta_2 &= E \left[-\frac{n+1}{4n} X_{(1)} + \frac{n+1}{4n} X_{(n)} \right],
 \end{aligned}$$

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 21

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 21.0.1 (Bayes Estimator)

The Bayes Estimator δ with respect to Λ is minimizing the Bayes Risk of δ

$$r(\Lambda, \delta) = \int R(\theta, \delta) d\Lambda(\theta) \quad (21.1)$$

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 $X \sim f(x | \theta)$.
- The prior distribution will be denoted by $\Pi \sim \pi(\theta | \lambda)$ 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 $\Pi \sim \pi(\theta | x, \lambda)$ or $\Lambda \sim \gamma(\lambda | x)$, that is

$$\pi(\theta | x, \lambda) = \frac{f(x | \theta) \pi(\theta | \lambda)}{m(x | \lambda)}, \quad (21.2)$$

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

Theorem 21.0.1

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

$$E\{L[\Theta, \delta(x)] | X = x\}. \quad (21.3)$$

Then, $\delta_\Lambda(x)$ is a Bayes Estimator.

Remark. Improper prior

Corollary 21.0.1

Suppose the assumptions of Theorem 21.0.1 hold.

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

$$\delta_\Lambda(x) = E[g(\Theta) | x]. \quad (21.4)$$

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

$$\delta_\Lambda(x) = \frac{E[w(\theta) g(\Theta) | x]}{E[w(\theta) | x]}. \quad (21.5)$$

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 maximizes $P(\Theta \in I | x)$.

Proof.

□

Theorem 21.0.2

Necessary 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.

21.1 Single-Prior Bayes

The Single-Prior Bayes model in a general form as

$$\begin{aligned} X | \theta &\sim f(x | \theta), \\ \Theta | \gamma &\sim \pi(\theta | \lambda), \end{aligned} \quad (21.6)$$

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

$$\int L(\theta, d(x)) \pi(\theta | x) d\theta, \quad (21.7)$$

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

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

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

$$E(\Theta | x) = \frac{\int \theta f(x | \theta) \pi(\theta | \gamma_0) d\theta}{\int f(x | \theta) \pi(\theta | \gamma_0) d\theta}. \quad (21.8)$$

Example. Consider

$$\begin{aligned} X_i &\stackrel{\text{i.i.d.}}{\sim} N(\mu, \Gamma^{-1}), \quad i = 1, 2, \dots, n \\ \mu &\sim N(0, 1), \\ \Gamma &\sim \text{Gamma}(2, 1), \end{aligned}$$

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

Proof.

$$\begin{aligned} p(\mathbf{X} | \mu, \Gamma) &= \Gamma^n (2\pi)^{-\frac{n}{2}} \exp \left[-2\Gamma^2 \sum_{i=1}^n (x_i - \mu)^2 \right], \\ p(\mu) &= \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{\mu^2}{2} \right), \\ p(\Gamma) &= \frac{1}{\Gamma(2)} \Gamma \exp(-\Gamma). \end{aligned}$$

Therefore,

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

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

For μ , we have

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

□

For exponential families

Theorem 21.1.1

21.2 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

$$\begin{aligned} X | \theta &\sim f(x | \theta), \\ \Theta | \gamma &\sim \pi(\theta | \gamma), \\ \Gamma &\sim \psi(\gamma), \end{aligned} \quad (21.9)$$

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

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

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

$$\int L(\theta, d(x)) \pi(\theta | x) d\theta, \quad (21.10)$$

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

$$\pi(\theta | x) = \frac{\int f(x | \theta) \pi(\theta | \gamma) \psi(\gamma) d\gamma}{\int \int f(x | \theta) \pi(\theta | \gamma) \psi(\gamma) d\theta d\gamma}.$$

Remark. The posterior distribution can also be writed as

$$\pi(\theta | x) = \int \pi(\theta | x, \gamma) \pi(\gamma | x) d\gamma,$$

where $\pi(\gamma | x)$ is the posterior distribution of Γ , unconditional on θ . The equation 21.10 can be writed as

$$\int L(\theta, d(x)) \pi(\theta | x) d\theta = \int \left[\int L(\theta, d(x)) \pi(\theta | x, \gamma) d\theta \right] \pi(\gamma | x) d\gamma.$$

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

Example (Poisson Hierarchy). Consider

$$\begin{aligned} X_i | \lambda &\stackrel{\text{i.i.d}}{\sim} \text{Poisson}(\lambda), \quad i = 1, 2, \dots, n \\ \lambda | b &\sim \text{Gamma}(a, b), \quad a \text{ known}, \\ \frac{1}{b} &\sim \text{Gamma}(k, \tau), \end{aligned} \quad (21.11)$$

calculate the Hierarchical Bayes estimator under squared error loss.

Theorem 21.2.1

For the Hierarchical Bayes model (21.9),

$$K[\pi(\lambda | x), \psi(\lambda)] < K[\pi(\theta | x), \pi(\theta)], \quad (21.12)$$

where K is the Kullback-Leibler information for discrimination between two densities.

Proof.

□

Remark.

21.3 Empirical Bayes

21.4 Bayes Prediction

Chapter 22

Hypothesis Testing

Part VIII

Nonparametric Statistics

Chapter 23

Statistical Functionals

23.1 Empirical Distribution Function

Let $X_1, \dots, X_n \sim F$ where $F(x) = \mathbb{P}(X \leq x)$ is a distribution function on the real line.

Definition 23.1.1 (Empirical Distribution Function)

The empirical distribution function \hat{F}_n is the CDF that puts mass $1/n$ at each data point X_i , that,

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x) \quad (23.1)$$

Chapter 24

Kernel Methods

24.1 Positive Definite Kernels

Definition 24.1.1 (Positive Definite Kernel)

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

1. symmetric, that is,

$$K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x}), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X} \quad (24.1)$$

2. positive definite, that is,

$$\sum_{i=1}^n \sum_{j=1}^n c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0, \quad (24.2)$$

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

24.1.1 Construction of the Reproducing Kernel Hilbert Space

Theorem 24.1.1 (Morse-Aronszajn's Theorem)

For any set \mathcal{X} , suppose $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 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 $\{K(\cdot, \mathbf{x})\}_{\mathbf{x} \in \mathcal{X}}$. For any $f, g \in \mathcal{H}_0$, suppose

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

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

$$\langle f, g \rangle = \sum_{i=1}^m \sum_{j=1}^n a_i b_j K(\mathbf{x}_i, \mathbf{y}_j) \quad (24.3)$$

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

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

And, we also have

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

Suppose

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

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

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

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

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

(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:

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

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

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

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

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

Hence, definition in equation 24.3 is a valid inner product, which is a valid pre-RKHS \mathcal{H}_0 . \square

24.1.2 Examples of Kernels

Example (Gaussian Kernel).

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right), \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \quad (24.4)$$

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.

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

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

$$\begin{aligned} \langle \mathbf{x}, \mathbf{y} \rangle^n &= \left(\sum_{i=1}^d x_i y_i \right)^n = \sum_{k_1+k_2+\dots+k_d=n} \left[\binom{n}{k_1, k_2, \dots, k_d} \prod_{i=1}^d (x_i y_i)^{k_i} \right] \\ &= \sum_{k_1+k_2+\dots+k_d=n} \left[\binom{n}{k_1, k_2, \dots, k_d}^{\frac{1}{2}} \prod_{i=1}^d x_i^{k_i} \cdot \binom{n}{k_1, k_2, \dots, k_d}^{\frac{1}{2}} \prod_{i=1}^d y_i^{k_i} \right] \end{aligned}$$

Therefore,

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

Let

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

then,

$$\begin{aligned} K(\mathbf{x}, \mathbf{y}) &= \sum_{n=0}^{+\infty} \sum_{k_1+k_2+\dots+k_d=n} c_{\sigma, n}(\mathbf{x}) f_{n, \mathbf{k}}(\mathbf{x}) \cdot c_{\sigma, n}(\mathbf{y}) f_{n, \mathbf{k}}(\mathbf{y}) \\ &= \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle \end{aligned}$$

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

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i=1}^n \sum_{j=1}^n c_i c_j \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \\ &= \left\langle \sum_{i=1}^n c_i \Phi(\mathbf{x}_i), \sum_{i=1}^n c_i \Phi(\mathbf{x}_i) \right\rangle \geq 0 \end{aligned}$$

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

□

Part IX

Convex Optimization

Chapter 25

Convex Sets

25.1 Affine and Convex Sets

25.1.1 Affine Sets

Definition 25.1.1 (Affine Set)

A nonempty set C is said to be **affine set**, if

$$\forall x_1, x_2 \in C, \theta \in \mathbf{R}, \theta x_1 + (1 - \theta)x_2 \in C.$$

25.1.2 Convex Sets

Definition 25.1.2 (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 25.1.3 (Convex Hull)

The **convex hull** of said to be set C , denoted by $\text{conv } C$ is a set of all convex combinations of points in C ,

$$\text{conv } C = \{\theta_1 x_1 + \dots + \theta_k x_k \mid x_i \in C; \theta_i \geq 0, i = 1, \dots, k; \theta_1 + \dots + \theta_k = 1\}.$$

Remark. The convex hull $\text{conv } C$ is always convex, which is the minimal convex set that contains C .

25.1.3 Cones

Definition 25.1.4 (Cone)

A nonempty set C is said to be **cone**, if

$$\forall x \in C, \theta \geq 0, \theta x \in C.$$

Definition 25.1.5 (Convex Cone)

A nonempty set C is said to be **convex cone**, if

$$\forall x_1, x_2 \in C, \theta_1, \theta_2 \geq 0, \theta_1 x_1 + \theta_2 x_2 \in C.$$

25.2 Some Important Examples

Definition 25.2.1 (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 25.2.2 (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 25.2.3 ((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 25.2.4 (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).

25.3 Generalized Inequalities

25.3.1 Definition of Generalized Inequalities

Definition 25.3.1 (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 25.3.2 (Generalized Inequalities)

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

$$y - x \in K, \quad (25.1)$$

which can be denoted by

$$x \preceq_K y \text{ or } y \succeq_K x. \quad (25.2)$$

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

$$y - x \in \text{int } K, \quad (25.3)$$

which can be denoted by

$$x \prec_K y \text{ or } y \succ_K x. \quad (25.4)$$

Remark. 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} .

25.3.2 Properties of Generalized Inequalities**Theorem 25.3.1 (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 26

Convex Optimization Problems

26.1 Generalized Inequality Constraints

Definition 26.1.1 (With Generalized Inequality Constraints)

A convex optimization problem with generalized inequality constraints is defined to be

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \preceq_{K_i} 0, \quad i = 1, \dots, m \\ & Ax = b \end{aligned} \tag{26.1}$$

where $f_0 : \mathbf{R}^n \rightarrow \mathbf{R}$, $K_i \in \mathbf{R}^{k_i}$ are proper convexes, and $f_i : \mathbf{R}^n \rightarrow \mathbf{R}^{k_i}$ are K_i -convex.

26.1.1 Conic Form Problems

Definition 26.1.2 (Conic Form Problem)

A conic form problem is defined to be

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & Fx + g \preceq_K 0 \\ & Ax = b \end{aligned} \tag{26.2}$$

26.1.2 Semidefinite Programming

26.2 Vector Optimization

Chapter 27

Unconstrained Minimization

27.1 Definition of Unconstrained Minimization

Definition 27.1.1 (Unconstrained Minimization Problem)

The unconstrained minimization problem is defined to be

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad (27.1)$$

where $f : \mathbf{R}^n \rightarrow \mathbf{R}$ is convex and twice continuously differentiable.

Assume the problem is solvable, i.e., there exists an optimal point \mathbf{x}^* , such that,

$$f(\mathbf{x}^*) = \inf_{\mathbf{x}} f(\mathbf{x})$$

and denote it by p^* . Since f is differentiable and convex, the point \mathbf{x}^* be the optimal. if and only if

$$\nabla f(\mathbf{x}^*) = 0 \quad (27.2)$$

Solving (27.1) is equal to finding the solution of (27.2), thus (27.1) can be solved by analytic solution of (27.2) in a few cases, but usually can be solved by an iterative algorithm, i.e.,

$$\exists \mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots \in \text{dom } f, \quad \text{s.t. } f(\mathbf{x}^{(k)}) \rightarrow p^*, \quad \text{as } k \rightarrow \infty$$

This algorithm is terminated when $f(\mathbf{x}^{(k)}) - p^* \leq \epsilon$, where $\epsilon > 0$ is some specified tolerance.

Remark. The initial point $\mathbf{x}^{(0)}$ must lie in $\text{dom } f$, and the sublevel set

$$S = \left\{ \mathbf{x} \in \text{dom } f \mid f(\mathbf{x}) \leq f(\mathbf{x}^{(0)}) \right\}$$

must be closed. Any closed function (Definition 1.1.2)

Example (Quadratic Minimization). The general convex quadratic minimization problem has the form

$$\min_{\mathbf{x}} \frac{1}{2} \mathbf{x}' \mathbf{P} \mathbf{x} + \mathbf{q}' \mathbf{x} + r \quad (27.3)$$

where $\mathbf{P} \in \mathbb{S}_+^n$, $\mathbf{q} \in \mathbb{R}^n$, and $r \in \mathbb{R}$. The optimality condition is

$$\mathbf{P} \mathbf{x}^* + \mathbf{q} = \mathbf{0} \quad (27.4)$$

which is a set of linear equations.

1. If $\mathbf{P} \succ 0$, exists a unique solution $\mathbf{x}^* = -\mathbf{P}^{-1}\mathbf{q}$.
2. If \mathbf{P} is not positive definite, any solution of (27.4) is optimal for (27.3).
3. If (27.4) does not have a solution, then (27.3) is unbounded.

Proof.

1. Obviously.
2. Since $\mathbf{P} \not\succ 0$, i.e.,

$$\exists \mathbf{v}, \quad \text{s.t. } \mathbf{v}'\mathbf{P}\mathbf{v} < 0$$

Let $\mathbf{x} = t\mathbf{v}$, we have

$$f(\mathbf{x}) = t^2 (\mathbf{v}'\mathbf{P}\mathbf{v}/2) + t (\mathbf{q}'\mathbf{v}) + r$$

which converges to $-\infty$ as $t \rightarrow \infty$.

3. Since (27.4) does not have a solution, i.e.,

$$\mathbf{q} \notin \mathcal{R}(\mathbf{P})$$

Let

$$\mathbf{q} = \tilde{\mathbf{q}} + \mathbf{v}$$

where $\tilde{\mathbf{q}}$ is the Euclidean projection of \mathbf{q} onto $\mathcal{R}(\mathbf{P})$, and $\mathbf{v} = \mathbf{q} - \tilde{\mathbf{q}}$. And \mathbf{v} is nonzero and orthogonal to $\mathcal{R}(\mathbf{P})$, i.e., $\mathbf{v}'\mathbf{P}\mathbf{v} = 0$. If we take $\mathbf{x} = t\mathbf{v}$, we have

$$f(\mathbf{x}) = t\mathbf{q}'\mathbf{v} + r = t(\tilde{\mathbf{q}} + \mathbf{v})'\mathbf{v} + r = t(\mathbf{v}'\mathbf{v}) + r$$

which is unbounded below. □

Remark. The least-squares problem is a special case of quadratic minimization, that,

$$\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2^2 = \mathbf{x}'(\mathbf{A}'\mathbf{A})\mathbf{x} - 2(\mathbf{A}'\mathbf{b})'\mathbf{x} + \mathbf{b}'\mathbf{b}$$

The optimality condition is

$$\mathbf{A}'\mathbf{Ax}^* = \mathbf{A}'\mathbf{b}$$

are called the normal equations of the least-squares problem.

Example (Unconstrained Geometric Programming). The unconstrained geometric program in convex form

$$\min_{\mathbf{x}} f(\mathbf{x}) = \log \left(\sum_{i=1}^m \exp(\mathbf{a}_i'\mathbf{x} + b_i) \right)$$

The optimality condition is

$$\nabla f(\mathbf{x}^*) = \frac{\sum_{i=1}^m \exp(\mathbf{a}_i'\mathbf{x}^* + b_i) \mathbf{a}_i}{\sum_{j=1}^m \exp(\mathbf{a}_j'\mathbf{x}^* + b_j)} = \mathbf{0}$$

which has no analytical solution, so we must resort to an iterative algorithm. For this problem, $\text{dom } f = \mathbb{R}^n$, so any point can be chosen as the initial point $\mathbf{x}^{(0)}$.

Example (Analytic Center of Linear Inequalities). Consider the optimization problem

$$\min_{\mathbf{x}} f(\mathbf{x}) = -\sum_{i=1}^m \log(\mathbf{b}_i - \mathbf{a}_i^T \mathbf{x})$$

where the domain of f is the open set

$$\text{dom } f = \{\mathbf{x} \mid \mathbf{a}_i'\mathbf{x} < \mathbf{b}_i, i = 1, \dots, m\}$$

Definition 27.1.2 (Strong Convexity)

27.2 **General Descent Method**

27.3 **Gradient Descent Method**

27.4 **Steepest Descent Method**

27.5 **Newton's Method**

Chapter 28

Exercises for Convex Optimization

28.1 Convex Sets

Exercise. 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 $A \succeq 0$.

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

$$\begin{aligned} & (\theta x + (1 - \theta)y)^T A (\theta x + (1 - \theta)y) + b^T (\theta x + (1 - \theta)y) + c \\ &= \theta^2 x^T A x + \theta(1 - \theta)(y^T A x + x^T A y) + (1 - \theta)^2 y^T A y + \theta b^T x + (1 - \theta)b^T y + c \\ &= \theta^2(x^T A x + b^T x + c) + (1 - \theta)^2(y^T A y + b^T y + c) - \theta^2(b^T x + c) \\ & \quad - (1 - \theta)^2(b^T y + c) + \theta(1 - \theta)(y^T A x + x^T A y) + \theta b^T x + (1 - \theta)b^T y + c \\ &\leq -\theta^2(b^T x + c) - (1 - \theta)^2(b^T y + c) + \theta(1 - \theta)(y^T A x + x^T A y) \\ & \quad + \theta b^T x + (1 - \theta)b^T y + c \\ &= \theta(1 - \theta)[(b^T x + c) + (b^T y + c) + x^T A x + y^T A y] \\ &\leq \theta(1 - \theta)(-x^T A x - y^T A y + x^T A x + y^T A y) \leq 0 \end{aligned}$$

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

□

Part X

Regression Analysis

Chapter 29

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.

29.1 Marginal Likelihood

29.2 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

$$\ell_c = \log(f_{Y|S_{\boldsymbol{\lambda}}}) \quad (29.1)$$

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

Remark. 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.

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

$$\ell(\boldsymbol{\theta}, \mathbf{y}) = \boldsymbol{\theta}'\mathbf{s} - b(\boldsymbol{\theta}) \quad (29.2)$$

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

$$\ell(\boldsymbol{\theta}, \mathbf{y}) = \boldsymbol{\varphi}'\mathbf{s}_1 + \boldsymbol{\lambda}'\mathbf{s}_2 - b(\boldsymbol{\varphi}, \boldsymbol{\lambda}) \quad (29.3)$$

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

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

$$\ell(\varphi \mid \mathbf{s}_2) = \varphi' \mathbf{s}_1 - b^*(\varphi, \boldsymbol{\lambda}) \quad (29.4)$$

which is independent of the nuisance parameter and may be used for inferences regarding φ .

Example. $Y_1 \sim P(\mu_1), Y_2 \sim P(\mu_2)$ 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.

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

$$\begin{aligned} \ell(\phi, \lambda_1) &\propto \log \left[e^{-(\mu_1 + \mu_2)} \mu_1^{y_1} \mu_2^{y_2} \right] \\ &= -(\mu_1 + \mu_2) + y_1 \log(\mu_1) + y_2 \log(\mu_2) \\ &= -\mu_1 \left(1 + \frac{\mu_2}{\mu_1} \right) + y_1 \log(\mu_1) + y_2 \log(\mu_1) \\ &\quad - y_2 [\log(\mu_1) - \log(\mu_2)] \\ &= -e^{\lambda_1} (1 + e^\varphi) + (y_1 + y_2) \lambda_1 - y_2 \varphi \\ &= s_1 \varphi + s_2 \lambda_1 - b(\varphi, \lambda_1) \end{aligned}$$

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,

$$\begin{aligned} \ell(\varphi \mid S_2 = s_2) &\propto y_1 \log\left(\frac{\mu_1}{\mu_1 + \mu_2}\right) + y_2 \log\left(\frac{\mu_2}{\mu_1 + \mu_2}\right) \\ &= y_1 \log\left(\frac{\mu_1}{\mu_1 + \mu_2}\right) + y_2 \log\left(\frac{\mu_1}{\mu_1 + \mu_2}\right) \\ &\quad - y_2 \left[\log\left(\frac{\mu_1}{\mu_1 + \mu_2}\right) - \log\left(\frac{\mu_2}{\mu_1 + \mu_2}\right) \right] \\ &= (y_1 + y_2) \log\left(\frac{1}{1 + e^\varphi}\right) - y_2 \varphi \\ &= s_1 \varphi - b^*(\varphi, s_2) \end{aligned}$$

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

□

29.3 Profile Likelihood

29.4 Quasi Likelihood

Chapter 30

Generalized Linear Model

30.1 Introduction

Definition 30.1.1 (Exponential Family)

An exponential family of probability distributions as those distributions whose density is defined to be

$$f(y \mid \theta, \phi) = \exp \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right] \quad (30.1)$$

Property. The exponential family have the following properties,

$$E(Y) = b'(\theta) \quad \text{Var}(Y) = b''(\theta)a(\phi).$$

Proof.

□

Table 30.1: Common Distributions of Exponential Family

Distribution	Parameter(s)	θ	ϕ	$b(\theta)$	$a(\phi)$	$c(y, \phi)$	$E(Y)$	$\text{Var}(Y)$
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	$\text{Bern}(p)$	$\log\left(\frac{p}{1-p}\right)$	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

Suppose the response Y has a distribution in the exponential family

$$f(y | \theta, \phi) = \exp \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right]$$

with link function g , such that,

$$E(Y | \mathbf{X}) = \mu = g^{-1}(\eta), \quad \eta = \mathbf{X}'\boldsymbol{\beta} \quad (30.2)$$

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**.

Remark. 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 30.2: 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: $0, 1, 2, \dots$	$\log(\mu)$	$\exp(\eta)$

Maximum Likelihood Suppose the log-likelihood function be

$$\ell(\boldsymbol{\beta} | \mathbf{X}, y) = \log[f(y | \theta, \phi)] = \log[f(y | g^{-1}(\eta), \phi)] \quad (30.3)$$

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

Let

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

be the score function and observed information matrix.

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

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

By mean value theorem,

$$\begin{aligned} U(\hat{\boldsymbol{\beta}}) - U(\boldsymbol{\beta}_0) &= \frac{\partial U(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ \Rightarrow -U(\boldsymbol{\beta}_0) &= -A(\boldsymbol{\beta}^*) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \end{aligned}$$

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

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

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:

$$\hat{\beta}_{t+1} = \hat{\beta}_t + A^{-1}(\hat{\beta}_t) U(\hat{\beta}_t) \Leftrightarrow A(\hat{\beta}_t) \hat{\beta}_{t+1} = A(\hat{\beta}_t) \hat{\beta}_t + U(\hat{\beta}_t) \quad (30.4)$$

where

$$U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} \quad (30.5)$$

is the score function and

$$A(\beta) = -\frac{\partial^2 \ell(\beta)}{\partial \beta' \partial \beta} \quad (30.6)$$

is the observed information matrix.

2. Fisher's Scoring Method:

$$\hat{\beta}_{t+1} = \hat{\beta}_t + I^{-1}(\hat{\beta}_t) U(\hat{\beta}_t) \Leftrightarrow I(\hat{\beta}_t) \hat{\beta}_{t+1} = I(\hat{\beta}_t) \hat{\beta}_t + U(\hat{\beta}_t) \quad (30.7)$$

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

$$I(\beta) = E[A(\beta)] = -E\left[\frac{\partial^2 \ell(\beta)}{\partial \beta' \partial \beta}\right] \quad (30.8)$$

is the Fisher information matrix.

Bayesian Methods

30.2 Binary Data

Suppose

$$Y \sim b(m, \pi), \quad i = 1, 2, \dots, n \quad (30.9)$$

with link function

$$\eta = g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \mathbf{x}'\beta \quad (30.10)$$

Remark.

The likelihood function is

$$f(\boldsymbol{\pi} \mid \mathbf{X}, \mathbf{y}) = \prod_{i=1}^n \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i} \quad (30.11)$$

and the log-likelihood function is

$$\begin{aligned} \ell(\beta) &= \log[f(\boldsymbol{\pi} \mid \mathbf{X}, \mathbf{y})] = \sum_{i=1}^n \ell_i(\beta) \\ &= \sum_{i=1}^n \left\{ \log \left[\binom{m_i}{y_i} \right] + y_i \log(\pi_i) + (m_i - y_i) \log(1 - \pi_i) \right\} \\ &= \sum_{i=1}^n \left[y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + m_i \log(1 - \pi_i) \right] + \sum_{i=1}^n \log \left[\binom{m_i}{y_i} \right] \end{aligned} \quad (30.12)$$

where

$$\pi_i = \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i' \boldsymbol{\beta})} \quad (30.13)$$

Thus,

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

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

30.3 Polytomous Data

Definition 30.3.1 (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 scales¹. 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.

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

$$\pi_j(\mathbf{x}_i) = P(Y = y_j \mid \mathbf{X} = \mathbf{x}_i) \quad (30.14)$$

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

$$r_j(\mathbf{x}_i) = P\left(Y \leq \sum_{r \leq j} y_r \mid \mathbf{X} = \mathbf{x}_i\right) \quad (30.15)$$

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

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

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

for non-negative integers y_1, \dots, y_k .

As for the link function, we have

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

Nominal Scale

$$\pi_j(\mathbf{x}_i) = \frac{\exp[\eta_j(\mathbf{x}_i)]}{\sum_{j=1}^k \exp[\eta_j(\mathbf{x}_i)]} \quad (30.16)$$

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:

$$\log \left[\frac{r_j(\mathbf{x}_i)}{1 - r_j(\mathbf{x}_i)} \right] = \theta_j - \mathbf{x}_i' \boldsymbol{\beta} \quad (30.17)$$

2. Complementary Log-Log Scale:

$$\log \{-\log[1 - r_j(\mathbf{x}_i)]\} = \theta_j - \mathbf{x}_i' \boldsymbol{\beta} \quad (30.18)$$

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.

$$\log \left[\frac{r_j(\mathbf{x}_i)}{1 - r_j(\mathbf{x}_i)} \right] = \varsigma_0 + \varsigma_1 \left(\frac{s_j + s_{j+1}}{2} \right) - \mathbf{x}_i' \boldsymbol{\beta} - \mathbf{x}_i' \boldsymbol{\xi} (c_j - \bar{c}) \quad (30.19)$$

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.

$$\pi_j(\mathbf{x}_i) = \frac{\exp[\eta_j(\mathbf{x}_i)]}{\sum_{j=1}^k \exp[\eta_j(\mathbf{x}_i)]} \quad (30.20)$$

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

3.

$$\sum_{j=1}^k \pi_j(\mathbf{x}_i) s_j = \mathbf{x}_i' \boldsymbol{\beta} \quad (30.21)$$

30.4 Count Data

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

$$\text{Var}(Y) = \sigma^2 E(Y) \quad (30.22)$$

with link function

$$\log(\mu) = \eta = \mathbf{x}' \boldsymbol{\beta} \quad (30.23)$$

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

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

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

and the log-likelihood function is

$$\ell(\boldsymbol{\beta}) \propto \sum_{i=1}^n (y_i \log(\mu_i) - \mu_i) \quad (30.24)$$

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

Chapter 31

Survival Analysis

31.1 General Formulation

Definition 31.1.1 (Survival Function)

The survival function^a is defined to be

$$S(t) = P(T > t) = \int_t^\infty f(u) du = 1 - F(t). \quad (31.1)$$

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

^aThe survival function is the probability that the time of death is later than some specified time t .

Definition 31.1.2 (Lifetime Distribution Function)

The lifetime distribution function is defined to be

$$F(t) = P(T \leq t) \quad (31.2)$$

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

$$f(t) = F'(t) = \frac{d}{dt} F(t) \quad (31.3)$$

^aThe function f is sometimes called the event density; it is the rate of death or failure events per unit time.

Definition 31.1.3 (Hazard Function)

The Hazard function^a is defined to be

$$\lambda(t) = \lim_{\varepsilon \rightarrow 0^+} \left[\frac{P(t \leq T < t + \varepsilon \mid T \geq t)}{\varepsilon} \right] = \frac{f(t)}{S(t)} \quad (31.4)$$

^aThe Hazard function is the event rate at time t conditional on survival until time t or later (that is, $T \geq t$).

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

1.

$$\lambda(t) = -\frac{d \log[S(t)]}{dt} \quad (31.5)$$

2.

$$S(t) = \exp \left[- \int_0^t \lambda(x) dx \right] \quad (31.6)$$

3.

$$f(t) = \lambda(t) \exp \left[- \int_0^t \lambda(x) dx \right] \quad (31.7)$$

Proof.

□

Example (Constant Hazards). Suppose

$$\lambda(t) = \lambda \quad (31.8)$$

then

$$\begin{aligned} S(t) &= \exp \left[- \int_0^t \lambda(x) dx \right] = \exp \left[- \int_0^t \lambda dx \right] = \exp(-\lambda t) \\ f(t) &= \lambda(t) \exp \left[- \int_0^t \lambda(x) dx \right] = \lambda \exp \left[- \int_0^t \lambda dx \right] = \lambda \exp(-\lambda t) \end{aligned}$$

which is the exponential distribution.

Example (Bathtub Hazards).

$$\lambda(t) = \alpha t + \frac{\beta}{1 + \gamma t} \quad (31.9)$$

31.2 Estimation of Survival Function

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

$$\begin{aligned} f(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}) &= \prod_{i=1}^n [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} \\ &= \prod_{i=1}^n \left(\frac{f(t_i)}{S(t_i)} \right)^{\delta_i} S(t_i) \\ &= \prod_{i=1}^n [\lambda(t_i)]^{\delta_i} S(t_i) \end{aligned} \quad (31.10)$$

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

Example. Suppose \mathbf{T} have exponential density, that,

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

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 $t_{(k)} \leq t < t_{(k+1)}$,

$$\begin{aligned}\hat{S}(t) &= \prod_{j=1}^k \left(\frac{n_j - d_j}{n_j} \right) \\ &= \left(1 - \frac{d_1}{n_1} \right) \left(1 - \frac{d_2}{n_2} \right) \cdots \left(1 - \frac{d_k}{n_k} \right) \\ &\approx [1 - \hat{\lambda}(t_1)] [1 - \hat{\lambda}(t_2)] \cdots [1 - \hat{\lambda}(t_k)]\end{aligned} \tag{31.11}$$

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

31.3 Proportional Hazards Model

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

Further, let $t_{(1)} \leq t_{(2)} \leq \dots \leq t_{(m)}$ be the ordered uncensored failure times corresponding to $\delta_{(j)} = 1, j = 1, 2, \dots, m$, and $x_{(1)}, x_{(2)}, \dots, x_{(m)}$ 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

$$\lambda(t) = \lambda_0(t) e^{\mathbf{x}'\boldsymbol{\beta}} \tag{31.12}$$

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

The exact likelihood function is

$$\ell[\boldsymbol{\beta}, \lambda_0(t)] = \prod_{i=1}^n [\lambda_i(t_i)]^{\delta_i} S(t_i) \tag{31.13}$$

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

$$\ell_p(\boldsymbol{\beta}) = \prod_{j=1}^m \frac{e^{\mathbf{x}'_{(j)}\boldsymbol{\beta}}}{\sum_{l \in R(t_{(j)})} e^{\mathbf{x}'_l\boldsymbol{\beta}}} = \prod_{i=1}^n \left[\frac{e^{\mathbf{x}'_i\boldsymbol{\beta}}}{\sum_{l \in R(t_i)} e^{\mathbf{x}'_l\boldsymbol{\beta}}} \right]^{\delta_i} \tag{31.14}$$

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 risk set.

Part XI

Machine Learning

Chapter 32

Support Vector Machine

Theorem 32.0.1

The minimizer of

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

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,

$$\begin{aligned} & E \{ [1 - Yg(X)]_+ \mid X = x \} \\ &= [1 - g(x)]_+ P(Y = 1 \mid X = x) + [1 + g(x)]_+ P(Y = -1 \mid X = x) \\ &= [1 - g(x)]_+ p(x) + [1 + g(x)]_+ [1 - p(x)] \\ &= \begin{cases} [1 - g(x)] p(x), & g(x) < -1 \\ 1 + [1 - 2p(x)] g(x), & -1 \leq g(x) \leq 1 \\ [1 + g(x)] [1 - p(x)], & g(x) > 1 \end{cases} \end{aligned}$$

When $g(x) < -1$,

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

When $g(x) > 1$,

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

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

$$\begin{aligned} & \arg \min_g E \{ [1 - Yg(X)]_+ \mid X = x \} \\ &= \arg \min_g \{ 1 + [1 - 2p(x)]g(x) \} \\ &= \begin{cases} -1, & p(x) < \frac{1}{2} \\ 0, & p(x) = \frac{1}{2} \\ 1, & p(x) > \frac{1}{2} \end{cases} \end{aligned}$$

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

Chapter 33

Linear Discriminant Analysis

Chapter 34

K-Nearest Neighbor

Chapter 35

Decision Tree

Part XII

Random Matrix Theory

Chapter 36

Sample Covariance Matrices

Suppose $\{\mathbf{X}\}$ be a sequence of random vectors defined in \mathbb{R}^n , and $(x_i)_{1 \leq i \leq n}$ be the components of the random vector \mathbf{X} , such that

$$E(\mathbf{X}) = 0, \quad E(\mathbf{X} \otimes \mathbf{X}) = \mathbf{I}_n$$

where \mathbf{X} is also called **isotropic** random vector.

Suppose $\{m_n\}$ be a sequence defined in \mathbb{N} such that

$$0 < \underline{\rho} := \liminf_{n \rightarrow \infty} \frac{n}{m_n} \leq \limsup_{n \rightarrow \infty} \frac{n}{m_n} =: \bar{\rho} < \infty$$

Let $\mathbf{X}_1, \dots, \mathbf{X}_{m_n}$ be i.i.d. copies of \mathbf{X} , and \mathbb{X} be the $m_n \times n$ random matrix with i.i.d. rows $\mathbf{X}_1, \dots, \mathbf{X}_{m_n}$, and their empirical covariance matrix is

$$\hat{\Sigma} := \frac{1}{m_n} \sum_{i=1}^{m_n} \mathbf{X}_i \otimes \mathbf{X}_i = \frac{1}{m_n} \mathbb{X}' \mathbb{X}$$

which is a $n \times n$ symmetric positive semidefinite random matrix, and

$$E(\hat{\Sigma}) = \mathbb{E}(\mathbf{X} \otimes \mathbf{X}) = \mathbf{I}_n$$

For convenience, we define the random matrix

$$\mathbf{A} := m_n \hat{\Sigma} = \mathbb{X}' \mathbb{X} = \sum_{i=1}^{m_n} \mathbf{X}_i \otimes \mathbf{X}_i$$

36.1 Eigenvalues and Singular Values

Theorem 36.1.1

The eigenvalues of \mathbf{A} are squares of the singular values of \mathbb{X} , in particular

$$\lambda_{\max}(\mathbf{A}) = s_{\max}(\mathbb{X})^2 = \max_{\|\mathbf{x}\|=1} \|\mathbb{X}\mathbf{x}\|^2 = \|\mathbb{X}\|_2^2$$

if $m_n \geq n$, then

$$\lambda_{\min}(\mathbf{A}) = s_{\min}(\mathbb{X})^2 = \min_{\|\mathbf{x}\|=1} \|\mathbb{X}\mathbf{x}\|^2 = \|\mathbb{X}^{-1}\|_2^{-2}$$

Proof.

□

36.2 Laguerre Orthogonal Ensemble

Definition 36.2.1 (Wishart Distribution)

Suppose \mathbb{X} be a $p \times n$ matrix, each column of which is independently drawn from a p -variate normal distribution with zero mean:

$$\mathbf{X}_i = (x_i^1, \dots, x_i^p)' \sim N_p(0, \Sigma)$$

Then the Wishart distribution is the probability distribution of the $p \times p$ random matrix,

$$\mathbf{M} = \mathbb{X}'\mathbb{X} = \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i' \quad (36.1)$$

and which can be denoted by

$$\mathbf{M} \sim W_p(\Sigma, n)$$

If $p = \Sigma = 1$, then this distribution is a chi-squared distribution with n degrees of freedom.

Theorem 36.2.1

If $n \geq p$, the probability density function of \mathbf{M} is

$$f(\mathbf{M}) = \frac{1}{2^{np/2} [\det(\Sigma)]^{n/2} \Gamma_p\left(\frac{n}{2}\right)} \det(\mathbf{M})^{(n-p-1)/2} \exp\left[-\frac{1}{2} \text{tr}(\Sigma^{-1}\mathbf{M})\right] \quad (36.2)$$

with respect to Lebesgue measure on the cone of symmetric positive definite matrices. Here, Γ_p is the multivariate gamma function defined as

$$\Gamma_p\left(\frac{n}{2}\right) = \pi^{p(p-1)/4} \prod_{j=1}^p \Gamma\left(\frac{n}{2} - \frac{j-1}{2}\right)$$

Remark. Specially, if the random variables $(x_i)_{1 \leq i \leq n}$ are i.i.d. standard Gaussians, then the distribution of the random matrix $\hat{\Sigma}$ can be derived from the Wishart distribution. The probability density function of $\hat{\Sigma}$ can be derived from (36.2), since

$$\mathbf{A} \sim W_n(\mathbf{I}_n, m_n), \quad \det(\hat{\Sigma}) = m_n^{-n} \det(\mathbf{A}), \quad \text{tr}(\hat{\Sigma}) = m_n^{-1} \text{tr}(\mathbf{A})$$

thus,

$$f(\hat{\Sigma}) = \frac{m_n^{-n(m_n-n-1)/2+1}}{2^{m_n n/2} \Gamma_n\left(\frac{m_n}{2}\right)} \det(\hat{\Sigma})^{(m_n-n-1)/2} \exp\left[-\frac{m_n}{2} \text{tr}(\hat{\Sigma})\right] \quad (36.3)$$

Theorem 36.2.2

If the random variables $(x_i)_{1 \leq i \leq n}$ are i.i.d. standard Gaussians, the joint probability density function of eigenvalues of $\widehat{\mathbf{\Sigma}}$ is

$$p(\boldsymbol{\lambda}) = \tilde{Q}_{m_n, n}^{-1} \exp\left(-\frac{m_n}{2} \sum_{k=1}^n \lambda_k\right) \prod_{k=1}^n \lambda_k^{(m_n - n - 1)/2} \prod_{i < j} |\lambda_i - \lambda_j| \quad (36.4)$$

where

$$0 \leq \lambda_1 \leq \dots \leq \lambda_n < \infty$$

and $\tilde{Q}_{m_n, n}$ is the normalization constant.

Proof. First, we will give the characteristic function of $\widehat{\mathbf{\Sigma}}$, i.e.,

$$\varphi_{\widehat{\mathbf{\Sigma}}}(\mathbf{P}) = E \left[\exp \left(i \sum_{1 \leq i \leq j \leq n} P_{ij} \widehat{\mathbf{\Sigma}}_{ji} \right) \right] = E \left[\exp \left(i \operatorname{tr} (\mathbf{P} \widehat{\mathbf{\Sigma}}) \right) \right]$$

where $\{P_{ij}\}_{1 \leq i \leq j \leq n} \in \mathbb{R}^{(n+1)n/2}$ and \mathbf{P} is a real symmetric matrix, that

$$\mathbf{P} = \left\{ \widehat{P}_{ij}, \widehat{P}_{ij} = \widehat{P}_{ji} \right\}_{i,j=1}^n, \quad \widehat{P}_{ij} = \begin{cases} P_{ii}, & i = j \\ P_{ij}/2, & i < j \end{cases}$$

Thus, we have

$$\begin{aligned} &= \int_{\mathbb{R}^{m_n \times n}} \exp \left(i \operatorname{tr} (\mathbf{P} \widehat{\mathbf{\Sigma}}) \right) \cdot (2\pi)^{-m_n n/2} \exp \left(-\frac{1}{2} \sum_{k=1}^{m_n} \sum_{i=1}^n \left(x_i^{(k)} \right)^2 \right) \prod_{k=1}^{m_n} \prod_{i=1}^n dx_i^{(k)} \\ &= \int_{\mathbb{R}^{m_n \times n}} (2\pi)^{-m_n n/2} \exp \left(-\frac{1}{2} \sum_{k=1}^{m_n} \sum_{i=1}^n \sum_{j=1}^n \mathbf{Q}_{ij} x_i^{(k)} x_j^{(k)} \right) \prod_{k=1}^{m_n} \prod_{i=1}^n dx_i^{(k)} \end{aligned}$$

where

$$\mathbf{Q} = \mathbf{I}_n - \frac{2i}{m_n} \mathbf{P}$$

Since $(x_i^{(k)})_{1 \leq i \leq n}$ are i.i.d. standard Gaussians,

$$\begin{aligned} &= \left[\int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp \left(-\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{Q}_{ij} x_i x_j \right) \prod_{i=1}^n dx_i \right]^{m_n} \\ &= \left[\int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp \left(-\frac{1}{2} \mathbf{X}' \mathbf{Q} \mathbf{X} \right) d\mathbf{X} \right]^{m_n} \\ &= \left[\det(\mathbf{Q})^{-\frac{1}{2}} \int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp \left(-\frac{1}{2} \left(\mathbf{Q}^{\frac{1}{2}} \mathbf{X} \right)' \left(\mathbf{Q}^{\frac{1}{2}} \mathbf{X} \right) \right) d\mathbf{Q}^{\frac{1}{2}} \mathbf{X} \right]^{m_n} \\ &= [\det(\mathbf{Q})]^{-m_n/2} \end{aligned}$$

thus,

$$[\det(\mathbf{Q})]^{-m_n/2} = \left[\det \left(\mathbf{I}_n - \frac{2i}{m_n} \mathbf{P} \right) \right]^{-m_n/2} = \prod_{k=1}^n \left(1 - \frac{2i}{m_n} p_k \right)^{-m_n/2} \quad (36.5)$$

where $\{p_k\}_{k=1}^n$ are the eigenvalues of \mathbf{P} .

Then, we will show that the characteristic function of (36.4) coincides with the above function. By the Wishart distribution, the probability density of the real symmetric and positive definite random matrix $\widehat{\mathbf{\Sigma}}$ is

$$\tilde{Q}_{m_n, n}^{-1} \exp \left[-\frac{m_n}{2} \operatorname{tr} (\widehat{\mathbf{\Sigma}}) \right] \left[\det (\widehat{\mathbf{\Sigma}}) \right]^{(m_n - n - 1)/2} d\widehat{\mathbf{\Sigma}} \quad (36.6)$$

where $\tilde{Q}_{m_n, n}$ is the normalization constant. Then, the characteristic function of (36.6), i.e.,

$$\tilde{Q}_{m_n, n}^{-1} \int_{\mathcal{S}_n^+} \exp \left[i \operatorname{tr} (\mathbf{P} \widehat{\mathbf{\Sigma}}) - \frac{m_n}{2} \operatorname{tr} (\widehat{\mathbf{\Sigma}}) \right] \left[\det (\widehat{\mathbf{\Sigma}}) \right]^{(m_n - n - 1)/2} d\widehat{\mathbf{\Sigma}}$$

where the integration is over the set \mathcal{S}_n^+ of $n \times n$ real symmetric and positive definite matrices. Since

$$\sum_{k=1}^n \lambda_k = \operatorname{tr} (\widehat{\mathbf{\Sigma}}), \quad \prod_{k=1}^n \lambda_k^{(m_n - n - 1)/2} = \left[\det (\widehat{\mathbf{\Sigma}}) \right]^{(m_n - n - 1)/2}$$

and

$$d\widehat{\mathbf{\Sigma}} = \prod_{i < j} |\lambda_i - \lambda_j| d\boldsymbol{\lambda} H_1(dO)$$

where H_1 is the normalized Haar measure of $O(n)$, and the integration over $\boldsymbol{\lambda}$ and $O \in O(n)$ are independent. Since the orthogonal invariance of the density of (36.6), and the characteristic function is

$$Q_{m_n, n}^{-1} \int_{(\mathbb{R}_+)^n} \exp \left[\sum_{k=1}^n \left(ip_k - \frac{m_n}{2} \right) \lambda_k \right] \prod_{k=1}^n \lambda_k^{(m_n - n - 1)/2} \prod_{i < j} |\lambda_i - \lambda_j| d\boldsymbol{\lambda} \quad (36.7)$$

where $Q_{m_n, n} = m_n! \tilde{Q}_{m_n, n}$.

If we viewed (36.5) and (36.7) as the function of $\{p_k\}_{k=1}^n \in \mathbb{R}^n$, then they can be **analytic continuation** to the domain

$$\{p_k + ip'_k, p'_k \geq 0\}_{k=1}^n$$

If we replace $\{p_k\}_{k=1}^n$ by $\{ip'_k, p'_k \geq 0\}_{k=1}^n$ on (36.5), since this is a set of uniqueness of both (36.5) and (36.7) analytic functions, we have

$$Q_{m_n, n}^{-1} \int_{(\mathbb{R}_+)^n} \exp \left[-\frac{m_n}{2} \sum_{k=1}^n q_k \lambda_k \right] \prod_{k=1}^n \lambda_k^{(m_n - n - 1)/2} \prod_{i < j} |\lambda_i - \lambda_j| d\boldsymbol{\lambda}$$

where $q_k = 1 + \frac{2p'_k}{m_n} \geq 1, k = 1, \dots, n$, and since

$$\forall i, j \quad \frac{q_i}{q_j} = \frac{1 + \frac{2p'_i}{m_n}}{1 + \frac{2p'_j}{m_n}} \rightarrow 1, \quad \text{as } m_n \rightarrow \infty$$

we have

$$\prod_{i < j} |q_i \lambda_i - q_j \lambda_j| = \prod_{i < j} q_i \left| \lambda_i - \frac{q_j}{q_i} \lambda_j \right| \rightarrow \prod_{k=1}^n q_k^{(n-1)/2} \prod_{i < j} |\lambda_i - \lambda_j|, \quad \text{as } m_n \rightarrow \infty$$

thus,

$$\prod_{k=1}^n q_k^{-m_n/2} \cdot Q_{m_n, n}^{-1} \int_{(\mathbb{R}_+)^n} \exp \left[-\frac{m_n}{2} \sum_{k=1}^n q_k \lambda_k \right] \prod_{k=1}^n (q_k \lambda_k)^{(m_n-n-1)/2} \cdot \prod_{i < j} |q_i \lambda_i - q_j \lambda_j| \, d\mathbf{q} \, d\boldsymbol{\lambda}$$

Since

$$\forall k \quad q_k \lambda_k \rightarrow \lambda_k, \quad \text{as} \quad m_n \rightarrow \infty$$

we can "lifting" from $\{\lambda_k\}_{k=1}^n$ to \mathcal{S}_n^+ bring the integral to

$$\prod_{k=1}^n \left(1 + \frac{2p'_k}{m_n} \right)^{-m_n/2} \tilde{Q}_n^{-1} \int_{\mathcal{S}_n^+} \exp \left[-\frac{m_n}{2} \text{tr}(\hat{\boldsymbol{\Sigma}}) \right] \left[\det(\hat{\boldsymbol{\Sigma}}) \right]^{(m_n-n-1)/2} d\hat{\boldsymbol{\Sigma}}$$

The integral here is equal to \tilde{Q}_n , the normalization constant of the probability measure (36.6). If we replace $\{p'_k\}_{k=1}^n$ back by $\{p_k\}_{k=1}^n$, then the above expression is

$$\prod_{k=1}^n \left(1 - \frac{2p_k}{m_n} \right)^{-m_n/2}$$

which coincides with (36.5). Thus the probability law of the Wishart matrices of $\boldsymbol{\Sigma}$ given by (36.6) implies that the corresponding joint probability density of eigenvalues is given by (36.4) for $\boldsymbol{\Sigma}$. \square

Definition 36.2.2 (Laguerre Orthogonal Ensemble)

For the $n \times n$ Laguerre orthogonal ensembles of statistics, the joint probability density function of eigenvalues is for arbitrary parameter $\beta > 0$ and $\alpha > -\frac{2}{\beta}$, is

$$p(\boldsymbol{\lambda}) = K_{\alpha, \beta} \exp \left(-\frac{\beta}{2} \sum_{k=1}^n \lambda_k \right) \prod_{k=1}^n \lambda_k^{\frac{\alpha\beta}{2}} \prod_{i < j} |\lambda_i - \lambda_j|^\beta \quad (36.8)$$

where

$$0 \leq \lambda_1 \leq \dots \leq \lambda_n < \infty$$

and $K_{n, m}$ are normalization constant.

And Equation (36.8) can be written in the standard Boltzmann-Gibbs form, that,

$$p(\boldsymbol{\lambda}) \propto \exp[-\beta E(\boldsymbol{\lambda})]$$

where

$$E(\boldsymbol{\lambda}) = \frac{1}{2} \sum_{k=1}^n (\lambda_k - \alpha \log \lambda_k) - \frac{1}{2} \sum_{i \neq j} |\lambda_i - \lambda_j| \quad (36.9)$$

Remark. For the (36.4), which can be written as (36.8) form, that,

$$p(\boldsymbol{\lambda}) \propto \exp[-\beta m_n E(\boldsymbol{\lambda})]$$

where $\beta = 1$ and

$$E(\boldsymbol{\lambda}) = \frac{m_n}{2} \sum_{k=1}^n \left[\lambda_k - \left(\frac{m_n - n - 1}{m_n} \right) \log \lambda_k \right] - \frac{1}{2m_n} \sum_{i \neq j} |\lambda_i - \lambda_j|$$

36.3 Marčenko-Pastur Theorem

In this section, we will investigate the empirical spectral measure of $\widehat{\Sigma}$, which converges to a nonrandom distribution — Marčenko-Pastur distribution. Before further proof, we will introduce some basic concepts and tools.

36.3.1 Preliminary

Empirical Spectral Measure

Definition 36.3.1 (Empirical Spectral Measure)

For a symmetric matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, the spectral measure or empirical spectral measure or empirical spectral distribution (ESD) $\mu_{\mathbf{M}}$ of \mathbf{M} is defined as the normalized counting measure of the eigenvalues $\lambda_1(\mathbf{M}), \dots, \lambda_n(\mathbf{M})$ of \mathbf{M} , i.e.,

$$\mu_{\mathbf{M}} := \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i(\mathbf{M})} \quad (36.10)$$

where δ_x is a Dirac measure for any (measurable) set, that

$$\delta_x(A) := \mathbf{1}_A(x) = \begin{cases} 0, & x \notin A \\ 1, & x \in A \end{cases}$$

Since $\int \mu_{\mathbf{M}}(dx) = 1$, the spectral measure $\mu_{\mathbf{M}}$ of a matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ (random or not) is a probability measure.

Remark. Many important statistics in multivariate analysis can be expressed as functionals of the ESD, such as, for \mathbf{M} be an $n \times n$ positive definite matrix, then

$$\det(\mathbf{M}) = \prod_{i=1}^n \lambda_i = \exp \left(n \int_0^\infty \log x \mu_{\mathbf{M}}(dx) \right) \quad (36.11)$$

Stieltjes Transform

Definition 36.3.2 (Resolvent)

For a symmetric matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, the resolvent $\mathbf{Q}_{\mathbf{M}}(z)$ of \mathbf{M} is defined as

$$\mathbf{Q}_{\mathbf{M}}(z) := (\mathbf{M} - z\mathbf{I}_n)^{-1} \quad (36.12)$$

where $z \in \mathbb{C}$ not eigenvalue of \mathbf{M} .

Definition 36.3.3 (Stieltjes Transform)

For a real probability measure μ with support $\text{supp}(\mu)$, the Stieltjes transform $m_{\mu}(z)$ is defined as

$$m_{\mu}(z) := \int \frac{1}{t - z} \mu(dt) \quad (36.13)$$

where $z \in \mathbb{C} \setminus \text{supp}(\mu)$.

Property. The Stieltjes transform m_μ has numerous interesting properties:

1. it is complex analytic on its domain of definition $\mathbb{C} \setminus \text{supp}(\mu)$.
2. it is bounded $|m_\mu(z)| \leq 1/\text{dist}(z, \text{supp}(\mu))$.
3. it satisfies $\Im[z] > 0 \Rightarrow \Im[m_\mu(z)] > 0$.
4. it is an increasing function on all connected components of its restriction to $\mathbb{R} \setminus \text{supp}(\mu)$.
5. if $\text{supp}(\mu)$ is bounded, $\lim_{x \rightarrow \pm\infty} m_\mu(x) = 0$.

Remark. Most of the results involve Stieltjes transforms $m_\mu(z)$ of a real probability measure with support $\text{supp}(\mu) \subset \mathbb{R}$. Since Stieltjes transforms are such that

$$m_\mu(z) > 0, \forall z < \inf \text{supp}(\mu), \quad m_\mu(z) < 0, \forall z > \sup \text{supp}(\mu), \quad \Im[z]\Im[m_\mu(z)] > 0, \text{ if } z \in \mathbb{C} \setminus \mathbb{R}$$

it will be convenient in the following to consider the set of scalar pairs

$$\mathcal{Z}(\mathcal{A}) = \{(z, m) \in \mathcal{A} \times \mathbb{C}, (\Im[z]\Im[m] > 0 \text{ if } \Im[z] \neq 0) \text{ or } (m > 0 \text{ if } z < \inf \mathcal{A}^c \cap \mathbb{R}) \\ \text{or } (m < 0 \text{ if } z > \sup \mathcal{A}^c \cap \mathbb{R})\}$$

As a transform, m_μ has an inverse formula to recover μ , as per the following result.

Theorem 36.3.1 (Inverse Stieltjes Transform)

For a, b continuity points of the probability measure μ , we have

$$\mu([a, b]) = \frac{1}{\pi} \lim_{y \downarrow 0} \int_a^b \Im[m_\mu(x + iy)] \, dx \quad (36.14)$$

Specially, if μ has a density f at x , then

$$f(x) = \frac{1}{\pi} \lim_{y \downarrow 0} \Im[m_\mu(x + iy)] \quad (36.15)$$

And, if μ has an isolated mass at x , then

$$\mu(\{x\}) = \lim_{y \downarrow 0} -ym_\mu(x + iy) \quad (36.16)$$

Proof.

$$\begin{aligned} \frac{1}{\pi} \int_a^b \Im[m_\mu(x + iy)] \, dx &= \frac{1}{\pi} \int_a^b \left\{ \int \Im \left[\frac{1}{(t-x) - iy} \right] \mu(dt) \right\} \, dx \\ &= \frac{1}{\pi} \int_a^b \left[\int \frac{y}{(t-x)^2 + y^2} \mu(dt) \right] \, dx \end{aligned}$$

By Fubini's theorem,

$$\begin{aligned} &= \frac{1}{\pi} \int \left[\int_a^b \frac{y}{(t-x)^2 + y^2} \, dx \right] \mu(dt) \\ &= \frac{1}{\pi} \int \left[\arctan \left(\frac{b-t}{y} \right) - \arctan \left(\frac{a-t}{y} \right) \right] \mu(dt) \end{aligned}$$

Since

$$\left| \frac{y}{(t-x)^2 + y^2} \right| \leq \frac{1}{y}, \quad \forall y > 0$$

by the dominated convergence theorem,

$$\frac{1}{\pi} \lim_{y \downarrow 0} \int_a^b \Im [m_\mu(x + iy)] dx = \frac{1}{\pi} \int \lim_{y \downarrow 0} \left[\arctan \left(\frac{b-t}{y} \right) - \arctan \left(\frac{a-t}{y} \right) \right] \mu(dt)$$

as $y \downarrow 0$, the difference in brackets converges either to $\pm\pi$ or 0 depending on the relative position of a, b and t , thus

$$= \int 1_{[a,b]} \mu(dt) = \mu([a, b])$$

Thus, if μ has a density f at x , then

$$f(x) = \frac{1}{\pi} \lim_{y \downarrow 0} \Im [m_\mu(x + iy)]$$

When μ has an isolated mass at x , i.e., $\mu(dt) = a\delta_x(t)$, similarly, since

$$|y(t-x)| \leq \frac{1}{2} (y^2 + (t-x)^2)$$

by dominated convergence theorem,

$$\lim_{y \downarrow 0} -iy m_\mu(x + iy) = -\lim_{y \downarrow 0} \int \frac{iy(t-x)\mu(dt)}{(t-x)^2 + y^2} + \lim_{y \downarrow 0} \int \frac{y^2 \mu(dt)}{(t-x)^2 + y^2} = a$$

□

Remark. The important relation between the empirical spectral measure $\mu_{\mathbf{M}}$ of $\mathbf{M} \in \mathbb{R}^{n \times n}$, the Stieltjes transform $m_{\mu_{\mathbf{M}}}(z)$ and the resolvent $\mathbf{Q}_{\mathbf{M}}(z)$ lies in the fact that

$$m_{\mu_{\mathbf{M}}}(z) = \frac{1}{n} \sum_{i=1}^n \int \frac{\delta_{\lambda_i(\mathbf{M})}(t)}{t-z} = \frac{1}{n} \sum_{i=1}^n \frac{1}{\lambda_i(\mathbf{M}) - z} = \frac{1}{n} \text{tr } \mathbf{Q}_{\mathbf{M}}(z) \quad (36.17)$$

The resolvent $\mathbf{Q}_{\mathbf{M}}$ provides access to scalar observations of the eigenspectrum of \mathbf{M} through its linear functionals. Cauchy's integral formula provides a connection between the linear functionals of the eigenvalues of \mathbf{M} and the Stieltjes transform $m_{\mu_{\mathbf{M}}}(z)$ through

$$\frac{1}{n} \sum_{i=1}^n f(\lambda_i(\mathbf{M})) = -\frac{1}{2\pi i n} \oint_{\Gamma} f(z) \text{tr}(\mathbf{Q}_{\mathbf{M}}(z)) dz = -\frac{1}{2\pi i} \oint_{\Gamma} f(z) m_{\mu_{\mathbf{M}}}(z) dz \quad (36.18)$$

for all f complex analytic in a compact neighborhood of $\text{supp}(\mu_{\mathbf{M}})$, by choosing the contour Γ to enclose $\text{supp}(\mu_{\mathbf{M}})$ (i.e., all the eigenvalues $\lambda_i(\mathbf{M})$).

Matrix Equivalents

Definition 36.3.4 (Deterministic Equivalent)

$\overline{\mathbf{Q}} \in \mathbb{R}^{n \times n}$ is said to be deterministic equivalent for the symmetric random matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$, if for a (sequences of) deterministic matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ of unit norms (operator and Euclidean, respectively),

$$\frac{1}{n} \text{tr } \mathbf{A}(\mathbf{Q} - \overline{\mathbf{Q}}) \rightarrow 0, \quad \mathbf{a}'(\mathbf{Q} - \overline{\mathbf{Q}})\mathbf{b} \rightarrow 0, \quad \text{as } n \rightarrow \infty \quad (36.19)$$

where the convergence is either in probability or almost sure.

Remark. A practical use of deterministic equivalents is to establish that, for a random matrix \mathbf{M} of interest, suppose

$$\frac{1}{n} \operatorname{tr} (\mathbf{Q}_{\mathbf{M}}(z) - \overline{\mathbf{Q}}(z)) \rightarrow 0, \quad \text{a.s.,} \quad \forall z \in \mathcal{C}, \mathcal{C} \subset \mathbb{C}$$

this convergence implies that the Stieltjes transform of $\mu_{\mathbf{M}}$ "converges" in the sense that

$$m_{\mu_{\mathbf{M}}}(z) - \bar{m}_n(z) \rightarrow 0$$

where $\bar{m}_n(z) = \frac{1}{n} \operatorname{tr} \overline{\mathbf{Q}}(z)$.

Definition 36.3.5 (Matrix Equivalents)

For $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times n}$ two random or deterministic matrices, we write

$$\mathbf{X} \leftrightarrow \mathbf{Y} \quad (36.20)$$

if, for all $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ of unit norms (respectively, operator and Euclidean), we have the simultaneous results

$$\frac{1}{n} \operatorname{tr} \mathbf{A}(\mathbf{X} - \mathbf{Y}) \rightarrow 0, \quad \mathbf{a}'(\mathbf{X} - \mathbf{Y})\mathbf{b} \rightarrow 0, \quad \|\mathbb{E}[\mathbf{X} - \mathbf{Y}]\| \rightarrow 0$$

where, for random quantities, the convergence is either in probability or almost sure.

Resolvent and Perturbation Identities

Lemma 36.3.1 (Resolvent Identity)

For invertible matrices \mathbf{A} and \mathbf{B} , we have

$$\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1} \quad (36.21)$$

Lemma 36.3.2 (Sherman-Morrison)

For $\mathbf{A} \in \mathbb{R}^{n \times n}$ invertible and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, then $\mathbf{A} + \mathbf{u}\mathbf{v}'$ is invertible if and only if $1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u} \neq 0$ and

$$(\mathbf{A} + \mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}}{1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}} \quad (36.22)$$

or,

$$(\mathbf{A} + \mathbf{u}\mathbf{v}')^{-1}\mathbf{u} = \frac{\mathbf{A}^{-1}\mathbf{u}}{1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}} \quad (36.23)$$

Lemma 36.3.3 (Quadratic-form-close-to-the-trace)

Let $\mathbf{x} \in \mathbb{R}^p$ have i.i.d. entries of zero mean, unit variance and $\mathbb{E}[|x_i|^K] \leq \nu_K$ for some $K \geq 1$. Then for $\mathbf{A} \in \mathbb{R}^{p \times p}$ and $k \geq 1$

$$\mathbb{E}[|\mathbf{x}'\mathbf{A}\mathbf{x} - \text{tr } \mathbf{A}|^k] \leq C_k \left[(\nu_4 \text{tr}(\mathbf{A}\mathbf{A}'))^{k/2} + \nu_{2k} \text{tr}(\mathbf{A}\mathbf{A}')^{k/2} \right]$$

for some constant $C_k > 0$ independent of p . In particular, if $\|\mathbf{A}\| \leq 1$ and the entries of \mathbf{x} have bounded eighth-order moment,

$$\mathbb{E}[(\mathbf{x}'\mathbf{A}\mathbf{x} - \text{tr } \mathbf{A})^4] \leq Cp^2$$

for some $C > 0$ independent of p , and consequently, as $p \rightarrow \infty$,

$$\frac{1}{p}\mathbf{x}'\mathbf{A}\mathbf{x} - \frac{1}{p}\text{tr } \mathbf{A} \xrightarrow{\text{a.s.}} 0$$

36.3.2 Marčenko-Pastur Theorem

With the above tools, we can prove the Marčenko-Pastur Theorem. Here, we only suppose \mathbf{X} having some smooth tail condition.

Theorem 36.3.2 (Marčenko-Pastur Theorem)

Consider the resolvent

$$\mathbf{Q}(z) = (\hat{\Sigma} - z\mathbf{I}_n)^{-1}$$

Then, if

$$\frac{n}{m_n} \rightarrow \rho \text{ with } \rho \in (0, \infty), \quad \text{as } n \rightarrow \infty$$

we have

$$\mathbf{Q}(z) \leftrightarrow \overline{\mathbf{Q}}(z), \quad \overline{\mathbf{Q}}(z) = m(z)\mathbf{I}_n$$

with $(z, m(z))$ the unique solution in $\mathcal{Z}(\mathbb{C} \setminus [(1 - \sqrt{\rho})^2, (1 + \sqrt{\rho})^2])$ be

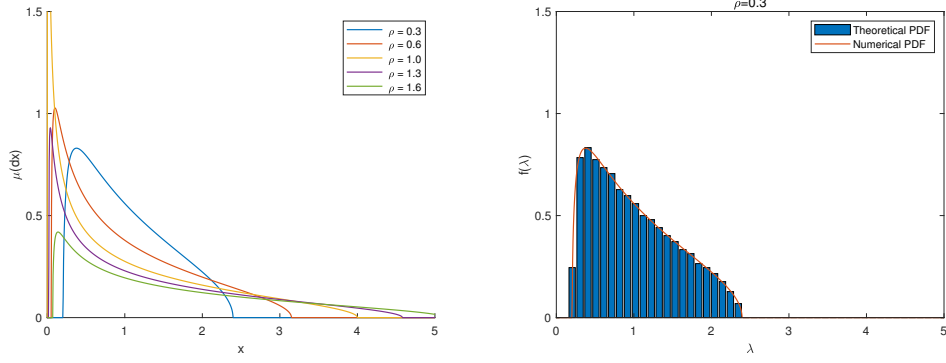
$$zcm^2(z) - (1 - c - z)m(z) + 1 = 0$$

where the function $m(z)$ is the Stieltjes transform of the probability measure μ given explicitly by

$$\mu(dx) = (1 - \rho^{-1})^+ \delta_0(x) + \frac{\sqrt{(x - a_-)^+ (a_+ - x)^+}}{2\pi\rho x} dx$$

where $a_{\pm} = (1 \pm \sqrt{\rho})^2$ and $(x)^+ = \max(0, x)$, and is known as the Marčenko-Pastur distribution. In particular, with probability one, the empirical spectral measure $\mu_{\hat{\Sigma}}$ converges weakly to μ .

Proof. (Intuitive Proof) Suppose $\overline{\mathbf{Q}}(z) = \mathbf{F}(z)^{-1}$ for some matrix $\mathbf{F}(z)$. To prove $\overline{\mathbf{Q}}(z)$ to be a



(a) The Marčenko-Pastur Distribution for $\rho = 0.3, 0.6, 1, 1.3, 1.6$ (b) Simulation Results of the Marčenko-Pastur Theorem When $\rho = 0.3$

Figure 36.1: Illustrations of the Marčenko-Pastur Theorem

deterministic equivalent for $\mathbf{Q}(z)$, particularly,

$$\frac{1}{n} \text{tr} \mathbf{A}(\mathbf{Q}(z) - \overline{\mathbf{Q}}(z)) \rightarrow 0 \quad \text{a.s.}$$

where \mathbf{A} is arbitrary, deterministic, and such that $\|\mathbf{A}\| = 1$. By Lemma 36.3.1, we have

$$\begin{aligned} \mathbf{Q}(z) - \overline{\mathbf{Q}}(z) &= \mathbf{Q}(z) \left(\mathbf{F}(z) + z\mathbf{I}_n - \widehat{\Sigma} \right) \overline{\mathbf{Q}}(z) \\ &= \mathbf{Q}(z) \left(\mathbf{F}(z) + z\mathbf{I}_n - \frac{1}{m_n} \sum_{i=1}^{m_n} \mathbf{X}_i \mathbf{X}_i' \right) \overline{\mathbf{Q}}(z) \end{aligned}$$

Thus, we turn to prove that,

$$\frac{1}{n} \text{tr} \left[(\mathbf{F}(z) + z\mathbf{I}_n) \overline{\mathbf{Q}}(z) \mathbf{A} \mathbf{Q}(z) \right] - \frac{1}{n} \cdot \frac{1}{m_n} \sum_{i=1}^{m_n} \mathbf{X}_i' \overline{\mathbf{Q}}(z) \mathbf{A} \mathbf{Q}(z) \mathbf{X}_i \rightarrow 0 \quad \text{a.s.}$$

By Lemma 36.3.2, we have

$$\mathbf{Q}(z) \mathbf{X}_i = \frac{\mathbf{Q}_{-i}(z) \mathbf{X}_i}{1 + \frac{1}{m_n} \mathbf{X}_i' \mathbf{Q}_{-i}(z) \mathbf{X}_i}$$

where

$$\mathbf{Q}_{-i}(z) = \left(\frac{1}{m_n} \sum_{j \neq i} \mathbf{X}_j \mathbf{X}_j' - z\mathbf{I}_n \right)^{-1}$$

is independent of \mathbf{X}_i . By Lemma 36.3.3, we have

$$\frac{1}{n} \mathbf{X}_i' \overline{\mathbf{Q}}(z) \mathbf{A} \mathbf{Q}(z) \mathbf{X}_i = \frac{\frac{1}{n} \mathbf{X}_i' \overline{\mathbf{Q}}(z) \mathbf{A} \mathbf{Q}_{-i}(z) \mathbf{X}_i}{1 + \frac{1}{m_n} \mathbf{X}_i' \mathbf{Q}_{-i}(z) \mathbf{X}_i} \simeq \frac{\frac{1}{n} \text{tr} [\overline{\mathbf{Q}}(z) \mathbf{A} \mathbf{Q}_{-i}(z)]}{1 + \frac{1}{m_n} \text{tr} [\mathbf{Q}_{-i}(z)]}$$

Hence, we need to prove the approximation that

$$\frac{1}{n} \text{tr} \left[(\mathbf{F}(z) + z\mathbf{I}_n) \overline{\mathbf{Q}}(z) \mathbf{A} \mathbf{Q}(z) \right] \simeq \frac{\frac{1}{n} \text{tr} [\overline{\mathbf{Q}}(z) \mathbf{A} \mathbf{Q}(z)]}{1 + \frac{1}{m_n} \text{tr} [\mathbf{Q}(z)]}$$

If $\mathbf{F}(z)$ exist, for the approximation above to hold, $\mathbf{F}(z)$ must be of the type

$$\mathbf{F}(z) \simeq \left(-z + \frac{1}{1 + \frac{1}{m_n} \operatorname{tr} \mathbf{Q}(z)} \right) \mathbf{I}_n$$

By Equation 36.17, we have,

$$m(z) \equiv \frac{1}{n} \operatorname{tr} [\overline{\mathbf{Q}}(z)] = \frac{1}{n} \operatorname{tr} [\mathbf{F}(z)^{-1}]$$

taking $\mathbf{A} = \mathbf{I}_n$, we have

$$\frac{1}{n} \operatorname{tr} [\mathbf{Q}(z)] \simeq \frac{1}{n} \operatorname{tr} [\overline{\mathbf{Q}}(z)] = m(z) = \frac{1}{-z + \frac{1}{1 + \frac{n}{m_n} \frac{1}{n} \operatorname{tr} [\mathbf{Q}(z)]}} \simeq \frac{1}{-z + \frac{1}{1 + \rho m(z)}}$$

As $n, m_n \rightarrow \infty$, $m(z)$ is solution to

$$m(z) = \frac{1}{-z + \frac{1}{1 + \rho m(z)}}$$

or equivalently

$$z \rho m^2(z) - (1 - \rho - z)m(z) + 1 = 0$$

This equation has two solutions defined via the two values of the complex square root function. Let

$$z = re^{i\theta} \text{ where } r \geq 0, \theta \in [0, 2\pi) \Rightarrow \sqrt{z} \in \left\{ \pm \sqrt{r} e^{i\theta/2} \right\}$$

and we can conclude that

$$m(z) = \frac{1 - \rho - z}{2\rho z} + \frac{\sqrt{((1 + \sqrt{\rho})^2 - z)((1 - \sqrt{\rho})^2 - z)}}{2\rho z}$$

only one of which is such that $\Im[z]\Im[m(z)] > 0$ as imposed by the definition of Stieltjes transforms. By the inverse Stieltjes transform theorem, Theorem 36.3.1, we find that $m(z)$ is the Stieltjes transform of the measure μ with

$$\mu([a, b]) = \frac{1}{\pi} \lim_{\epsilon \downarrow 0} \int_a^b \Im[m(x + i\epsilon)] dx$$

for all continuity points $a, b \in \mathbb{R}$ of μ . This term under the square root in $m(z)$ being negative only in the set

$$[(1 - \sqrt{\rho})^2, (1 + \sqrt{\rho})^2]$$

(and thus of non-real square root), the latter defines the support of the continuous part of the measure μ with density

$$\frac{\sqrt{((1 + \sqrt{\rho})^2 - x)(x - (1 - \sqrt{\rho})^2)}}{2\rho\pi x}$$

at point x in the set. The case $x = 0$ brings a discontinuity in μ with weight equal to

$$\mu(\{0\}) = -\lim_{y \downarrow 0} \Im y m(iy) = \frac{\rho - 1}{2\rho} \pm \frac{\rho - 1}{2\rho}$$

where the sign is established by a second order development of $zm(z)$ in the neighborhood of zero: that is, "+" for $c > 1$ inducing a mass $1 - 1/\rho$ for $p > n$, or "-" for $c < 1$ in which case $\mu(\{0\}) = 0$ and μ has no mass at zero. \square

Remark. The asymptotic phenomenon holds not only in the Gaussian case, which also holds

1. if $(x_i)_{1 \leq i \leq n}$ are i.i.d. with finite second moment.
2. if \mathbf{X} is isotropic and log-concave¹ random vector.

36.4 Limits of Extreme Eigenvalues

The weak convergence in Theorem 36.3.2 does not provide much information at the edge on the behavior of the extremal atoms, and what one can actually extract is that

$$\limsup_{n \rightarrow \infty} \lambda_{\min}(\hat{\Sigma}) \leq (1 - \sqrt{\rho})^2 \leq (1 + \sqrt{\rho})^2 \leq \liminf_{n \rightarrow \infty} \lambda_{\max}(\hat{\Sigma}) \quad \text{a.s.} \quad (36.24)$$

where the first inequality is considered only in the case where $m_n \geq n$.

The weak convergence above does not provide much information at the edge on the behavior of the extremal atoms. Now, we have more exact result, that if $(X_{n,k})_{n \geq 1, 1 \leq k \leq n}$ are i.i.d. with finite fourth moment then,

$$(1 - \sqrt{\rho})^2 = \lim_{n \rightarrow \infty} \lambda_{\min}(\hat{\Sigma}) \leq \lim_{n \rightarrow \infty} \lambda_{\max}(\hat{\Sigma}) = (1 + \sqrt{\rho})^2 \quad \text{a.s.} \quad (36.25)$$

where the first inequality is considered only in the case where $m_n \geq n$.

Remark. The convergence of the smallest eigenvalue in the left hand side of (36.25) holds if $(x_i)_{1 \leq i \leq n}$ are i.i.d. with finite second moment.

Theorem 36.4.1

If $\bar{\rho} < 1$ (in particular $m_n > n$ for $n \gg 1$) and if the centered isotropic random vector \mathbf{X} is log-concave or if $(x_i)_{1 \leq i \leq n}$ are i.i.d. then

$$\liminf_{n \rightarrow \infty} \frac{E(\lambda_{\min}(\mathbf{A}_n))}{(\sqrt{m_n} - \sqrt{n})^2} \geq 1 \quad (36.26)$$

If additionally $\lim_{n \rightarrow \infty} \frac{n}{m_n} = \rho$ with $\rho \in (0, 1)$, in other words $\underline{\rho} = \bar{\rho} \in (0, 1)$, then

$$\lambda_{\min}(\hat{\Sigma}_n) \xrightarrow{p} (1 - \sqrt{\rho})^2 \quad \text{as } n \rightarrow \infty \quad (36.27)$$

Proof.

□

Theorem 36.4.2

If the centered isotropic random vector \mathbf{X} is log-concave or if $(x_i)_{1 \leq i \leq n}$ are i.i.d. with finite 4-th moment then

$$\limsup_{n \rightarrow \infty} \frac{E(\lambda_{\max}(\mathbf{A}_n))}{(\sqrt{m_n} + \sqrt{n})^2} \leq 1 \quad (36.28)$$

If additionally $\lim_{n \rightarrow \infty} \frac{n}{m_n} = \rho$ with $\rho \in (0, 1)$, in other words $\underline{\rho} = \bar{\rho} \in (0, 1)$, then

$$\lambda_{\max}(\hat{\Sigma}_n) \xrightarrow{p} (1 + \sqrt{\rho})^2 \quad \text{as } n \rightarrow \infty \quad (36.29)$$

Proof.

□

¹A probability measure μ on \mathbb{R}^n with density φ is log-concave when $\varphi = e^{-V}$ with V convex.

Part XIII

Applications

Chapter 37

Missingness Data

37.1 The Problem of Missing Data

We concern the problem the analysis of such a data matrix when some of the entries in the matrix are not observed (Figure 37.1).

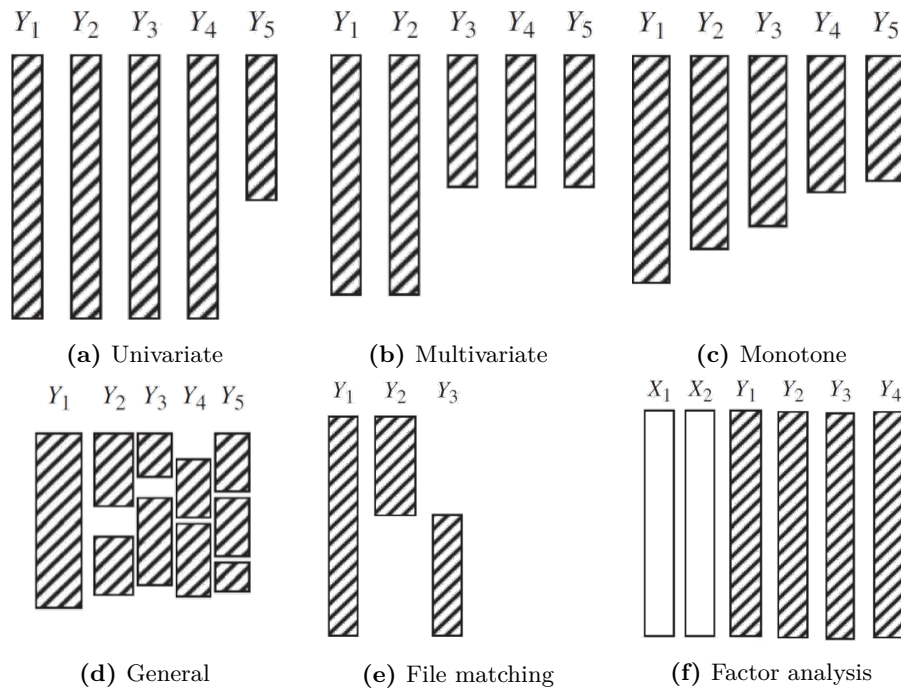


Figure 37.1: Examples of missingness patterns

Notations for missing data are as follows

- $Y = (y_{ij})$ denote the $(n \times p)$ rectangular data matrix, where only a portion of Y are observed and $y_{ij} = \star$ indicates this entry is missing;
- $M = (m_{ij})$ denote the *missingness indicator matrix* for y_{ij} , taking $m_{ij} = 0$ for y_{ij} is observed, and $m_{ij} = 1$ for y_{ij} is missing.

- In order to simplify, let $y_i = (y_{i1}, y_{i2}, \dots, y_{ip})$, $m_i = (m_{i1}, m_{i2}, \dots, m_{ip})$ and $y_{(0)i}$ be the components of y_i that are observed for unit i , $y_{(1)i}$ be the components of y_i that are missing for unit i .

37.1.1 Missingness Mechanisms

The missingness mechanism is characterized by the conditional distribution of m_i given y_i , say

$$f_{M|Y}(m_i | y_i, \phi), \quad (37.1)$$

where ϕ denotes unknown parameters.

Definition 37.1.1 (Missing Completely at Random, MCAR)

If missingness does not depend on the value of the data, missing or observed, i.e., if for all y_i and any distinct values y^* in the sample space of Y ,

$$f_{M|Y}(m_i | y_i, \phi) = f_{M|Y}(m_i | y^*, \phi), \quad (37.2)$$

then the data are called missing completely at random, MCAR.

Definition 37.1.2 (Missing at Random, MAR)

If missingness depends on y_i only through the observed components $y_{(0)i}$, i.e., if for all y_i and any distinct values $y_{(1)}^*$ in the sample space of $y_{(1)}$,

$$f_{M|Y}(m_i | y_{(0)i}, y_{(1)i}, \phi) = f_{M|Y}(m_i | y_{(0)i}, y_{(1)}^*, \phi), \quad (37.3)$$

then the data are called missing at random, MAR.

Definition 37.1.3 (Missing Not at Random, MNAR)

If missingness depends on y_i the missing components $y_{(1)i}$, i.e., if some y_i and some values $y_{(1)}^*$ in the sample space of $y_{(1)}$,

$$f_{M|Y}(m_i | y_{(0)i}, y_{(1)i}, \phi) \neq f_{M|Y}(m_i | y_{(0)i}, y_{(1)}^*, \phi), \quad (37.4)$$

then the data are called missing not at random, MNAR.

37.1.2 Commonly Used Methods for Missing Data

1. Complete-case Analysis: discard incompletely recorded units, only use the units with the complete data.
2. Weighting Procedures: randomization inferences from sample survey data without nonresponse commonly weight sampled units by their design weights.
3. Imputation Methods: impute the missing values, and the resultant completed data are analyzed by standard methods.
4. **Model-based Methods:** A broad class of procedures is generated by defining a model for the complete data and basing inferences on the likelihood or posterior distribution under that model, with parameters estimated by procedures such as maximum likelihood.

5. Hybrid Approaches: approaches based on estimating equations have been proposed that combine the aspects of modeling and weighting.

37.2 Likelihood-Based Inference with Missing Data

We can model the density of the joint distribution of Y and M using the "selection model" factorization

$$p(Y = y, M = m \mid \theta, \psi) = f_Y(y \mid \theta) f_{M|Y}(m \mid y, \psi),$$

where θ is the parameter vector governing the data model, and ψ is the parameter vector governing the model for the missingness mechanism.

The full likelihood based on the observed values $(y_{(0)}, m)$ and the assumed joint distribution model above is defined to be

$$L_{\text{full}}(\theta, \psi \mid y_{(0)}, m) = \int f_Y(y_{(0)}, y_{(1)} \mid \theta) f_{M|Y}(m \mid y_{(0)}, y_{(1)}, \psi) dy_{(1)} \quad (37.5)$$

The likelihood of θ ignoring the missingness mechanism is defined to be

$$L_{\text{ign}}(\theta \mid y_{(0)}) = \int f_Y(y_{(0)}, y_{(1)} \mid \theta) dy_{(1)} \quad (37.6)$$

37.2.1 Ignorable Missingness Mechanism

Definition 37.2.1 (Ignorable missingness mechanism)

The missingness mechanism is called ignorable if for any given \tilde{m} and $\tilde{y}_{(0)}$ the inferences for θ based on the ignorable likelihood equation evaluated at $m = \tilde{m}$ and \tilde{y}_0 are the same as the full likelihood equation.

Remark (Another definition of ignorable missingness mechanism).

$$\frac{L_{\text{full}}(\theta, \psi \mid \tilde{y}_{(0)}, \tilde{m})}{L_{\text{full}}(\theta^*, \psi \mid \tilde{y}_{(0)}, \tilde{m})} = \frac{L_{\text{ign}}(\theta \mid \tilde{y}_{(0)})}{L_{\text{ign}}(\theta^* \mid \tilde{y}_{(0)})} \quad \forall \theta, \theta^*, \psi. \quad (37.7)$$

Theorem 37.2.1

The missingness mechanism is ignorable for direct likelihood inference on $(\tilde{m}, \tilde{y}_{(0)})$ if

1. Parameter distinctness: The parameters θ and ψ are distinct, that is, $\Omega_{\theta, \psi} = \Omega_{\theta} \times \Omega_{\psi}$.
2. Factorization of the full likelihood: The full likelihood, with $(y_0, m) = (\tilde{y}_0, \tilde{m})$ factors as

$$L_{\text{full}}(\theta, \psi \mid \tilde{y}_{(0)}, \tilde{m}) = L_{\text{ign}}(\theta \mid \tilde{y}_{(0)}) \times L_{\text{rest}}(\psi \mid \tilde{y}_{(0)}, \tilde{m}) \quad \forall \theta, \psi \in \Omega_{\theta, \psi} \quad (37.8)$$

Corollary 37.2.1

If the missing data are MAR at $(\tilde{m}, \tilde{y}_{(0)})$, and θ and ψ are distinct, the missingness mechanism is ignorable for likelihood inference.

Proof. Since,

$$f_{M|Y}(\tilde{m} | \tilde{y}_{(0)}, y_{(1)}, \psi) = f_{M|Y}(\tilde{m} | \tilde{y}_{(0)}, y_{(1)}^*, \psi) \quad \forall y_{(1)}, y_{(1)}^*, \psi \quad (37.9)$$

therefore,

$$\begin{aligned} f(\tilde{y}_{(0)}, \tilde{m} | \theta, \psi) &= f_{M|Y}(\tilde{m} | \tilde{y}_{(0)}, \psi) \times \int f_Y(\tilde{y}_{(0)}, y_{(1)} | \theta) dy_{(1)} \\ &= f_{M|Y}(\tilde{m} | \tilde{y}_{(0)}, \psi) \times f_Y(\tilde{y}_{(0)} | \theta) \end{aligned} \quad (37.10)$$

yields the factored likelihood equation 37.8. \square

Ignorable Missingness Mechanism v.s. Nonignorable Missingness Mechanism

Example (Exponential Sample). The joint density of n independent and identically distributed scalar units from the exponential distribution with mean $\theta > 0$ is

$$f_Y(y | \theta) = \theta^{-n} \exp \left\{ - \sum_{i=1}^n \frac{y_i}{\theta} \right\}. \quad (37.11)$$

The loglikelihood function is

$$\ell_Y(\theta | y) = \ln \left\{ \theta^{-n} \exp \left(- \sum_{i=1}^n \frac{y_i}{\theta} \right) \right\} = -n \ln \theta - \sum_{i=1}^n \frac{y_i}{\theta}. \quad (37.12)$$

Differentiating with respect to θ gives the likelihood equation

$$-\frac{n}{\theta} + \sum_{i=1}^n \frac{y_i}{\theta^2} = 0. \quad (37.13)$$

Thus, we obtain the ML estimates

$$\hat{\theta} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i. \quad (37.14)$$

Example (Incomplete Exponential Sample). Suppose we have an incomplete univariate exponential sample with $y_{(0)} = (y_1, \dots, y_r)^T$ observed and $y_{(1)} = (y_{r+1}, \dots, y_n)^T$ missing. Thus, $m = (m_1, \dots, m_n)^T$, where $m_i = 0, i = 1, \dots, r$ and $m_i = 1, i = r+1, \dots, n$.

The likelihood ignoring the missingness mechanism is

$$L_{\text{ign}}(\theta | y_{(0)}) = \theta^{-r} \exp \left(- \sum_{i=1}^r \frac{y_i}{\theta} \right). \quad (37.15)$$

If each unit is observed with probability ψ that does not depend on Y , that is,

$$f_{M|Y}(m | y, \psi) = \frac{n!}{r!(n-r)!} \psi^r (1-\psi)^{n-r} \quad (37.16)$$

then,

$$f(y_{(0)}, m | \theta, \psi) = \frac{n!}{r!(n-r)!} \psi^r (1-\psi)^{n-r} \theta^{-r} \exp \left(- \sum_{i=1}^r \frac{y_i}{\theta} \right) \quad (37.17)$$

Because the missing data are MAR, if ψ and θ are distinct, then likelihood-based inferences about θ can be based on the ignorable likelihood, the ML estimate of θ is

$$\hat{\theta} = \frac{1}{r} \sum_{i=1}^r y_i. \quad (37.18)$$

If each unit is observed only if values less than c , that is

$$f_{M|Y}(m | y, \psi) = \prod_{i=1}^n f(m_i | y_i, \psi), \quad (37.19)$$

where

$$f(m_i | y_i, \psi) = \begin{cases} 1, & y_i \geq c \\ 0, & \text{otherwise} \end{cases} \quad (37.20)$$

Hence,

$$\begin{aligned} L_{\text{full}}(\theta | y_{(0)}, m) &= \prod_{i=1}^r f_Y(y_i | \theta) \Pr(y_i < c | y_i, \theta) \times \prod_{i=r+1}^n \Pr(y_i \geq c | \theta) \\ &= \theta^{-r} \exp\left(-\sum_{i=1}^r \frac{y_i}{\theta}\right) \times \exp\left(-\frac{(n-r)c}{\theta}\right) \end{aligned} \quad (37.21)$$

Maximizing above equation with respect to θ gives the ML estimate

$$\hat{\theta} = \frac{1}{r} \left[\sum_{i=1}^r y_i + (n-r)c \right]. \quad (37.22)$$

The inflation of the sample mean in this expression reflects the censoring of the missing values.

37.2.2 Expectation-Maximization Algorithm

Let $\theta^{(i)}$ be the current estimate of the parameter θ . The E step of EM finds the expected complete-data loglikelihood if θ were $\theta^{(t)}$:

$$Q(\theta | \theta^{(t)}) = \int \ell(\theta | Y_{(0)}, Y_{(1)}) f(Y_{(1)} | Y_{(0)}, \theta^{(t)}) dY_{(1)}. \quad (37.23)$$

The M step of EM determines $\theta^{(t+1)}$ by maximizing this expected completedata loglikelihood:

$$Q(\theta^{(t+1)} | \theta^{(t)}) \geq Q(\theta | \theta^{(t)}), \quad \forall \theta. \quad (37.24)$$

Hence, EM algorithm for likelihood-based inference with missing data is

1. replace missing values by estimated values
2. estimate parameters
3. re-estimate the missing values assuming the new parameter estimates are correct
4. re-estimate parameters, and so forth, iterating until apparent convergence

Convergence Properties of EM Algorithm with Missing Data**Theorem 37.2.2**

Every GEM algorithm increases $\ell(\theta | Y_{(0)})$ at each iteration, that is,

$$\ell(\theta^{(t+1)} | Y_{(0)}) \geq \ell(\theta^{(t)} | Y_{(0)}) \quad (37.25)$$

, with equality if and only if

$$Q(\theta^{(t+1)} | \theta^{(t)}) = Q(\theta^{(t)} | \theta^{(t)}) \quad (37.26)$$

Proof. The distribution of the complete data Y can be factored as follows:

$$f(Y | \theta) = f(Y_{(0)}, Y_{(1)} | \theta) = f(Y_{(0)} | \theta) f(Y_{(1)} | Y_{(0)}, \theta) \quad (37.27)$$

The corresponding decomposition of the loglikelihood is

$$\ell(\theta | Y) = \ell(\theta | Y_{(0)}, Y_{(1)}) = \ell(\theta | Y_{(0)}) + \ln f(Y_{(1)} | Y_{(0)}, \theta) \quad (37.28)$$

Let,

$$\ell(\theta | Y_{(0)}) = \ell(\theta | Y) - \ln f(Y_{(1)} | Y_{(0)}, \theta) \quad (37.29)$$

The expectation of both sides of above equation over the distribution of the missing data $Y_{(1)}$, given the observed data $Y_{(0)}$ and a current estimate of θ , say $\theta^{(t)}$, is

$$\ell(\theta | Y_{(0)}) = Q(\theta | \theta^{(t)}) - H(\theta | \theta^{(t)}) \quad (37.30)$$

, where

$$\begin{aligned} Q(\theta | \theta^{(t)}) &= \int \ell(\theta | Y_{(0)}, Y_{(1)}) f(Y_{(1)} | Y_{(0)}, \theta^{(t)}) dY_{(1)} \\ H(\theta | \theta^{(t)}) &= \int \ln f(Y_{(1)} | Y_{(0)}, \theta) f(Y_{(1)} | Y_{(0)}, \theta^{(t)}) dY_{(1)} \end{aligned} \quad (37.31)$$

Since,

$$\begin{aligned} & H(\theta^{(t)}, \theta^{(t)}) - H(\theta, \theta^{(t)}) \\ &= \int \ln f(Y_{(1)} | Y_{(0)}, \theta^{(t)}) f(Y_{(1)} | Y_{(0)}, \theta^{(t)}) dY_{(1)} \\ &\quad - \int \ln f(Y_{(1)} | Y_{(0)}, \theta) f(Y_{(1)} | Y_{(0)}, \theta^{(t)}) dY_{(1)} \\ &= \int \ln \left[\frac{f(Y_{(1)} | Y_{(0)}, \theta^{(t)})}{f(Y_{(1)} | Y_{(0)}, \theta)} \right] f(Y_{(1)} | Y_{(0)}, \theta^{(t)}) dY_{(1)} \\ &= \int -\ln \left[\frac{f(Y_{(1)} | Y_{(0)}, \theta)}{f(Y_{(1)} | Y_{(0)}, \theta^{(t)})} \right] f(Y_{(1)} | Y_{(0)}, \theta^{(t)}) dY_{(1)} \\ &\geq -\ln \int \frac{f(Y_{(1)} | Y_{(0)}, \theta)}{f(Y_{(1)} | Y_{(0)}, \theta^{(t)})} f(Y_{(1)} | Y_{(0)}, \theta^{(t)}) dY_{(1)} = 0 \end{aligned} \quad (37.32)$$

Therefore,

$$H(\theta | \theta^{(t)}) \leq H(\theta^{(t)} | \theta^{(t)}) \quad (37.33)$$

Hence, the difference in values of $\ell(\theta | Y_{(0)})$ at successive iterates is given by

$$\begin{aligned} \ell(\theta^{(t+1)} | Y_{(0)}) - \ell(\theta^{(t)} | Y_{(0)}) &= \left[Q(\theta^{(t+1)} | \theta^{(t)}) - Q(\theta^{(t)} | \theta^{(t)}) \right] \\ &\quad - \left[H(\theta^{(t+1)} | \theta^{(t)}) - H(\theta^{(t)} | \theta^{(t)}) \right] \\ &\geq 0 \end{aligned} \quad (37.34)$$

□

Theorem 37.2.3

Suppose a sequence of EM iterates is such that

1. $D^{10}Q(\theta^{(t+1)} | \theta^{(t)}) = 0$, where "D" here denotes derivative, and D^{10} means the derivative with respect to the first argument, that is, define

$$D^{10}Q(\theta^{(t+1)} | \theta^{(t)}) = \left. \frac{\partial}{\partial \theta} Q(\theta | \theta^{(t)}) \right|_{\theta=\theta^{(t+1)}} = 0. \quad (37.35)$$

2. $\theta^{(t)}$ converges to θ^* .

3. $f(Y_{(1)} | Y_{(0)}, \theta)$ is smooth in θ , where smooth is defined in the proof.

Then

$$D\ell(\theta^* | Y_{(0)}) \equiv \left. \frac{\partial}{\partial \theta} \ell(\theta | Y_{(0)}) \right|_{\theta=\theta^*} = 0, \quad (37.36)$$

so that if the $\theta^{(t)}$ converge, they converge to a stationary point.

Proof.

$$\begin{aligned} D\ell(\theta^{(t+1)} | Y_{(0)}) &= D^{10}Q(\theta^{(t+1)} | \theta^{(t)}) - D^{10}H(\theta^{(t+1)} | \theta^{(t)}) \\ &= -D^{10}H(\theta^{(t+1)} | \theta^{(t)}) \\ &= -\left. \frac{\partial}{\partial \theta} \int [\ln f(Y_{(1)} | Y_{(0)}, \theta)] f(Y_{(1)} | Y_{(0)}, \theta^{(t)}) dY_{(1)} \right|_{\theta=\theta^{(t+1)}} \end{aligned} \quad (37.37)$$

which assuming sufficient smoothness to interchange the order of differentiation and integration,

$$\begin{aligned} &= -\left. \int \frac{\partial}{\partial \theta} f(Y_{(1)} | Y_{(0)}, \theta) dY_{(1)} \right|_{\theta=\theta^{(t+1)}} \\ &= -\int \frac{\partial}{\partial \theta} f(Y_{(1)} | Y_{(0)}, \theta) dY_{(1)} \Big|_{\theta=\theta^{(t+1)}} = 0 \end{aligned} \quad (37.38)$$

□

Examples of EM Algorithm with Missing Data

Example (Multivariate Normal Sample). Let $y = (y_{ij})$, where $i = 1, \dots, n$, $j = 1, \dots, p$, be a matrix representing an independent and identically distributed sample of n units from the multivariate normal distribution with mean vector $\mu = (\mu_1, \dots, \mu_p)$ and covariance matrix $\Sigma = (\sigma_{jk})$, $j = 1, \dots, p$; $k = 1, \dots, p$. Thus, y_{ij} represents the value of the j th variable for the i th unit in the sample. The density of y is

$$f_Y(y | \mu, \Sigma) = (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (y_i - \mu) \Sigma^{-1} (y_i - \mu)^T \right\}. \quad (37.39)$$

The loglikelihood of $\theta = (\mu, \Sigma)$ is then

$$\ell_Y(\mu, \Sigma | y) = -\frac{n}{2} \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \quad (37.40)$$

Maximizing above equation with respect to θ and Σ gives the ML estimate

$$\hat{\mu} = \bar{y}, \quad \hat{\Sigma} = \frac{n-1}{n} S, \quad (37.41)$$

where $\bar{y} = (\bar{y}_1, \dots, \bar{y}_p)$ is the row vector of sample means, and $S = (s_{jk})$ is the $(p \times p)$ sample covariance matrix with (j, k) th element $s_{jk} = \frac{1}{n-1} \sum_{i=1}^n (y_{ij} - \bar{y}_i)(y_{ik} - \bar{y}_k)$

Example (Incomplete Multivariate Normal Sample). Suppose $Y = (Y_{(0)}, Y_{(1)})$, where Y represents a random sample of size n on (Y_1, \dots, Y_p) , $Y_{(0)}$ the set of observed values, and $Y_{(1)}$ the missing data. Also, let $y_{(0),i}$ represent the set of variables with values observed for unit i , $i = 1, \dots, n$.

The loglikelihood based on the observed data is then

$$\ell(\mu, \Sigma | Y_{(0)}) = -\frac{1}{2} \sum_{i=1}^n \ln |\Sigma_{(0),i}| - \frac{1}{2} \sum_{i=1}^n (y_{(0),i} - \mu_{(0),i})^T \Sigma_{(0),i}^{-1} (y_{(0),i} - \mu_{(0),i}) \quad (37.42)$$

, where $\mu_{(0),i}$ and $\Sigma_{(0),i}$ are the mean and covariance matrix of the observed components of Y for unit i .

The exponential family form of multivariate normal distribution with (μ, Σ) is

$$f_Y(y | \mu, \Sigma) = (2\pi)^{-np/2} |\Lambda|^{n/2} \exp \left[\eta^T \sum_{i=1}^n y_i - \frac{1}{2} \sum_{i=1}^n \text{tr}(\Lambda y_i y_i^T) - \frac{n}{2} \eta^T \Lambda \eta \right] \quad (37.43)$$

, where $\Lambda = \Sigma^{-1}$ and $\eta = \Sigma^{-1} \mu$. And

$$\ln f_Y(y | \mu, \Sigma) = -\frac{np}{2} \ln(2\pi) + \frac{n}{2} \ln |\Lambda| - \frac{n}{2} \eta^T \Lambda \eta + \eta^T \sum_{i=1}^n y_i - \frac{1}{2} \sum_{i=1}^n \text{tr}(\Lambda y_i y_i^T) \quad (37.44)$$

Hence,

$$\begin{aligned} Q(\theta | \theta^{(t)}) &= E_{Y_{(0)}, \theta^{(t)}} [\ell(\theta | Y_{(0)}, Y_{(1)})] \\ &= -\frac{np}{2} \ln(2\pi) + \frac{n}{2} \ln |\Lambda| - \frac{n}{2} \eta^T \Lambda \eta \\ &\quad + \eta^T E \left(\sum_{i=1}^n y_i \right) - \frac{1}{2} \sum_{i=1}^n \text{tr}(\Lambda E(y_i y_i^T)) \end{aligned} \quad (37.45)$$

Therefore, the EM algorithm for incomplete multivariate normal sample is,

- E-step:

$$\begin{aligned} E \left(\sum_{i=1}^n y_{ij} | Y_{(0)}, \theta^{(t)} \right) &= \sum_{i=1}^n y_{ij}^{(t+1)}, \quad j = 1, \dots, p \\ E \left(\sum_{i=1}^n y_{ij} y_{ik} | Y_{(0)}, \theta^{(t)} \right) &= \sum_{i=1}^n (y_{ij}^{(t+1)} y_{ik}^{(t+1)} + c_{jki}^{(t+1)}), \quad j, k = 1, \dots, p \end{aligned} \quad (37.46)$$

where

$$\begin{aligned} y_{ij}^{(t+1)} &= \begin{cases} y_{ij}, & \text{if } y_{ij} \text{ is observed} \\ E(y_{ij} \mid y_{(0),i}, \theta^{(t)}), & \text{if } y_{ij} \text{ is missing} \end{cases} \\ c_{jki}^{(t+1)} &= \begin{cases} 0, & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed} \\ \text{Cov}(y_{ij}, y_{ik} \mid y_{(0),i}, \theta^{(t)}), & \text{if } y_{ij} \text{ and } y_{ik} \text{ are missing} \end{cases} \end{aligned} \quad (37.47)$$

- M-step:

$$\begin{aligned} \mu_j^{(t+1)} &= n^{-1} \sum_{i=1}^n y_{ij}^{(t+1)}, \quad j = 1, \dots, p \\ \sigma_{jk}^{(t+1)} &= n^{-1} E \left(\sum_{i=1}^n y_{ij} y_{ik} \mid Y_{(0)}, \theta^{(t)} \right) - \mu_j^{(t+1)} \mu_k^{(t+1)} \\ &= n^{-1} \sum_{i=1}^n \left[\left(y_{ij}^{(t+1)} - \mu_j^{(t+1)} \right) \left(y_{ik}^{(t+1)} - \mu_k^{(t+1)} \right) + c_{jki}^{(t+1)} \right], \quad j, k = 1, \dots, p \end{aligned} \quad (37.48)$$

Example (Missing Outcomes in Multiple Linear Regression). Suppose a scalar outcome variable Y is regressed on p predictor variables X_1, \dots, X_p , $y_i, i = 1, \dots, m$ are missing, where

$$E(Y \mid X_1, \dots, X_p) = \beta_0 + \sum_{j=1}^p \beta_j X_j \quad (37.49)$$

$$\text{Var}(Y \mid X_1, \dots, X_p) = \sigma^2$$

We assume the joint distribution of the data (including outcomes and predictors) is multi-variate normal with

$$\begin{aligned} \mu &= (\mu_1, \dots, \mu_p, \mu_y) \\ \Sigma &= \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \sigma_{yy} \end{pmatrix} \end{aligned} \quad (37.50)$$

, and that the missing data mechanism is ignorable.

Standard regression theory gives

$$\begin{aligned} \beta &= \Sigma_{yx} \Sigma_{xx}^{-1}; \quad \beta_0 = \mu_y - \sum_{j=1}^p \beta_j \mu_j; \\ \sigma^2 &= \sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \end{aligned} \quad (37.51)$$

The loglikelihood based on the observed data of $\theta = (\beta, \sigma^2)$, where $\beta = (\beta_0, \beta_1, \dots, \beta_p)$, given observed data $\{(x_i, y_i), i = 1, \dots, n\}$ is

$$\ell(\beta, \sigma^2 \mid X, Y_{(0)}) = -\frac{n-m}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=m+1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 \quad (37.52)$$

, where only using the observed data.

EM algorithms can be applied to all observations and will obtain iteratively the same ML estimates as would have been obtained noniteratively using only the complete observations.

- E-step:

$$\begin{aligned} E(y_i \mid X, Y_{(0)}, \theta^{(t)}) &= \begin{cases} y_i, & \text{if } y_i \text{ is observed} \\ \beta^{(t)} \tilde{x}_i^T, & \text{if } y_i \text{ is missing} \end{cases} \\ E(y_i^2 \mid X, Y_{(0)}, \theta^{(t)}) &= \begin{cases} y_i^2, & \text{if } y_i \text{ is observed} \\ (\beta^{(t)} \tilde{x}_i^T)^2 + \sigma^{(t)2}, & \text{if } y_i \text{ is missing} \end{cases} \end{aligned} \quad (37.53)$$

, where $\tilde{x}_i = (1, x_i)$.

- M-step:

$$\begin{aligned}\beta^{(t+1)} &= (X^T X)^{-1} X^T Y^{(t+1)} \\ \sigma^{(t+1)^2} &= n^{-1} \left[\sum_{i=m+1}^n \left(y_i - \beta^{(t)} x_i \right)^2 + m \sigma^{(t)^2} \right]\end{aligned}\quad (37.54)$$

, where $X = (1, X_1, X_2, \dots, X_p)$

37.3 Missing Not At Random Models

Here, we based on

$$L_{\text{full}}(\theta, \psi \mid Y_{(0)}, X, M) \propto f(Y_{(0)}, M \mid X, \theta, \psi) \quad (37.55)$$

regarded as a function of the parameters θ, ψ for fixed observed data $Y_{(0)}$ and missingness pattern M ; here $f(Y_{(0)}, M \mid X, \theta, \psi)$ is obtained by integrating $Y_{(1)}$ out of the joint density $f(Y, M \mid X, \theta, \psi)$ based on a joint model for Y and M given X .

The EM algorithm has the following form for MNAR selection models are as followed,

- E-step:

$$\begin{aligned}Q(\theta, \psi \mid \theta^{(t)}, \psi^{(t)}) &= \int \ell(\theta, \psi \mid X, Y_{(0)}, Y_{(1)}, M) \\ &\quad \cdot f(Y_{(1)} \mid X, Y_{(0)}, M, \theta = \theta^{(t)}, \psi = \psi^{(t)}) dY_{(1)}\end{aligned}\quad (37.56)$$

- M-step:

$$Q(\theta^{(t+1)}, \psi^{(t+1)} \mid \theta^{(t)}, \psi^{(t)}) \geq Q(\theta, \psi \mid \theta^{(t)}, \psi^{(t)}) \quad \text{for all } \theta, \psi \quad (37.57)$$

37.3.1 Normal Models for MNAR Missing Data

1. Follow up a sample of nonrespondents, and incorporate this information into the main analysis.
2. Adopt a Bayesian approach, assigning the parameters prior distributions. Bayesian inference does not generally require that the data provide information for all the parameters, although inferences tend to be sensitive to the choice of prior distribution.
3. Impose additional restrictions on model parameters.
4. Conduct analysis to assess sensitivity of inferences for quantities of interest to different choices of the values of parameters poorly estimated from the data.
5. Selectively discard data to avoid modeling the missingness mechanism.

Chapter 38

Treatment-effects Analysis

38.1 Evaluations

38.1.1 Average Treatment Effect

Definition 38.1.1 (Average Treatment Effect)

$$\mathbb{E}(Y_1 - Y_2) \tag{38.1}$$

38.1.2 Mann-Whitney Statistic

Definition 38.1.2 (Mann-Whitney Statistic)

$$\Pr(Y_1 < Y_2) \tag{38.2}$$

38.1.3 Distribution-type Index

Definition 38.1.3 (Distribution-type Index)

$$F(x) := \Pr(Y_1 - Y_2 = x) \tag{38.3}$$

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.