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 Kernel Methods **75**
23.1 Positive Definite Kernels . . . 75
23.1.1 Construction of the Reproducing Kernel Hilbert Space . . . 75
23.1.2 Examples of Kernels . . . 77

IX Convex Optimization **79**

24 Convex Sets **80**
24.1 Affine and Convex Sets . . . 80
24.1.1 Affine Sets . . . 80
24.1.2 Convex Sets . . . 80
24.1.3 Cones . . . 80
24.2 Some Important Examples . . . 81
24.3 Generalized Inequalities . . . 81
24.3.1 Definition of Generalized Inequalities . . . 81
24.3.2 Properties of Generalized Inequalities . . . 82

25 Convex Optimization Problems **83**
25.1 Generalized Inequality Constraints . . . 83
25.1.1 Conic Form Problems . . . 83
25.1.2 Semidefinite Programming . . . 83
25.2 Vector Optimization . . . 83

26 Unconstrained Minimization **84**
26.1 Definition of Unconstrained Minimization . . . 84
26.2 General Descent Method . . . 86
26.3 Gradient Descent Method . . . 86
26.4 Steepest Descent Method . . . 86
26.5 Newton's Method . . . 86

27 Exercises for Convex Optimization **87**
27.1 Convex Sets . . . 87

X Regression Analysis **88**

28 Modified Likelihood **89**
28.1 Marginal Likelihood . . . 89
28.2 Conditional Likelihood . . . 89

28.3 Profile Likelihood . . . 90
28.4 Quasi Likelihood . . . 90

29 Generalized Linear Model **91**
29.1 Introduction . . . 91
29.2 Binary Data . . . 94
29.3 Polytomous Data . . . 95
29.4 Count Data . . . 96

30 Survival Analysis **97**
30.1 General Formulation . . . 97
30.2 Estimation of Survival Function . . . 98
30.3 Proportional Hazards Model . . . 99

XI Machine Learning **100**

31 Support Vector Machine **101**

32 Linear Discriminant Analysis **103**

33 K-Nearest Neighbor **104**

34 Decision Tree **105**

XII Random Matrix Theory **106**

35 Sample Covariance Matrices **107**
35.1 Eigenvalues and Singular Values . . . 107
35.2 Laguerre Orthogonal Ensemble . . . 108
35.3 Marčenko-Pastur Theorem . . . 112
35.3.1 Preliminary . . . 112
35.3.2 Marčenko-Pastur Theorem . . . 116
35.4 Limits of Extreme Eigenvalues . . . 119

XIII Applications **120**

36 Missingness Data **121**
36.1 The Problem of Missing Data . . . 121
36.1.1 Missingness Mechanisms . . . 122
36.1.2 Commonly Used Methods for Missing Data . . . 122
36.2 Likelihood-Based Inference with Missing Data . . . 123
36.2.1 Ignorable Missingness Mechanism . . . 123
36.2.2 Expectation-Maximization Algorithm . . . 125
36.3 Missing Not At Random Models . . . 130
36.3.1 Normal Models for MNAR Missing Data . . . 130

37 Treatment-effects Analysis **131**
37.1 Evaluations 131
37.1.1 Average Treatment Effect 131
37.1.2 Mann-Whitney Statistic 131
37.1.3 Distribution-type Index 131

Part I

Calculus

Chapter 1

Limit Theory

1.1 Function

Definition 1.1.1 (Mapping)

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

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

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

2. For every class

Chapter 2

Differential Calculus

Chapter 3

Integral Calculus

Part II

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

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \tag{5.1}$$

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 $q_i, i = 1, \ldots, n$. 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

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}' \tag{5.2}$$

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

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' \tag{5.3}$$

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

$$\sigma_i = \mathbf{\Sigma}_{ii}, \quad i = 1, 2, \ldots, r$$

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

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

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

Let

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

then

$$\mathbf{A}'\boldsymbol{u}^{(i)} = \frac{\mathbf{A}'\mathbf{A}\boldsymbol{v}^{(i)}}{\sigma_i} = \sigma_i \boldsymbol{v}^{(i)} \Rightarrow$$
$$\mathbf{A}\mathbf{A}'\boldsymbol{u}^{(i)} = \sigma_i \mathbf{A}\boldsymbol{v}^{(i)} = \sigma_i^2 \boldsymbol{u}^{(i)}$$

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

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

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

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 $\boldsymbol{u}^{(i)}, i = 1, \ldots, r$ with orthonormal vectors $\boldsymbol{u}^{(i)}, i = r+1, \ldots, m$ that span $\text{null}(\mathbf{A}\mathbf{A}')$, and together $\boldsymbol{u}^{(i)}, i = 1, \ldots, n$ 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

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

we get

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

where

$$\mathbf{\Sigma} = \begin{pmatrix} \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \\ \mathbf{0} \end{pmatrix}, \quad \sigma_i = 0 \text{ for } r < i \le n$$

Consequentially, 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 $\mathbf{A} \in \mathbb{R}^{m\times n}$,

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

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,

$$\|\mathbf{U}\mathbf{x}\|_2^2 = \mathbf{x}'\mathbf{U}'\mathbf{U}\mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2^2, \quad \forall \mathbf{x} \in \mathbb{R}^m$$

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,

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

which takes "=", if $\boldsymbol{y} = (1, 0, \ldots, 0)'$. □

Theorem 5.2.2

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

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

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

Remark. If $\operatorname{rank}(\mathbf{A}) < n$, 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 of $\mathcal{S}$ of subsets of Ω is an **semi-algebra** on Ω that satisfy

1. if $A, B \in \mathcal{S}$, then $A \cap B \in \mathcal{S}$.
2. if $A \in \mathcal{S}$, then A^C is a finite disjoint union of sets in $\mathcal{S}$, i.e.,

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

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

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

Lemma 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 $A_i, B_i \in \mathcal{S}$.

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

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

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

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

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

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

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

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

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

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

Theorem 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 $\mu : \mathcal{A} \to \mathbf{R}$ with

1. $\mu(A) \geq \mu(\emptyset) = 0$ for all $A \in \mathcal{A}$.
2. if $A_i \in \mathcal{A}$ is a countable sequence of disjoint sets, then

$$\mu(\cup_i A_i) = \sum_i \mu(A_i).$$

Definition 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 $P(\Omega) = 1$.

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

1. **monotonieity** if $A \subset B$, then $\mu(A) \leq \mu(B)$.
2. **subadditivity** if $A \subset \cup_{m=1}^{\infty} A_m$, then $\mu(A) \leq \sum_{m=1}^{\infty} u(A_m)$.
3. **continuity from below** if $A_i \uparrow A$ (i.e. $A_1 \subset A_2 \subset \ldots$ and $\cup_i A_i = A$), then $\mu(A_i) \uparrow \mu(A)$.
4. **continuity from above** if $A_i \downarrow A$ (i.e. $A_1 \supset A_2 \supset \ldots$ 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 (Ω, A, μ) be a probability space. If f is a real-valued function that is μ-integrable, and if φ is a convex function on the real line, then:

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

Proof. Let $x_0 = \int_\Omega f \mathrm{d}\mu$. 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

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

□

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,

$$\int_\Omega |f \cdot g| \, \mathrm{d}\mu \leq \|f\|_p \|g\|_q. \tag{7.2}$$

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,

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p. \tag{7.3}$$

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

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

Therefore,

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

By Hölder's Inequality (7.1.2),

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

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 $(\Omega, \mathcal{F}, P)$ consisting of:

1. the sample space Ω: an arbitrary non-empty set.
2. the σ-algebra $\mathcal{F} \subseteq 2^{\Omega}$: a set of subsets of Ω, called events.
3. the probability measure $P : \mathcal{F} \to [0, 1]$: a function on $\mathcal{F}$ which is a measure function.

8.2 Random Variables

Definition 8.2.1 (Random Variable)

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

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

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

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 $P : \mathcal{R} \to \mathbb{R}$ by setting

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

Definition 8.3.2 (Distribution Function)

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

$$F(x) = P(X \leq x). \tag{8.3}$$

Definition 8.3.3 (Density Function)

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

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

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. $\lim_{x\to\infty} F(x) = 1, \lim_{x\to-\infty} F(x) = 0$.
3. F is right continuous, i.e., $\lim_{y\downarrow x} F(y) = F(x)$.
4. If $F(x-) = \lim_{y\uparrow x} F(y)$, then $F(x-) = P(X < x)$.
5. $P(X = x) = F(x) - F(x-)$.

Proof. □

Theorem 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

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

Theorem 8.4.3 (Monotone Convergence theorem)

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

$$EX_n \uparrow EX. \tag{8.5}$$

Theorem 8.4.4 (Dominated Convergence theorem)

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

$$EX_n \to EX. \tag{8.6}$$

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 $C, D \in \mathcal{R}$

$$P(X \in C, Y \in D) = P(X \in C)P(Y \in D). \tag{8.7}$$

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 $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n$ are independent if whenever $A_i \in \mathcal{F}_i$ for $i = 1, \ldots, n$, we have
$$P\left(\cap_{i=1}^n A_i\right) = \prod_{i=1}^n P\left(A_i\right). \tag{8.8}$$
2. Random variables $X_1, \ldots, X_n$ are independent if whenever $B_i \in \mathcal{R}$ for $i = 1, \ldots, n$ we have
$$P\left(\cap_{i=1}^n \left\{X_i \in B_i\right\}\right) = \prod_{i=1}^n P\left(X_i \in B_i\right). \tag{8.9}$$
3. Sets $A_1, \ldots, A_n$ are independent if whenever $I \subset \{1, \ldots, n\}$ we have
$$P\left(\cap_{i \in I} A_i\right) = \prod_{i \in I} P\left(A_i\right). \tag{8.10}$$

8.5.2 Sufficient Conditions for Independence

8.5.3 Independence, Distribution, and Expectation

Theorem 8.5.3

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

Theorem 8.5.4

If $X_1, \ldots, X_n$ are independent and have

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

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

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

$$P(X + Y \leq z) = \int F(z - y)\mathrm{d}G(y). \tag{8.12}$$

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

$$h(x) = \int f(x - y)\mathrm{d}G(y). \tag{8.13}$$

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

$$h(x) = \int f(x - y)g(y)\mathrm{d}y. \tag{8.14}$$

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

$$P(X + Y = n) = \sum_m P(X = m)P(Y = n - m). \tag{8.15}$$

8.6 Moments

Lemma 8.6.1

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

$$E(Y^p) = \int_0^\infty py^{p-1}P(Y > y)\mathrm{d}y. \tag{8.16}$$

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

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

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

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

Proof. □

Theorem 8.7.3

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

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

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 $\varphi_n(t) \to \varphi_\infty(t)$ for all t.
2. If $\varphi_n(t)$ converges pointwise to a limit $\varphi(t)$ that is continuous at 0, then the associated sequence of distributions μ_n is tight and converges weakly to the measure μ with characteristic function φ.

Proof. □

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

$$\varphi^{(n)}(t) = \int (ix)^n e^{itx} \mu(\mathrm{d}x). \tag{8.20}$$

Theorem 8.7.6

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

$$\varphi(t) = 1 + itEX - t^2 E\left(X^2\right)/2 + o\left(t^2\right). \tag{8.21}$$

Theorem 8.7.7

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

$$E|X|^2 < \infty. \tag{8.22}$$

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. $\lim_{n\to\infty} E\left(|X_n - X|^r\right) = 0$.

Convergence in the r-th mean is denoted by

$$X_n \xrightarrow{L^r} X. \tag{9.1}$$

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

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

Convergence in probability is denoted by

$$X_n \xrightarrow{p} X. \tag{9.3}$$

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

$$\lim_{n\to\infty} F_n(x) = F(x), \tag{9.4}$$

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

$$X_n \xrightarrow{d} X, \text{ or } X_n \Rightarrow X. \tag{9.5}$$

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 $Y_n, 1 \le n \le \infty$, with distribution F_n so that

$$Y_n \xrightarrow{a.s.} Y_\infty. \tag{9.6}$$

Theorem 9.4.1 (Portmanteau Lemma)

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

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,

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

If in addition g is bounded, then

$$Eg(X_n) \to Eg(X). \tag{9.8}$$

Proof. □

Theorem 9.4.3

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

$$X_n \xrightarrow{d} X, \tag{9.9}$$

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

$$X_n \xrightarrow{p} c. \tag{9.10}$$

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

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

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

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

then, by the union bound,

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

So, we got

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

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

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

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

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

which means,

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

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

$$\lim_{n\to\infty} P(X_n \leq c+\varepsilon) = 1 \Rightarrow \lim_{n\to\infty} P(X_n > c+\varepsilon) = 0$$
$$\lim_{n\to\infty} P(X_n \leq c-\varepsilon) = 0.$$

Therefore,

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

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 $\lim_{k\to\infty} F_{n(k)}(y) = F(y)$ 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

$$\limsup_{n\to\infty} 1 - F_n\left(M_\epsilon\right) + F_n\left(-M_\epsilon\right) \leq \epsilon. \tag{9.11}$$

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

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

Almost sure convergence is denoted by

$$X_n \overset{a.s.}{\to} X. \tag{9.13}$$

Remark.

Theorem 9.5.1

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

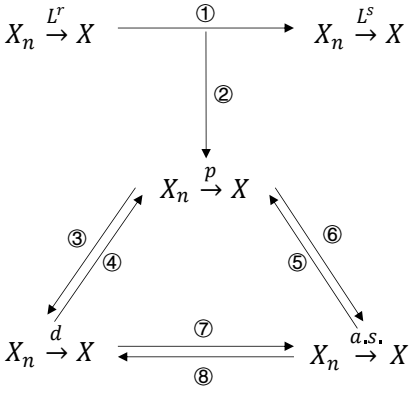
$$X_n \overset{p}{\to} X. \tag{9.14}$$

Proof. □

Theorem 9.5.2

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

$$X_{n(m_k)} \overset{a.s.}{\to} X. \tag{9.15}$$



① $if\ r \geq s \geq 1, X_n \overset{L^r}{\longrightarrow} X \Rightarrow X_n \overset{L^s}{\longrightarrow} X$

② $X_n \overset{L^r}{\longrightarrow} X \Rightarrow X_n \overset{p}{\to} X$

③ $X_n \overset{p}{\longrightarrow} X \Rightarrow X_n \overset{d}{\to} X$

④ $if\ X = c, X_n \overset{d}{\longrightarrow} X \Rightarrow X_n \overset{p}{\to} X$

⑤ $X_n \overset{a.s.}{\longrightarrow} X \Rightarrow X_n \overset{p}{\to} X$

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

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

⑧ $X_n \overset{a.s.}{\longrightarrow} X \Rightarrow X_n \overset{d}{\to} X$

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

$$\frac{A_n}{B_n} \xrightarrow{p} 0. \tag{9.16}$$

Smaller order in probability is denoted by

$$A_n = o_p(B_n). \tag{9.17}$$

Particularly,

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

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

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

Smaller order than or equal to in probability is denoted by

$$A_n = O_p(B_n). \tag{9.20}$$

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

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

Same order in probability is denoted by

$$A_n \asymp_p B_n. \tag{9.22}$$

Chapter 10

Law of Large Numbers

10.1 Weak Law of Large Numbers

Lemma 10.1.1

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

$$Z_n \xrightarrow{p} 0. \tag{10.1}$$

Proof. □

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

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

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

Proof. □

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

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

$$\frac{S_n - \mu_n}{b_n} \xrightarrow{p} 0. \tag{10.3}$$

Proof. □

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

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

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

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

$$\frac{S_n - a_n}{b_n} \xrightarrow{p} 0. \tag{10.4}$$

Proof. □

Theorem 10.1.4 (Weak Law of Large Numbers by Feller)

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

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

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

$$S_n/n - \mu_n \xrightarrow{p} 0. \tag{10.6}$$

Proof. □

Theorem 10.1.5 (Weak Law of Large Numbers)

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

$$S_n/n \xrightarrow{p} \mu. \tag{10.7}$$

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

$$P\left(A_n \text{ i.o. }\right) = 0. \tag{10.8}$$

Lemma 10.2.2 (The Second Borel-Cantelli Lemma)

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

$$P\left(A_n \text{ i.o. }\right) = 1. \tag{10.9}$$

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

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

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

$$S_n/n \xrightarrow{a.s.} \mu. \tag{10.11}$$

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

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

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

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

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

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

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

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

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

Since

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

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

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

By analogous argument,

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

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

Chapter 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)

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \text{ as } n \to \infty. \tag{11.1}$$

Proof. □

Lemma 11.1.2

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

$$(1 + c_j)^{a_j} \to e^{\lambda}. \tag{11.2}$$

Proof. □

Theorem 11.1.1 (The De Moivre-Laplace Theorem)

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

$$P\left(a \leq S_m / \sqrt{m} \leq b\right) \to \int_a^b (2\pi)^{-1/2} e^{-x^2/2} \mathrm{d}x. \tag{11.3}$$

Proof. If n and k and integers

$$P(S_{2n} = 2k) = \begin{pmatrix} 2n \\ n+k \end{pmatrix} 2^{-2n}$$

By lemma 11.1.1, we have

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

Hence,

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

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

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

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

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

Putting things together, we have

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

Therefore,

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

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

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

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

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

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

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

□

11.1.2 Classic Central Limit Theorem

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

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

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

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

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

Theorem 11.1.4 (Berry-Esseen Theorem)

Let $X_1, X_2, \ldots, X_n$ 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),

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

Corollary 11.1.1

Under the assumptions of Theorem 51 ,

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

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

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

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

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

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,

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

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

by Slutsky's Theorem (9.4.4),

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

□

Chapter 13

Exercises for Probability Theory and Examples

13.1 Measure Theory

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

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

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

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

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

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

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

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

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

□

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

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

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

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

□

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

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

Proof.

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

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

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

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

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

and by Exercise 13.3

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

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

Hence, let $M \to \infty$,

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

which means,

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

□

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

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

Exercise. Show that if $|X_i| \leq M$ and $\sum_n \mathrm{Var}\,(X_n) = \infty$, then

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

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

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

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

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

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

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

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.,

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

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 $\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n)$ for $n \geq 1$ and take $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Show X_n is a martingale, and X_n^2 is a submartingale.

Proof. It is obvious that,

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

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

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

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

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:

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

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

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

and the transition probability is

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

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

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

Proof. Since ξ_m are independent with distribution μ,

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

□

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

$$Z_{n+1} = \begin{cases} \xi_1^{n+1} + \cdots + \xi_{Z_n}^{n+1} & Z_n > 0 \\ 0 & Z_n = 0 \end{cases} \tag{15.3}$$

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,

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

□

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

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

Definition 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

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

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

$$P\left(X_j \in B_j, 0 \le j \le n\right) = \int_{B_0} \mu\left(\mathrm{d}x_0\right) \int_{B_1} p\left(x_0, \mathrm{d}x_1\right) \cdots \int_{B_n} p\left(x_{n-1}, \mathrm{d}x_n\right) F \tag{15.5}$$

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

$$T_y^k = \inf\left\{n > T_y^{k-1} : X_n = y\right\} \tag{15.6}$$

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

$$\rho_{xy} = P_x\left(T_y < \infty\right) \tag{15.7}$$

and we have

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

Proof. □

Let

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

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.

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

□

Definition 15.3.3 (Closed State Set)

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

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

Definition 15.3.4 (Irreducible State Set)

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

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

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 $w \in C_x$. □

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. □

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

1. x is transient, if
$$\exists y, \rho_{xy} > 0, \text{ s.t. } \quad \rho_{yx} = 0$$
2. x is recurrent, if
$$\forall y, \rho_{xy} > 0, \text{ s.t. } \quad \rho_{yx} > 0$$

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 $0 = t_0 \leq t_1 \leq \ldots \leq t_n$ the increments
$$B(t_1) - B(t_0), \ldots, B(t_n) - B(t_{n-1})$$
are independent
2. for any $s, t \geq 0$ and Borel sets $A \in \mathbb{R}$,
$$P(B(s+t) - B(s) \in A) = \int_A (2\pi t)^{-1/2} \exp\left(-x^2/2t\right) \mathrm{d}x \tag{16.1}$$
3. the sample paths $t \to B(t)$ are a.s. continuous

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

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

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

2. Suppose $0 \leq s \leq t$,
$$\mathrm{Cov}(B_s, B_t) = E[(B_s - EB_s)(B_t - EB_t)] = EB_sB_t$$
Let $B_t = (B_t - B_s) + B_s$, we have
$$\begin{aligned} EB_sB_t &= E[B_s \cdot ((B_t - B_s) + B_s)] \\ &= E[B_s \cdot (B_t - B_s)] + EB_s^2 \end{aligned}$$
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
$$\mathrm{Cov}(B_s, B_t) = EB_s^2 = s$$

And

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

□

A second equivalent definition of Brownian motion are as followed,

Definition 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 $EB_sB_t = s \wedge t$
3. the sample paths $t \to B(t)$ are a.s. continuous

[a] Gaussian process, i.e., all its finite dimensional distributions are multivariate normal.

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

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

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

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

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

is a martingale.

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

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

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

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

□

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

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

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

Theorem 16.3.1

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

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

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

it follows that

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

□

Definition 16.3.1 (Variation)

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

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

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

$$\Delta t = \max_{i=0,\ldots,n-1} |t_{i+1} - t_i| \tag{16.5}$$

Theorem 16.3.2

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

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

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

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

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

By Theorem 16.3.1, we have

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

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

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

Since

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

hence,

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

while

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

□

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

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

which can be denoted by

$$\int_0^T f(t)\,\mathrm{d}B(t) \tag{16.7}$$

Property. The Itô Stochastic Integral has the following properties:

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

$$\int_0^t (\alpha f(r) + \beta g(r))\,\mathrm{d}B(r) = \alpha \int_0^t f(r)\,\mathrm{d}B(r) + \beta \int_0^t g(r)\,\mathrm{d}B(r) \tag{16.8}$$

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

$$E\left(\left|\int_0^t f(r)\,\mathrm{d}B(r)\right|^2\right) = E\left(\int_0^t |f(r)|^2\,\mathrm{d}r\right) \tag{16.9}$$

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

$$E\left(\int_0^t f(r)\,\mathrm{d}B(r) \mid \mathcal{F}_s\right) = \int_0^s f(r)\,\mathrm{d}B(r) \tag{16.10}$$

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

$$\xi(T) = \xi(0) + \int_0^T a(t)\,\mathrm{d}t + \int_0^T b(t)\,\mathrm{d}B(t) \quad \text{a.s.} \tag{16.11}$$

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

$$\int_0^T |a(t)|\,\mathrm{d}t < \infty \quad \text{a.s.} \tag{16.12}$$

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

$$\mathrm{d}\xi(t) = a(t)\,\mathrm{d}t + b(t)\,\mathrm{d}B(t) \tag{16.13}$$

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

$$\xi(t) = \xi(0) + \int_0^t a(s)\,\mathrm{d}s + \int_0^t b(s)\,\mathrm{d}B(s)$$

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

$$\begin{aligned}\mathrm{d}F(t,\xi(t)) = &\left(F'_t(t,\xi(t)) + F'_x(t,\xi(t))a(t) + \frac{1}{2}F''_{xx}(t,\xi(t))b(t)^2\right)\mathrm{d}t \\ &+ F'_x(t,\xi(t))b(t)\,\mathrm{d}B(t)\end{aligned} \tag{16.14}$$

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

$$\mathrm{d}F(t,B(t)) = \left(F'_t(t,B(t)) + \frac{1}{2}F''_{xx}(t,B(t))\right)\mathrm{d}t + F'_x(t,B(t))\,\mathrm{d}B(t) \tag{16.15}$$

Example (Exponential Martingale)**.** Show that the exponential martingale

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

is an Itô process, and satisfies the equation

$$\mathrm{d}X(t) = X(t)\,\mathrm{d}B(t)$$

Proof. Let $F(t,x) = \mathrm{e}^x\mathrm{e}^{-\frac{t}{2}}$, 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

$$\begin{aligned}\mathrm{d}X(t) =&\mathrm{d}F(t,B(t)) = \left(F'_t(t,B(t)) + \frac{1}{2}F''_{xx}(t,B(t))\right)\mathrm{d}t + F'_x(t,B(t))\,\mathrm{d}B(t) \\ =&\left(-\frac{1}{2}F(t,B(t)) + \frac{1}{2}F(t,B(t))\right)\mathrm{d}t + F(t,B(t))\,\mathrm{d}B(t) \\ =&X(t)\,\mathrm{d}B(t)\end{aligned}$$

□

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

$$\delta : \mathbf{X} \to \mathbb{R}. \tag{18.1}$$

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

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

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

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

that is,

$$\hat{\theta}_n \xrightarrow{p} \theta. \tag{18.4}$$

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

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

Efficiency

Definition 18.3.5 (Efficiency)

Robustness

Definition 18.3.6 (Robustness)

Chapter 19

Maximum Likelihood Estimator

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

We assume that

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

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

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

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

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

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

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

Define

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

19.1 Consistency of MLE

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

$$\frac{1}{n}\sum^{n} l(\theta;X_i) \xrightarrow{p} E[l(\theta;X)]. \tag{19.5}$$

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

$$\hat{\theta}(\mathbf{X}_n) \xrightarrow{p} \theta_0. \tag{19.6}$$

Proof. $\forall \varepsilon > 0$, let

$$\Theta(\epsilon) = \{\theta : \|\theta - \theta_0\| < \epsilon\}.$$

Since $\Theta(\epsilon)$ is an open set, then $\Theta \cap \Theta(\epsilon)^C$ is a compact set (Assumption 2).
Since $Q_0(\theta)$ is a continuous function (Assumption 3), then

$$\theta^* := \sup_{\theta \in \Theta \cap \Theta(\epsilon)^C} \{Q_0(\theta)\}$$

is a achieved for a θ in the compact set.
Since θ_0 is the unique maximized, let

$$Q_0(\theta_0) - Q_0(\theta^*) = \delta > 0.$$

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

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

By Assumption 1,

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

If both A_n and B_n hold, then

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

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

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

which means,

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

□

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

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

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

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

   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

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

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

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

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 $\bar{X}/\sigma$.

2. ξ **is known**:
   Since $s^r = \sum (x_i - \xi)^r$ is the complete sufficient statistic of X_i.
   Assume
   $$E\left[\frac{s^r}{\sigma^r}\right] = \frac{1}{K_{n,r}},$$
   where $K_{n,r}$ is a constant depends on n, r.
   Since $s^2/\sigma^2 \sim \mathrm{Ga}(n/2, 1/2) = \chi^2(n)$, then
   $$E\left[\frac{s^r}{\sigma^r}\right] = E\left[\left(\frac{s^2}{\sigma^2}\right)^{\frac{r}{2}}\right] = \int_0^\infty x^{\frac{r}{2}} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \mathrm{d}x = \frac{\Gamma\left(\frac{n+r}{2}\right)}{\Gamma(\frac{n}{2})} \cdot 2^{\frac{r}{2}}.$$
   therefore,
   $$K_{n,r} = \frac{\Gamma(\frac{n}{2})}{2^{\frac{r}{2}} \cdot \Gamma\left(\frac{n+r}{2}\right)}.$$
   Hence,
   $$E\left[s^r K_{n,r}\right] = \sigma^r \text{ and } E[\xi s^{-1} K_{n,-1}] = \xi/\sigma,$$
   which means the UMVU estimator of σ^r is $s^r K_{n,r}$ and the UMVU estimator of ξ/σ is $\xi s^{-1} K_{n,-1}$.
3. **Both** ξ **and** σ **is unknown**:
   Since $(\bar{X}, s_x^r)$ are the complete sufficient statistic of X_i, where $s_x^2 = \sum \left(x_i - \bar{X}\right)^r$.
   Since $s_x^2/\sigma^2 \sim \chi^2(n-1)$, then
   $$E\left[\frac{s_x^r}{\sigma^r}\right] = \frac{1}{K_{n-1,r}}.$$
   Hence,
   $$E\left[s_x^r K_{n-1,r}\right] = \sigma^r,$$
   which means the UMVU estimator of σ^r is $s_x^r K_{n-1,r}$, and
   $$E(\bar{X}^r) = \xi^r,$$
   which means the UMVU estimator of ξ^r is $\bar{X}^r$.
   Since $\bar{X}$ and s_x^r are independent, then
   $$E[\bar{X} s_x^{-1} K_{n-1,-1}] = \xi/\sigma$$
   which means the UMVU estimator of ξ/σ is $\bar{X} s_x^{-1} K_{n-1,-1}$.

□

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

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

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

.

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

$$r(\Lambda, \delta) = \int R(\theta, \delta) \, \mathrm{d}\Lambda(\theta) \tag{21.1}$$

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

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

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

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

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

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
$$\delta_\Lambda(x) = E[g(\Theta) \mid x]. \tag{21.4}$$
2. If $L(\theta, d) = w(\theta)[d - g(\theta)]^2$, then
$$\delta_\Lambda(x) = \frac{E[w(\theta) g(\Theta) \mid x]}{E[w(\theta) \mid x]}. \tag{21.5}$$
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 maxmizes $P(\Theta \in I \mid x)$.

Proof. □

Theorem 21.0.2

Necessiary condition for Bayes Estimator

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

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

21.1 Single-Prior Bayes

The Single-Prior Bayes model in a general form as

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

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

$$\int L(\theta, d(x)) \pi(\theta \mid x) \, \mathrm{d}\theta, \tag{21.7}$$

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

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

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

$$E(\Theta \mid x) = \frac{\int \theta f(x \mid \theta)\, \pi(\theta \mid \gamma_0)\, \mathrm{d}\theta}{\int f(x \mid \theta)\, \pi(\theta \mid \gamma_0)\, \mathrm{d}\theta}. \tag{21.8}$$

Example. Consider

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

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

Proof.

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

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

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

□

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

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

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 ofthen done in practice.

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

$$\int L(\theta, d(x))\, \pi(\theta \mid x)\, \mathrm{d}\theta, \tag{21.10}$$

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

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

Remark. The posterior distribution can also be writed as

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

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

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

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

Example (Poisson Hierarchy)**.** Consider

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

calculate the Hierarchical Bayes estimator under squared error loss.

Theorem 21.2.1

For the Hierarchical Bayes model (21.9),

$$K[\pi(\lambda \mid x), \psi(\lambda)] < K[\pi(\theta \mid x), \pi(\theta)], \tag{21.12}$$

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

Kernel Methods

23.1 Positive Definite Kernels

Definition 23.1.1 (Positive Definite Kernel)

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

1. symmetric, that is,
$$K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x}), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X} \tag{23.1}$$
2. positive definite, that is,
$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0, \tag{23.2}$$
holds for any $x_1, \ldots, x_n \in \mathcal{X}$, given $n \in \mathbb{N}, c_1, \ldots, c_n \in \mathbb{R}$.

23.1.1 Construction of the Reproducing Kernel Hilbert Space

Theorem 23.1.1 (Morse-Aronszajn's Theorem)

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

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

Consider the vector space $\mathcal{H}_0 \subset \mathcal{R}^{\mathcal{X}}$ 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
$$\langle f, g \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j K(\mathbf{x}_i, \mathbf{y}_j) \tag{23.3}$$

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-Schwarwz Inequality, we have,

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

therefore, if $\|f\|_{\mathcal{H}_0} = 0$, then $f = 0$
Hence, definition in equation 23.3 is a valid inner product, which is a valid pre-RKHS $\mathcal{H}_0$.
□

23.1.2 Examples of Kernels

Example (Gaussian Kernel)**.**

$$K(\mathbf{x},\mathbf{y})=\exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right),\quad \mathbf{x},\mathbf{y}\in\mathbb{R}^d \tag{23.4}$$

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

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

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

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

then,

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

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

□

Part IX

Convex Optimization

Chapter 24

Convex Sets

24.1 Affine and Convex Sets

24.1.1 Affine Sets

Definition 24.1.1 (Affine Set)

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

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

24.1.2 Convex Sets

Definition 24.1.2 (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 24.1.3 (Convex Hull)

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

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

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

24.1.3 Cones

Definition 24.1.4 (Cone)

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

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

Definition 24.1.5 (Convex Cone)

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

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

24.2 Some Important Examples

Definition 24.2.1 (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 24.2.2 (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 24.2.3 ((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 24.2.4 (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).

24.3 Generalized Inequalities

24.3.1 Definition of Generalized Inequalities

Definition 24.3.1 (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 24.3.2 (Generalized Inequalities)

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

$$y - x \in K, \tag{24.1}$$

which can be denoted by

$$x \preceq_K y \text{ or } y \succeq_K x. \tag{24.2}$$

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

$$y - x \in \text{ int } K, \tag{24.3}$$

which can be denoted by

$$x \prec_K y \text{ or } y \succ_K x. \tag{24.4}$$

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 $\mathbf{R}$.

24.3.2 Properties of Generalized Inequalities

Theorem 24.3.1 (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 25

Convex Optimization Problems

25.1 Generalized Inequality Constraints

Definition 25.1.1 (With Generalized Inequality Constraints)

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

$$\begin{array}{ll} \min_x & f_0(x) \\ \text{s.t.} & f_i(x) \preceq_{K_i} 0, \quad i = 1, \ldots, m \\ & Ax = b \end{array} \tag{25.1}$$

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

25.1.1 Conic Form Problems

Definition 25.1.2 (Conic Form Problem)

A conic form problem is defined to be

$$\begin{array}{ll} \min & c^T x \\ \text{s.t.} & Fx + g \preceq_K 0 \\ & Ax = b \end{array} \tag{25.2}$$

25.1.2 Semidefinite Programming

25.2 Vector Optimization

Chapter 26

Unconstrained Minimization

26.1 Definition of Unconstrained Minimization

Definition 26.1.1 (Unconstrained Minimization Problem)

The unconstrained minimization problem is defined to be

$$\min_{\mathbf{x}} f(\mathbf{x}) \tag{26.1}$$

where $f : \mathbf{R}^n \to \mathbf{R}$ 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

$$\nabla f(\mathbf{x}^*) = 0 \tag{26.2}$$

Solving (26.1) is equal to finding the solution of (26.2), thus (26.1) can be solved by analytic solution of (26.2) in a few cases, but usually can be solved by an iterative algorithm, i.e.,

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

where the initial point $\mathbf{x}^{(0)}$ must lie in $\operatorname{dom} f$, and the sublevel set

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

must be closed.

Remark.

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

$$\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}'\mathbf{P}\mathbf{x} + \mathbf{q}'\mathbf{x} + r \tag{26.3}$$

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

The optimality condition is

$$\mathbf{P}\mathbf{x}^* + \mathbf{q} = \mathbf{0} \tag{26.4}$$

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 (26.4) is optimal for (26.3).
   Since $\mathbf{P} \not\succeq 0$, 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 $t \to \infty$.
3. If (26.4) does not have a solution, then (26.3) is unbounded.
   Since (26.4) 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,

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \mathbf{x}' (\mathbf{A}'\mathbf{A}) \mathbf{x} - 2 (\mathbf{A}'\mathbf{b})' \mathbf{x} + \mathbf{b}'\mathbf{b} \tag{26.5}$$

The optimality condition is

$$\mathbf{A}'\mathbf{A}\mathbf{x}^* = \mathbf{A}'\mathbf{b} \tag{26.6}$$

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

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

$$\min_{\mathbf{x}} f(\mathbf{x}) = \log \left(\sum_{i=1}^{m} \exp(\mathbf{a}_i'\mathbf{x} + b_i) \right) \tag{26.7}$$

The optimality condition is

$$\nabla f(x^*) = \frac{\sum_{i=1}^{m} \exp(\mathbf{a}_i'\mathbf{x}^* + b_i) \mathbf{a}_i}{\sum_{j=1}^{m} \exp(\mathbf{a}_j'\mathbf{x}^* + b_j)} = \mathbf{0} \tag{26.8}$$

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

Example (Analytic Center of Linear Inequalities)**.**

Definition 26.1.2 (Strong Convexity)

26.2 General Descent Method

26.3 Gradient Descent Method

26.4 Steepest Descent Method

26.5 Newton's Method

Chapter 27

Exercises for Convex Optimization

27.1 Convex Sets

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

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

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 $x, y \in C$.

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

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

□

Part X

Regression Analysis

Chapter 28

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.

28.1 Marginal Likelihood

28.2 Conditional Likelihood

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

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

$$\ell_c = \log\left(f_{Y|S_{\boldsymbol{\lambda}}}\right) \tag{28.1}$$

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

$$\ell(\boldsymbol{\theta}, \mathbf{y}) = \boldsymbol{\theta}'\mathbf{s} - b(\boldsymbol{\theta}) \tag{28.2}$$

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

$$\ell(\boldsymbol{\theta}, \mathbf{y}) = \boldsymbol{\varphi}'\mathbf{s}_1 + \boldsymbol{\lambda}'\mathbf{s}_2 - b(\boldsymbol{\varphi}, \boldsymbol{\lambda}) \tag{28.3}$$

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

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

$$\ell(\boldsymbol{\varphi} \mid \mathbf{s}_2) = \boldsymbol{\varphi}'\mathbf{s}_1 - b^*(\boldsymbol{\varphi}, \boldsymbol{\lambda}) \tag{28.4}$$

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

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

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

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,

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

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

□

28.3 Profile Likelihood

28.4 Quasi Likelihood

Chapter 29

Generalized Linear Model

29.1 Introduction

Definition 29.1.1 (Exponential Family)

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

$$f\left(y \mid \theta, \phi\right) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \tag{29.1}$$

Property. The exponential family have the following properties,

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

Proof. □

Table 29.1: 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	Bern(p)	$\log\left(\frac{p}{1-p}\right)$	1	$\log\left(1+e^{\theta}\right)$	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

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

with link function g, such that,

$$E(Y \mid \mathbf{X}) = \mu = g^{-1}(\eta), \quad \eta = \mathbf{X}'\boldsymbol{\beta} \tag{29.2}$$

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 29.2: Commonly Used Link Functions

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

Maximum Likelihood Suppose the log-likelihood function be

$$\ell(\boldsymbol{\beta} \mid \mathbf{X}, y) = \log[f(y \mid \theta, \phi)] = \log\left[f\left(y \mid g^{-1}(\eta), \phi\right)\right] \tag{29.3}$$

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

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

By mean value theorem,

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

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

$$\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t + A^{-1}\left(\hat{\boldsymbol{\beta}}_t\right) U\left(\hat{\boldsymbol{\beta}}_t\right) \Leftrightarrow A\left(\hat{\boldsymbol{\beta}}_t\right) \hat{\boldsymbol{\beta}}_{t+1} = A\left(\hat{\boldsymbol{\beta}}_t\right) \hat{\boldsymbol{\beta}}_t + U\left(\hat{\boldsymbol{\beta}}_t\right) \tag{29.4}$$

where

$$U(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \tag{29.5}$$

is the score function and

$$A(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}' \partial \boldsymbol{\beta}} \tag{29.6}$$

is the observed information matrix.

2. Fisher's Scoring Method:

$$\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t + I^{-1}\left(\hat{\boldsymbol{\beta}}_t\right) U\left(\hat{\boldsymbol{\beta}}_t\right) \Leftrightarrow I\left(\hat{\boldsymbol{\beta}}_t\right) \hat{\boldsymbol{\beta}}_{t+1} = I\left(\hat{\boldsymbol{\beta}}_t\right) \hat{\boldsymbol{\beta}}_t + U\left(\hat{\boldsymbol{\beta}}_t\right) \tag{29.7}$$

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

$$I(\boldsymbol{\beta}) = E[A(\boldsymbol{\beta})] = -E\left[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}' \partial \boldsymbol{\beta}}\right] \tag{29.8}$$

is the Fisher information matrix.

Bayesian Methods

29.2 Binary Data

Suppose

$$Y \sim b(m, \pi), \quad i = 1, 2, \ldots, n \tag{29.9}$$

with link function

$$\eta = g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \mathbf{x}'\boldsymbol{\beta} \tag{29.10}$$

Remark.

The likelihood function is

$$f(\boldsymbol{\pi} \mid \mathbf{X}, \mathbf{y}) = \prod_{i=1}^{n} \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i} \tag{29.11}$$

and the log-likelihood function is

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \log[f(\boldsymbol{\pi} \mid \mathbf{X}, \mathbf{y})] = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \left\{ \log\left[\binom{m_i}{y_i}\right] + y_i \log(\pi_i) + (m_i - y_i)\log(1-\pi_i) \right\} \\ &= \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i) \right] + \sum_{i=1}^{n} \log\left[\binom{m_i}{y_i}\right] \end{aligned} \tag{29.12}$$

where

$$\pi_i = \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)} \tag{29.13}$$

Thus,

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

29.3 Polytomous Data

Definition 29.3.1 (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 scales [1]. For pure scales, there are several types:

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

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

$$\pi_j\left(\mathbf{x}_i\right) = P\left(Y = y_j \mid \mathbf{X} = \mathbf{x}_i\right) \tag{29.14}$$

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

$$r_j\left(\mathbf{x}_i\right) = P\left(Y \leq \sum_{r\leq j} y_r \mid \mathbf{X} = \mathbf{x}_i\right) \tag{29.15}$$

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

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

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

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

As for the link function, we have

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

Nominal Scale

$$\pi_j(\mathbf{x}_i) = \frac{\exp[\eta_j(\mathbf{x}_i)]}{\sum_{j=1}^{k}\exp[\eta_j(\mathbf{x}_i)]} \tag{29.16}$$

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:

$$\log\left[\frac{r_j(\mathbf{x}_i)}{1-r_j(\mathbf{x}_i)}\right] = \theta_j - \mathbf{x}_i'\boldsymbol{\beta} \tag{29.17}$$

2. Complementary Log-Log Scale:

$$\log\{-\log[1-r_j(\mathbf{x}_i)]\} = \theta_j - \mathbf{x}_i'\boldsymbol{\beta} \tag{29.18}$$

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.

$$\log\left[\frac{r_j(\mathbf{x}_i)}{1-r_j(\mathbf{x}_i)}\right] = \varsigma_0 + \varsigma_1\left(\frac{s_j+s_{j+1}}{2}\right) - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'\boldsymbol{\xi}(c_j - \bar{c}) \tag{29.19}$$

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.

$$\pi_j(\mathbf{x}_i) = \frac{\exp[\eta_j(\mathbf{x}_i)]}{\sum_{j=1}^{k}\exp[\eta_j(\mathbf{x}_i)]} \tag{29.20}$$

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

3.

$$\sum_{j=1}^{k}\pi_j(\mathbf{x}_i)s_j = \mathbf{x}_i\boldsymbol{\beta} \tag{29.21}$$

29.4 Count Data

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

$$\text{Var}(Y) = \sigma^2 E(Y) \tag{29.22}$$

with link function

$$\log(\mu) = \eta = \mathbf{x}'\boldsymbol{\beta} \tag{29.23}$$

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

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

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

and the log-likelihood function is

$$\ell(\boldsymbol{\beta}) \propto \sum_{i=1}^{n}(y_i\log(\mu_i) - \mu_i) \tag{29.24}$$

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

Chapter 30

Survival Analysis

30.1 General Formulation

Definition 30.1.1 (Survival Function)

The survival function[a] is defined to be

$$S(t) = P(T > t) = \int_t^\infty f(u)\,\mathrm{d}u = 1 - F(t). \tag{30.1}$$

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

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

Definition 30.1.2 (Lifetime Distribution Function)

The lifetime distribution function is defined to be

$$F(t) = P(T \leq t) \tag{30.2}$$

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

$$f(t) = F'(t) = \frac{\mathrm{d}}{\mathrm{d}t}F(t) \tag{30.3}$$

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

Definition 30.1.3 (Hazard Function)

The Hazard function[a] is defined to be

$$\lambda(t) = \lim_{\varepsilon \to 0^+} \left[\frac{P(t \leq T < t + \varepsilon \mid T \geq t)}{\varepsilon} \right] = \frac{f(t)}{S(t)} \tag{30.4}$$

[a] The 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.
$$\lambda(t) = -\frac{\mathrm{d}\log[S(t)]}{\mathrm{d}t} \tag{30.5}$$

2.
$$S(t) = \exp\left[-\int_0^t \lambda(x)\,\mathrm{d}x\right] \tag{30.6}$$

3.
$$f(t) = \lambda(t)\exp\left[-\int_0^t \lambda(x)\,\mathrm{d}x\right] \tag{30.7}$$

Proof. □

Example (Constant Hazards)**.** Suppose

$$\lambda(t) = \lambda \tag{30.8}$$

then

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

which is the exponential distribution.

Example (Bathtub Hazards)**.**

$$\lambda(t) = \alpha t + \frac{\beta}{1+\gamma t} \tag{30.9}$$

30.2 Estimation of Survival Function

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

$$\begin{aligned} f\left(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}\right) &= \prod_{i=1}^{n} \left[f\left(t_i\right)\right]^{\delta_i} \left[S\left(t_i\right)\right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left(\frac{f\left(t_i\right)}{S\left(t_i\right)}\right)^{\delta_i} S\left(t_i\right) \\ &= \prod_{i=1}^{n} \left[\lambda\left(t_i\right)\right]^{\delta_i} S\left(t_i\right) \end{aligned} \tag{30.10}$$

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

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

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

Thus,

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

Hence,

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

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

$$\begin{aligned}\hat{S}(t) &= \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right) \\ &= \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right) \cdots \left(1 - \frac{d_k}{n_k}\right) \\ &\approx \left[1 - \hat{\lambda}\left(t_1\right)\right]\left[1 - \hat{\lambda}\left(t_2\right)\right] \ldots \left[1 - \hat{\lambda}\left(t_k\right)\right]\end{aligned} \tag{30.11}$$

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

30.3 Proportional Hazards Model

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

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

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

$$\lambda(t) = \lambda_0(t) \mathrm{e}^{\mathbf{x}'\boldsymbol{\beta}} \tag{30.12}$$

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

The exact likelihood function is

$$\ell\left[\beta, \lambda_0(t)\right] = \prod_{i=1}^{n} \left[\lambda_i\left(t_i\right)\right]^{\delta_i} S\left(t_i\right) \tag{30.13}$$

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

$$\ell_p(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{\mathrm{e}^{\mathbf{x}'_{(j)}\boldsymbol{\beta}}}{\sum_{l \in R\left(t_{(j)}\right)} \mathrm{e}^{\mathbf{x}'_l \boldsymbol{\beta}}} = \prod_{i=1}^{n} \left[\frac{\mathrm{e}^{\mathbf{x}'_i \boldsymbol{\beta}}}{\sum_{l \in R(t_i)} \mathrm{e}^{\mathbf{x}'_l \boldsymbol{\beta}}}\right]^{\delta_i} \tag{30.14}$$

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

Part XI

Machine Learning

Chapter 31

Support Vector Machine

Theorem 31.0.1

The minimizer of

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

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

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

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

Proof. For the hinge loss function, that,

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

When $g(x)<-1$,

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

When $g(x)>1$,

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

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

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

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

Chapter 32

Linear Discriminant Analysis

Chapter 33

K-Nearest Neighbor

Chapter 34

Decision Tree

Part XII

Random Matrix Theory

Chapter 35

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

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

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

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

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

$$E\left(\widehat{\mathbf{\Sigma}}\right) = \mathbb{E}(\mathbf{X}\otimes\mathbf{X}) = \mathbf{I}_n$$

For convenience, we define the random matrix

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

35.1 Eigenvalues and Singular Values

Theorem 35.1.1

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

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

if $m_n \geq n$, then

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

Proof. □

35.2 Laguerre Orthogonal Ensemble

Definition 35.2.1 (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:

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

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

$$\mathbf{M} = \mathbb{X}'\mathbb{X} = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' \tag{35.1}$$

and which can be denoted by

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

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

Theorem 35.2.1

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

$$f\left(\mathbf{M}\right) = \frac{1}{2^{np/2}\left[\det\left(\mathbf{\Sigma}\right)\right]^{n/2}\Gamma_p\left(\frac{n}{2}\right)} \det\left(\mathbf{M}\right)^{(n-p-1)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{M}\right)\right] \tag{35.2}$$

with respect to Lebesque 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 $\widehat{\mathbf{\Sigma}}$ can be derived from the Wishart distribution. The probability denisty function of $\widehat{\mathbf{\Sigma}}$ can be derived from (35.2), since

$$\mathbf{A} \sim W_n\left(\mathbf{I}_n, m_n\right), \quad \det\left(\widehat{\mathbf{\Sigma}}\right) = m_n^{-n} \det\left(\mathbf{A}\right), \quad \operatorname{tr}\left(\widehat{\mathbf{\Sigma}}\right) = m_n^{-1} \operatorname{tr}\left(\mathbf{A}\right)$$

thus,

$$f\left(\widehat{\mathbf{\Sigma}}\right) = \frac{m_n^{-n(m_n-n-1)/2+1}}{2^{m_n n/2}\Gamma_n\left(\frac{m_n}{2}\right)} \det\left(\widehat{\mathbf{\Sigma}}\right)^{(m_n-n-1)/2} \exp\left[-\frac{m_n}{2}\operatorname{tr}\left(\widehat{\mathbf{\Sigma}}\right)\right] \tag{35.3}$$

Theorem 35.2.2

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

$$p(\boldsymbol{\lambda}) = \widetilde{Q}_{m_n,n}^{-1} \exp\left(-\frac{m_n}{2}\sum_{k=1}^{n}\lambda_k\right)\prod_{k=1}^{n}\lambda_k^{(m_n-n-1)/2}\prod_{i<j}|\lambda_i-\lambda_j| \tag{35.4}$$

where

$$0 \le \lambda_1 \le \ldots \le \lambda_n < \infty$$

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

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

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

where $\{P_{ij}\}_{1\le i\le j\le n}\in\mathbb{R}^{(n+1)n/2}$ 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

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

where

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

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

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

thus,

$$[\det(\mathbf{Q})]^{-m_n/2} = \left[\det\left(\mathbf{I}_n - \frac{2\imath}{m_n}\mathbf{P}\right)\right]^{-m_n/2} = \prod_{k=1}^{n}\left(1-\frac{2\imath}{m_n}p_k\right)^{-m_n/2} \tag{35.5}$$

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

Then, we will show that the characteristic function of (35.4) conincides with the above function. By the Wishart distribution, the probability denisty of the real symmetric and positive definite random matrix $\widehat{\mathbf{\Sigma}}$ is

$$\widetilde{Q}_{m_n,n}^{-1} \exp\left[-\frac{m_n}{2}\operatorname{tr}\left(\widehat{\mathbf{\Sigma}}\right)\right]\left[\det\left(\widehat{\mathbf{\Sigma}}\right)\right]^{(m_n-n-1)/2} \mathrm{d}\widehat{\mathbf{\Sigma}} \tag{35.6}$$

where $\widetilde{Q}_{m_n,n}$ is the normalization constant. Then, the characteristic function of (35.6), i.e.,

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

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

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

and

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

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 (35.6), and the characteristic function is

$$Q_{m_n,n}^{-1} \int_{(\mathbb{R}_+)^n} \exp\left[\sum_{k=1}^n \left(\imath p_k - \frac{m_n}{2}\right)\lambda_k\right] \prod_{k=1}^n \lambda_k^{(m_n-n-1)/2} \prod_{i<j} |\lambda_i - \lambda_j| \, \mathrm{d}\boldsymbol{\lambda} \tag{35.7}$$

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

If we viewed (35.5) and (35.7) as the function of $\{p_k\}_{k=1}^n \in \mathbb{R}^n$, then they can be **analytic continuation** to the domain

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

If we replace $\{p_k\}_{k=1}^n$ by $\{\imath p_k', p_k' \geq 0\}_{k=1}^n$ on (35.5), since this is a set of uniqueness of both (35.5) and (35.7) analytic functions, we have

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

where $q_k = 1 + \frac{2p_k'}{m_n} \geq 1, k = 1, \ldots, n$, and since

$$\forall i, j \quad \frac{q_i}{q_j} = \frac{1 + \frac{2p_i'}{m_n}}{1 + \frac{2p_j'}{m_n}} \to 1, \quad \text{as} \quad m_n \to \infty$$

we have

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

thus,

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

Since

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

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

$$\prod_{k=1}^{n} \left(1 + \frac{2p_k'}{m_n}\right)^{-m_n/2} \widetilde{Q}_n^{-1} \int_{\mathcal{S}_n^+} \exp\left[-\frac{m_n}{2} \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}\right)\right] \left[\det\left(\widehat{\boldsymbol{\Sigma}}\right)\right]^{(m_n-n-1)/2} \mathrm{d}\widehat{\boldsymbol{\Sigma}}$$

The integral here is equal to $\widetilde{Q}_n$, the normalization constant of the probability measure (35.6). If we replace $\{\imath p_k'\}_{k=1}^n$ back by $\{p_k\}_{k=1}^n$, then the above expression is

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

which coincides with (35.5). Thus the probability law of the Wishart matrices of $\boldsymbol{\Sigma}$ given by (35.6) implies that the corresponding joint probability density of eigenvalues is given by (35.4) for $\boldsymbol{\Sigma}$. □

Definition 35.2.2 (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

$$p(\boldsymbol{\lambda}) = K_{\alpha,\beta} \exp\left(-\frac{\beta}{2}\sum_{k=1}^{n}\lambda_k\right) \prod_{k=1}^{n} \lambda_k^{\frac{\alpha\beta}{2}} \prod_{i<j} |\lambda_i - \lambda_j|^{\beta} \tag{35.8}$$

where

$$0 \le \lambda_1 \le \ldots \le \lambda_n < \infty$$

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

And Equation (35.8) can be written in the standard Boltzmann-Gibbs form, that,

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

where

$$E(\boldsymbol{\lambda}) = \frac{1}{2}\sum_{k=1}^{n}(\lambda_k - \alpha\log\lambda_k) - \frac{1}{2}\sum_{i\neq j}|\lambda_i - \lambda_j| \tag{35.9}$$

Remark. For the (35.4), which can be written as (35.8) form, that,

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

where $\beta = 1$ and

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

35.3 Marčenko-Pastur Theorem

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

35.3.1 Preliminary

Empirical Spectral Measure

Definition 35.3.1 (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 $\lambda_1(\mathbf{M}), \ldots, \lambda_n(\mathbf{M})$ of $\mathbf{M}$, i.e.,

$$\mu_{\mathbf{M}} := \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{M})} \tag{35.10}$$

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 $\int \mu_{\mathbf{M}}(\mathrm{d}x) = 1$, 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

$$\det(\mathbf{M}) = \prod_{i=1}^{n} \lambda_i = \exp\left(n \int_0^\infty \log x \mu_{\mathbf{M}}(\mathrm{d}x)\right) \tag{35.11}$$

Stieltjes Transform

Definition 35.3.2 (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

$$\mathbf{Q}_{\mathbf{M}}(z) := (\mathbf{M} - z\mathbf{I}_n)^{-1} \tag{35.12}$$

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

Definition 35.3.3 (Stieltjes Transform)

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

$$m_\mu(z) := \int \frac{1}{t-z} \mu(\mathrm{d}t) \tag{35.13}$$

where $z \in \mathbb{C} \backslash \mathrm{supp}(\mu)$.

Property. The Stieltjes transform m_μ has numerous interesting properties:

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

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

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

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

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

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

Theorem 35.3.1 (Inverse Stieltjes Transform)

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

$$\mu\left([a,b]\right) = \frac{1}{\pi}\lim_{y\downarrow 0}\int_a^b \Im\left[m_\mu(x+\imath y)\right]\,\mathrm{d}x \tag{35.14}$$

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

$$f(x) = \frac{1}{\pi}\lim_{y\downarrow 0}\Im\left[m_\mu(x+\imath y)\right] \tag{35.15}$$

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

$$\mu(\{x\}) = \lim_{y\downarrow 0} -\imath y m_\mu(x+\imath y) \tag{35.16}$$

Proof.

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

By Fubini's theorem,

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

Since

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

by the dominated convergence theorem,

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

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

$$= \int 1_{[a,b]}\mu(\mathrm{d}t) = \mu\left([a,b]\right)$$

Thus, if μ has a density f at x, then

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

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

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

by dominated convergence theorem,

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

□

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

$$m_{\mu_{\mathbf{M}}}(z) = \frac{1}{n}\sum_{i=1}^n \int \frac{\delta_{\lambda_i(\mathbf{M})}(t)}{t-z} = \frac{1}{n}\sum_{i=1}^n \frac{1}{\lambda_i(\mathbf{M})-z} = \frac{1}{n}\operatorname{tr}\mathbf{Q}_{\mathbf{M}}(z) \tag{35.17}$$

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

$$\frac{1}{n}\sum_{i=1}^n f\left(\lambda_i(\mathbf{M})\right) = -\frac{1}{2\pi\imath n}\oint_\Gamma f(z)\operatorname{tr}\left(\mathbf{Q}_{\mathbf{M}}(z)\right)\mathrm{d}z = -\frac{1}{2\pi\imath}\oint_\Gamma f(z)m_{\mu_{\mathbf{M}}}(z)\mathrm{d}z \tag{35.18}$$

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

Matrix Equivalents

Definition 35.3.4 (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),

$$\frac{1}{n}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\overline{\mathbf{Q}})\to 0,\quad \mathbf{a}'(\mathbf{Q}-\overline{\mathbf{Q}})\mathbf{b}\to 0,\quad \text{as } n\to\infty \tag{35.19}$$

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

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

this convergence implies that the Stieltjes transform of μ_{M} "converges" in the sense that

$$m_{\mu_{\mathrm{M}}}(z) - \bar{m}_n(z) \to 0$$

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

Definition 35.3.5 (Matrix Equivalents)

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

$$\mathbf{X} \leftrightarrow \mathbf{Y} \tag{35.20}$$

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

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

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

Resolvent and Perturbation Identities

Lemma 35.3.1 (Resolvent Identity)

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

$$\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B}-\mathbf{A})\mathbf{B}^{-1} \tag{35.21}$$

Lemma 35.3.2 (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

$$(\mathbf{A}+\mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}}{1+\mathbf{v}'\mathbf{A}^{-1}\mathbf{u}} \tag{35.22}$$

or,

$$(\mathbf{A}+\mathbf{u}\mathbf{v}')^{-1}\mathbf{u} = \frac{\mathbf{A}^{-1}\mathbf{u}}{1+\mathbf{v}'\mathbf{A}^{-1}\mathbf{u}} \tag{35.23}$$

Lemma 35.3.3 (Quadratic-form-close-to-the-trace)

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

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

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

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

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

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

35.3.2 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 35.3.2 (Marčenko-Pastur Theorem)

Consider the resolvent

$$\mathbf{Q}(z) = \left(\widehat{\mathbf{\Sigma}} - z\mathbf{I}_n\right)^{-1}$$

Then, if

$$\frac{n}{m_n} \to \rho \text{ with } \rho \in (0,\infty), \quad \text{as } n \to \infty$$

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 $\mathcal{Z}\left(\mathbb{C}\backslash\left[(1-\sqrt{\rho})^2, (1+\sqrt{\rho})^2\right]\right)$ be

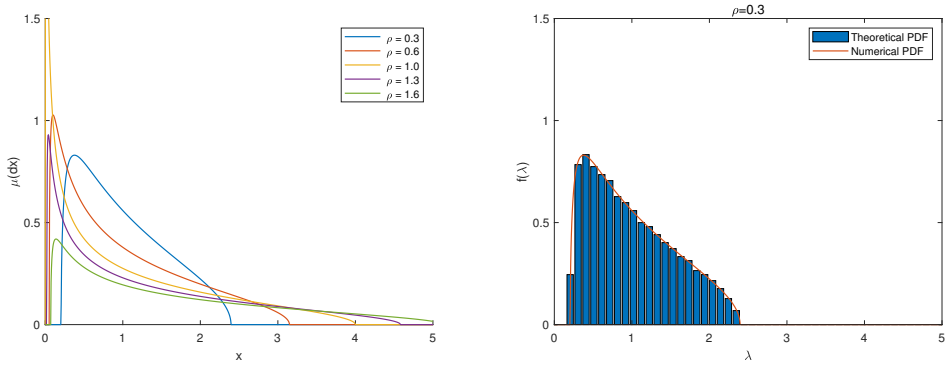
$$zcm^2(z) - (1-c-z)m(z) + 1 = 0$$

where the function $m(z)$ is the Stieltjes transform of the probability measure μ given explicitly by

$$\mu(\mathrm{d}x) = \left(1-\rho^{-1}\right)^+\delta_0(x) + \frac{\sqrt{(x-a_-)^+(a_+-x)^+}}{2\pi\rho x}\,\mathrm{d}x$$

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 $\mu_{\widehat{\mathbf{\Sigma}}}$ 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 35.1: Illustrations of the Marčenko-Pastur Theorem

deterministic equivalent for $\mathbf{Q}(z)$, particularly,

$$\frac{1}{n}\operatorname{tr}\mathbf{A}(\mathbf{Q}(z)-\overline{\mathbf{Q}}(z))\to 0 \quad \text{a.s.}$$

where $\mathbf{A}$ is arbitrary, deterministic, and such that $\|\mathbf{A}\|=1$. By Lemma 35.3.1, we have

$$\begin{aligned}\mathbf{Q}(z)-\overline{\mathbf{Q}}(z)=&\mathbf{Q}(z)\left(\mathbf{F}(z)+z\mathbf{I}_n-\widehat{\mathbf{\Sigma}}\right)\overline{\mathbf{Q}}(z)\\ =&\mathbf{Q}(z)\left(\mathbf{F}(z)+z\mathbf{I}_n-\frac{1}{m_n}\sum_{i=1}^{m_n}\mathbf{X}_i\mathbf{X}_i'\right)\overline{\mathbf{Q}}(z)\end{aligned}$$

Thus, we turn to prove that,

$$\frac{1}{n}\operatorname{tr}\left[(\mathbf{F}(z)+z\mathbf{I}_n)\,\overline{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)\right]-\frac{1}{n}\cdot\frac{1}{m_n}\sum_{i=1}^{m_n}\mathbf{X}_i'\overline{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)\mathbf{X}_i\to 0 \quad \text{a.s.}$$

By Lemma 35.3.2, 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 35.3.3, we have

$$\frac{1}{n}\mathbf{X}_i'\overline{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)\mathbf{X}_i=\frac{\frac{1}{n}\mathbf{X}_i'\overline{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}_{-i}(z)\mathbf{X}_i}{1+\frac{1}{m_n}\mathbf{X}_i'\mathbf{Q}_{-i}(z)\mathbf{X}_i}\simeq\frac{\frac{1}{n}\operatorname{tr}\left[\overline{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}_{-i}(z)\right]}{1+\frac{1}{m_n}\operatorname{tr}\left[\mathbf{Q}_{-i}(z)\right]}$$

Hence, we need to prove the approximation that

$$\frac{1}{n}\operatorname{tr}\left[(\mathbf{F}(z)+z\mathbf{I}_n)\,\overline{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)\right]\simeq\frac{\frac{1}{n}\operatorname{tr}\left[\overline{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)\right]}{1+\frac{1}{m_n}\operatorname{tr}\left[\mathbf{Q}(z)\right]}$$

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 35.17, we have,

$$m(z) \equiv \frac{1}{n} \operatorname{tr} \left[\overline{\mathbf{Q}}(z)\right] = \frac{1}{n} \operatorname{tr} \left[\mathbf{F}(z)^{-1}\right]$$

taking $\mathbf{A} = \mathbf{I}_n$, we have

$$\frac{1}{n} \operatorname{tr} \left[\mathbf{Q}(z)\right] \simeq \frac{1}{n} \operatorname{tr} \left[\overline{\mathbf{Q}}(z)\right] = m(z) = \frac{1}{-z + \frac{1}{1 + \frac{n}{m_n} \frac{1}{n} \operatorname{tr}[\mathbf{Q}(z)]}} \simeq \frac{1}{-z + \frac{1}{1 + \rho m(z)}}$$

As $n, m_n \to \infty$, $m(z)$ 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

$$z = r\mathrm{e}^{\imath\theta} \text{ where } r \geq 0, \theta \in [0, 2\pi) \Rightarrow \sqrt{z} \in \left\{\pm\sqrt{r}\mathrm{e}^{\imath\theta/2}\right\}$$

and we can conclude that

$$m(z) = \frac{1 - \rho - z}{2\rho z} + \frac{\sqrt{\left((1 + \sqrt{\rho})^2 - z\right)\left((1 - \sqrt{\rho})^2 - z\right)}}{2\rho z}$$

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 35.3.1, we find that $m(z)$ is the Stieltjes transform of the measure μ with

$$\mu([a, b]) = \frac{1}{\pi} \lim_{\epsilon \downarrow 0} \int_a^b \Im[m(x + \imath\epsilon)] \, \mathrm{d}x$$

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

$$\left[(1 - \sqrt{\rho})^2, (1 + \sqrt{\rho})^2\right]$$

(and thus of non-real square root), the latter defines the support of the continuous part of the measure μ with density

$$\frac{\sqrt{\left((1 + \sqrt{\rho})^2 - x\right)\left(x - (1 - \sqrt{\rho})^2\right)}}{2\rho\pi x}$$

at point x in the set. The case $x = 0$ brings a discontinuity in μ with weight equal to

$$\mu(\{0\}) = -\lim_{y \downarrow 0} \imath y m(\imath y) = \frac{\rho - 1}{2\rho} \pm \frac{\rho - 1}{2\rho}$$

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. □

Remark. The asymptotic phenomenon holds not only in the Gaussian case, which also holds
1. if $(x_i)_{1\le i\le n}$ are i.i.d. with finite second moment.
2. if $\mathbf{X}$ is isotropic and log-concave[1] random vector.

35.4 Limits of Extreme Eigenvalues

The weak convergence in Theorem 35.3.2 does not provide much information at the edge on the behavior of the extremal atoms, and what one can actually extract is that

$$\limsup_{n\to\infty} \lambda_{\min}\left(\widehat{\mathbf{\Sigma}}\right) \le (1-\sqrt{\rho})^2 \le (1+\sqrt{\rho})^2 \le \liminf_{n\to\infty} \lambda_{\max}\left(\widehat{\mathbf{\Sigma}}\right) \quad \text{a.s.} \tag{35.24}$$

where the first inequality is considered only in the case where $m_n \ge n$.

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 $(X_{n,k})_{n\ge 1, 1\le k\le n}$ are i.i.d. with finite fourth moment then,

$$(1-\sqrt{\rho})^2 = \lim_{n\to\infty} \lambda_{\min}\left(\widehat{\mathbf{\Sigma}}\right) \le \lim_{n\to\infty} \lambda_{\max}\left(\widehat{\mathbf{\Sigma}}\right) = (1+\sqrt{\rho})^2 \quad \text{a.s.} \tag{35.25}$$

where the first inequality is considered only in the case where $m_n \ge n$.

Remark. The convergence of the smallest eigenvalue in the left hand side of (35.25) holds if $(x_i)_{1\le i\le n}$ are i.i.d. with finite second moment.

Theorem 35.4.1

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 $(x_i)_{1\le i\le n}$ are i.i.d. then

$$\liminf_{n\to\infty} \frac{E\left(\lambda_{\min}\left(\mathbf{A}_n\right)\right)}{\left(\sqrt{m_n}-\sqrt{n}\right)^2} \ge 1 \tag{35.26}$$

If additionally $\lim_{n\to\infty} \frac{n}{m_n} = \rho$ with $\rho \in (0,1)$, in other words $\underline{\rho} = \bar{\rho} \in (0,1)$, then

$$\lambda_{\min}\left(\widehat{\mathbf{\Sigma}}_n\right) \xrightarrow{p} (1-\sqrt{\rho})^2 \quad \text{as } n\to\infty \tag{35.27}$$

Proof. □

Theorem 35.4.2

If the centered isotropic random vector $\mathbf{X}$ is log-concave or if $(x_i)_{1\le i\le n}$ are i.i.d. with finite 4-th moment then

$$\limsup_{n\to\infty} \frac{E\left(\lambda_{\max}\left(\mathbf{A}_n\right)\right)}{\left(\sqrt{m_n}+\sqrt{n}\right)^2} \le 1 \tag{35.28}$$

If additionally $\lim_{n\to\infty} \frac{n}{m_n} = \rho$ with $\rho \in (0,1)$, in other words $\underline{\rho} = \bar{\rho} \in (0,1)$, then

$$\lambda_{\max}\left(\widehat{\mathbf{\Sigma}}_n\right) \xrightarrow{p} (1+\sqrt{\rho})^2 \quad \text{as } n\to\infty \tag{35.29}$$

Proof. □

[1] A probability measure μ on $\mathbb{R}^n$ with density φ is log-concave when $\varphi = e^{-V}$ with V convex.

Part XIII

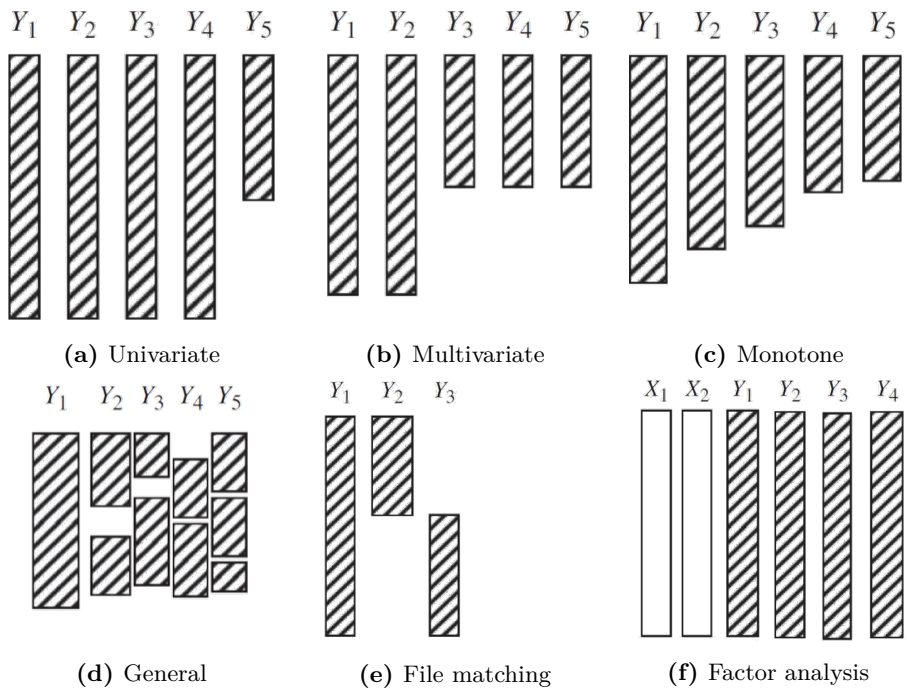
Applications

Chapter 36

Missingness Data

36.1 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 36.1).



(a) Univariate **(b)** Multivariate **(c)** Monotone

(d) General **(e)** File matching **(f)** Factor analysis

Figure 36.1: 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 $y_i = (y_{i1}, y_{i2}, \ldots, y_{ip})$, $m_i = (m_{i1}, m_{i2}, \ldots, m_{ip})$ 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.

36.1.1 Missingness Mechanisms

The missingness mechanism is characterized by the conditional distribution of m_i given y_i, say

$$f_{M|Y}\left(m_i \mid y_i, \phi\right), \tag{36.1}$$

where ϕ denotes unknown parameters.

Definition 36.1.1 (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,

$$f_{M|Y}\left(m_i \mid y_i, \phi\right) = f_{M|Y}\left(m_i \mid y^*, \phi\right), \tag{36.2}$$

then the data are called missing completely at random, MCAR.

Definition 36.1.2 (Missing at Random, MAR)

If missingness depends on y_i only through the observed components $y_{(0)}$, i.e., if for all y_i and any distinct values $y^*_{(1)}$ in the sample space of $y_{(1)}$,

$$f_{M|Y}\left(m_i \mid y_{(0)i}, y_{(1)i}, \phi\right) = f_{M|Y}\left(m_i \mid y_{(0)i}, y^*_{(1)}, \phi\right), \tag{36.3}$$

then the data are called missing at random, MAR.

Definition 36.1.3 (Missing Not at Random, MNAR)

If missingness depends on y_i the missing components $y_{(1)}$, i.e., if some y_i and some values $y^*_{(1)}$ in the sample space of $y_{(1)}$,

$$f_{M|Y}\left(m_i \mid y_{(0)i}, y_{(1)i}, \phi\right) \neq f_{M|Y}\left(m_i \mid y_{(0)i}, y^*_{(1)}, \phi\right), \tag{36.4}$$

then the data are called missing not at random, MNAR.

36.1.2 Commonly Used Methods for Missing Data

1. Complete-case Analysis: discard incompletely recorde 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.

36.2 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 $\left(y_{(0)}, m\right)$ and the assumed joint distribution model above is defined to be

$$L_{\text{full}}\ \left(\theta, \psi \mid y_{(0)}, m\right) = \int f_Y\left(y_{(0)}, y_{(1)} \mid \theta\right) f_{M|Y}\left(m \mid y_{(0)}, y_{(1)}, \psi\right) \mathrm{d}y_{(1)} \tag{36.5}$$

The likelihood of θ ignoring the missingness mechanism is defined to be

$$L_{\text{ign}}\left(\theta \mid y_{(0)}\right) = \int f_Y\left(y_{(0)}, y_{(1)} \mid \theta\right) \mathrm{d}y_{(1)} \tag{36.6}$$

36.2.1 Ignorable Missingness Mechanism

Definition 36.2.1 (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).

$$\frac{L_{\text{full}}\ \left(\theta, \psi \mid \tilde{y}_{(0)}, \tilde{m}\right)}{L_{\text{full}}\ \left(\theta^*, \psi \mid \tilde{y}_{(0)}, \tilde{m}\right)} = \frac{L_{\text{ign}}\ \left(\theta \mid \tilde{y}_{(0)}\right)}{L_{\text{ign}}\ \left(\theta^* \mid \tilde{y}_{(0)}\right)} \quad \forall \theta, \theta^*, \psi. \tag{36.7}$$

Theorem 36.2.1

The missingness mechanism is ignorable for direct likelihood inference on $\left(\tilde{m}, \tilde{y}_{(0)}\right)$ 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

$$L_{\text{full}}\left(\theta, \psi \mid \tilde{y}_{(0)}, \tilde{m}\right) = L_{\text{ign}}\left(\theta \mid \tilde{y}_{(0)}\right) \times L_{\text{rest}}\left(\psi \mid \tilde{y}_{(0)}, \tilde{m}\right) \quad \forall \theta, \psi \in \Omega_{\theta,\psi} \tag{36.8}$$

Corollary 36.2.1

If the missing data are MAR at $\left(\tilde{m}, \tilde{y}_{(0)}\right)$, and θ and ψ are distinct, the missingness mechanism is ignorable for likelihood inference.

Proof. Since,

$$f_{M|Y}\left(\tilde{m} \mid \tilde{y}_{(0)}, y_{(1)}, \psi\right) = f_{M|Y}\left(\tilde{m} \mid \tilde{y}_{(0)}, y^*_{(1)}, \psi\right) \quad \forall y_{(1)}, y^*_{(1)}, \psi \tag{36.9}$$

therefore,

$$\begin{aligned} f\left(\tilde{y}_{(0)}, \tilde{m} \mid \theta, \psi\right) &= f_{M|Y}\left(\tilde{m} \mid \tilde{y}_{(0)}, \psi\right) \times \int f_Y\left(\tilde{y}_{(0)}, y_{(1)} \mid \theta\right) \mathrm{d}y_{(1)} \\ &= f_{M|Y}\left(\tilde{m} \mid \tilde{y}_{(0)}, \psi\right) \times f_Y\left(\tilde{y}_{(0)} \mid \theta\right) \end{aligned} \tag{36.10}$$

yields the factored likelihood equation 36.8. □

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

$$f_Y(y \mid \theta) = \theta^{-n} \exp\left\{-\sum_{i=1}^{n} \frac{y_i}{\theta}\right\}. \tag{36.11}$$

The loglikelihood fuction is

$$\ell_Y(\theta \mid y) = \ln\left\{\theta^{-n} \exp\left(-\sum_{i=1}^{n} \frac{y_i}{\theta}\right)\right\} = -n \ln \theta - \sum_{i=1}^{n} \frac{y_i}{\theta}. \tag{36.12}$$

Differentiating with respect to θ gives the likelihood equation

$$-\frac{n}{\theta} + \sum_{i=1}^{n} \frac{y_i}{\theta^2} = 0. \tag{36.13}$$

Thus, we obtain the ML estimates

$$\hat{\theta} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{36.14}$$

Example (Incomplete Exponential Sample)**.** Suppose we have an incomplete univariate exponential sample with $y_{(0)} = (y_1, \ldots, y_r)^{\mathrm{T}}$ observed and $y_{(1)} = (y_{r+1}, \ldots, y_n)^{\mathrm{T}}$ missing. Thus, $m = (m_1, \ldots, m_n)^{\mathrm{T}}$, where $m_i = 0, i = 1, \ldots, r$ and $m_i = 1, i = r+1, \ldots, n$.

The likelihood ignoring the missingness mechanism is

$$L_{\text{ign}}\left(\theta \mid y_{(0)}\right) = \theta^{-r} \exp\left(-\sum_{i=1}^{r} \frac{y_i}{\theta}\right). \tag{36.15}$$

If each unit is observed with probability ψ that does not depend on Y, that is,

$$f_{M|Y}(m \mid y, \psi) = \frac{n!}{r!(n-r)!} \psi^r (1-\psi)^{n-r} \tag{36.16}$$

then,

$$f\left(y_{(0)}, m \mid \theta, \psi\right) = \frac{n!}{r!(n-r)!} \psi^r (1-\psi)^{n-r} \theta^{-r} \exp\left(-\sum_{i=1}^{r} \frac{y_i}{\theta}\right) \tag{36.17}$$

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

$$\hat{\theta} = \frac{1}{r}\sum_{i=1}^{r} y_i. \tag{36.18}$$

If each unit is observed only if values less than c, that is

$$f_{M|Y}(m \mid y, \psi) = \prod_{i=1}^{n} f\left(m_i \mid y_i, \psi\right), \tag{36.19}$$

where

$$f\left(m_i \mid y_i, \psi\right) = \begin{cases} 1, & y_i \geq c \\ 0, & \text{otherwise} \end{cases} \tag{36.20}$$

Hence,

$$\begin{aligned} L_{\text{full}}\ \left(\theta \mid y_{(0)}, m\right) &= \prod_{i=1}^{r} f_Y\left(y_i \mid \theta\right) \Pr\left(y_i < c \mid y_i, \theta\right) \times \prod_{i=r+1}^{n} \Pr\left(y_i \geq c \mid \theta\right) \\ &= \theta^{-r} \exp\left(-\sum_{1}^{r} \frac{y_i}{\theta}\right) \times \exp\left(-\frac{(n-r)c}{\theta}\right) \end{aligned} \tag{36.21}$$

Maximizing above equation with respect to θ gives the ML estimate

$$\hat{\theta} = \frac{1}{r}\left[\sum_{i=1}^{r} y_i + (n-r)c\right]. \tag{36.22}$$

The inflation of the sample mean in this expression reflects the censoring of the missing values.

36.2.2 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 $\theta^{(t)}$:

$$Q\left(\theta \mid \theta^{(t)}\right) = \int \ell\left(\theta \mid Y_{(0)}, Y_{(1)}\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)}. \tag{36.23}$$

The M step of EM determines $\theta^{(t+1)}$ by maximizing this expected completedata loglikelihood:

$$Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) \geq Q\left(\theta \mid \theta^{(t)}\right), \quad \forall\theta. \tag{36.24}$$

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 36.2.2

Every GEM algorithm increases $\ell\left(\theta \mid Y_{(0)}\right)$ at each iteration, that is,

$$\ell\left(\theta^{(t+1)} \mid Y_{(0)}\right) \geq \ell\left(\theta^{(t)} \mid Y_{(0)}\right) \tag{36.25}$$

, with equality if and only if

$$Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) = Q\left(\theta^{(t)} \mid \theta^{(t)}\right) \tag{36.26}$$

Proof. The distribution of the complete data Y can be factored as follows:

$$f(Y \mid \theta) = f\left(Y_{(0)}, Y_{(1)} \mid \theta\right) = f\left(Y_{(0)} \mid \theta\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) \tag{36.27}$$

The corresponding decomposition of the loglikelihood is

$$\ell(\theta \mid Y) = \ell\left(\theta \mid Y_{(0)}, Y_{(1)}\right) = \ell\left(\theta \mid Y_{(0)}\right) + \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) \tag{36.28}$$

Let,

$$\ell\left(\theta \mid Y_{(0)}\right) = \ell(\theta \mid Y) - \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) \tag{36.29}$$

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

$$\ell\left(\theta \mid Y_{(0)}\right) = Q\left(\theta \mid \theta^{(t)}\right) - H\left(\theta \mid \theta^{(t)}\right) \tag{36.30}$$

, where

$$\begin{aligned} Q\left(\theta \mid \theta^{(t)}\right) &= \int \ell\left(\theta \mid Y_{(0)}, Y_{(1)}\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\ H\left(\theta \mid \theta^{(t)}\right) &= \int \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \end{aligned} \tag{36.31}$$

Since,

$$\begin{aligned} & H\left(\theta^{(t)}, \theta^{(t)}\right) - H\left(\theta, \theta^{(t)}\right) \\ &= \int \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\ &\quad - \int \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta)\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\ &= \int \ln \left[\frac{f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right)}{f\left(Y_{(1)} \mid Y_{(0)}, \theta\right)}\right] f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\ &= \int -\ln \left[\frac{f\left(Y_{(1)} \mid Y_{(0)}, \theta\right)}{f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right)}\right] f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\ &\geq -\ln \int \frac{f\left(Y_{(1)} \mid Y_{(0)}, \theta\right)}{f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right)} f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} = 0 \end{aligned} \tag{36.32}$$

Therefore,

$$H\left(\theta \mid \theta^{(t)}\right) \leq H\left(\theta^{(t)} \mid \theta^{(t)}\right) \tag{36.33}$$

Hence, the difference in values of $\ell\left(\theta \mid Y_{(0)}\right)$ at successive iterates is given by

$$\begin{aligned}\ell\left(\theta^{(t+1)} \mid Y_{(0)}\right) - \ell\left(\theta^{(t)} \mid Y_{(0)}\right) &= \left[Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) - Q\left(\theta^{(t)} \mid \theta^{(t)}\right)\right] \\ &\quad - \left[H\left(\theta^{(t+1)} \mid \theta^{(t)}\right) - H\left(\theta^{(t)} \mid \theta^{(t)}\right)\right] \\ &\geq 0\end{aligned} \tag{36.34}$$

□

Theorem 36.2.3

Suppose a sequence of EM iterates is such that

1. $D^{10}Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) = 0$, where "D" here denotes derivative, and D^{10} means the derivative with respect to the first argument, that is, define

$$D^{10}Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) = \frac{\partial}{\partial\theta}Q\left(\theta \mid \theta^{(t)}\right)\Big|_{\theta=\theta^{(t+1)}} = 0. \tag{36.35}$$

2. $\theta^{(t)}$ converges to θ^*.
3. $f\left(Y_{(1)} \mid Y_{(0)}, \theta\right)$ is smooth in θ, where smooth is defined in the proof.

Then

$$D\ell\left(\theta^* \mid Y_{(0)}\right) \equiv \frac{\partial}{\partial\theta}\ell\left(\theta \mid Y_{(0)}\right)\Big|_{\theta=\theta^*} = 0, \tag{36.36}$$

so that if the $\theta^{(t)}$ converge, they converge to a stationary point.

Proof.

$$\begin{aligned}D\ell\left(\theta^{(t+1)} \mid Y_{(0)}\right) &= D^{10}Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) - D^{10}H\left(\theta^{(t+1)} \mid \theta^{(t)}\right) \\ &= -D^{10}H\left(\theta^{(t+1)} \mid \theta^{(t)}\right) \\ &= -\frac{\partial}{\partial\theta}\int\left[\ln f\left(Y_{(1)} \mid Y_{(0)}, \theta\right)\right] f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) dY_{(1)}\Big|_{\theta=\theta^{(t+1)}}\end{aligned} \tag{36.37}$$

which assuming sufficient smoothness to interchange the order of differentiation and integration,

$$\begin{aligned}&= -\int\frac{\partial}{\partial\theta}f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) dY_{(1)}\Big|_{\theta=\theta^{(t+1)}} \\ &= -\int\frac{\partial}{\partial\theta}f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) dY_{(1)}\Big|_{\theta=\theta^{(t+1)}} = 0\end{aligned} \tag{36.38}$$

□

Examples of EM Algorithm with Missing Data

Example (Multivariate Normal Sample)**.** Let $y = (y_{ij})$, where $i = 1, \ldots, n$, $j = 1, \ldots, p$, be a matrix representing an independent and identically distributed sample of n units from the multivariate normal distribution with mean vector $\mu = (\mu_1, \ldots, \mu_p)$ and covariance matrix $\Sigma = (\sigma_{jk})$, $j = 1, \ldots, p$; $k = 1, \ldots, p$. Thus, y_{ij} represents the value of the jth variable for the ith unit in the sample. The density of y is

$$f_Y(y \mid \mu, \Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \mu\right)\Sigma^{-1}\left(y_i - \mu\right)^{\mathrm{T}}\right\}. \tag{36.39}$$

The loglikelihood of $\theta = (\mu, \Sigma)$ is then

$$\ell_Y(\mu, \Sigma \mid y) = -\frac{n}{2} \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu) \Sigma^{-1} (y_i - \mu)^{\mathrm{T}} \tag{36.40}$$

Maximizing above equation with respect to θ and Σ gives the ML estimate

$$\hat{\mu} = \bar{y}, \quad \hat{\Sigma} = \frac{n-1}{n} S, \tag{36.41}$$

where $\bar{y} = (\bar{y}_1, \ldots, \bar{y}_p)$ 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 $Y = \left(Y_{(0)}, Y_{(1)}\right)$, where Y represents a random sample of size n on $(Y_1, \ldots, Y_p), Y_{(0)}$ 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, $\quad i = 1, \ldots, n$.

The loglikelihood based on the observed data is then

$$\ell\left(\mu, \Sigma \mid Y_{(0)}\right) = -\frac{1}{2} \sum_{i=1}^{n} \ln \left|\Sigma_{(0),i}\right| - \frac{1}{2} \sum_{i=1}^{n} \left(y_{(0),i} - \mu_{(0),i}\right)^{\mathrm{T}} \Sigma_{(0),i}^{-1} \left(y_{(0),i} - \mu_{(0),i}\right) \tag{36.42}$$

, 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

$$f_Y(y \mid \mu, \Sigma) = (2\pi)^{-np/2} |\Lambda|^{n/2} \exp\left[\eta^T \sum_{i=1}^{n} y_i - \frac{1}{2} \sum_{i=1}^{n} \operatorname{tr}\left(\Lambda y_i y_i^T\right) - \frac{n}{2} \eta^T \Lambda \eta\right] \tag{36.43}$$

, where $\Lambda = \Sigma^{-1}$ and $\eta = \Sigma^{-1}\mu$. And

$$\ln f_Y(y \mid \mu, \Sigma) = -\frac{np}{2} \ln(2\pi) + \frac{n}{2} \ln |\Lambda| - \frac{n}{2} \eta^T \Lambda \eta + \eta^T \sum_{i=1}^{n} y_i - \frac{1}{2} \sum_{i=1}^{n} \operatorname{tr}\left(\Lambda y_i y_i^T\right) \tag{36.44}$$

Hence,

$$\begin{aligned} Q\left(\theta \mid \theta^{(t)}\right) &= E_{Y_{(0)}, \theta^{(t)}}\left[\ell\left(\theta \mid Y_{(0)}, Y_{(1)}\right)\right] \\ &= -\frac{np}{2} \ln(2\pi) + \frac{n}{2} \ln |\Lambda| - \frac{n}{2} \eta^T \Lambda \eta \\ &\quad + \eta^T E\left(\sum_{i=1}^{n} y_i\right) - \frac{1}{2} \sum_{i=1}^{n} \operatorname{tr}\left(\Lambda E\left(y_i y_i^T\right)\right) \end{aligned} \tag{36.45}$$

Therefore, the EM algorithm for incomplete multivariate normal sample is,

- E-step:

$$\begin{aligned} E\left(\sum_{i=1}^{n} y_{ij} \mid Y_{(0)}, \theta^{(t)}\right) &= \sum_{i=1}^{n} y_{ij}^{(t+1)}, \quad j = 1, \ldots, p \\ E\left(\sum_{i=1}^{n} y_{ij} y_{ik} \mid Y_{(0)}, \theta^{(t)}\right) &= \sum_{i=1}^{n} \left(y_{ij}^{(t+1)} y_{ik}^{(t+1)} + c_{jki}^{(t+1)}\right), \quad j, k = 1, \ldots, p \end{aligned} \tag{36.46}$$

where

$$\begin{aligned} y_{ij}^{(t+1)} &= \begin{cases} y_{ij}, & \text{if } y_{ij} \text{ is observed} \\ E\left(y_{ij} \mid y_{(0),i}, \theta^{(t)}\right), & \text{if } y_{ij} \text{ is missing} \end{cases} \\ c_{jki}^{(t+1)} &= \begin{cases} 0, & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed} \\ \operatorname{Cov}\left(y_{ij}, y_{ik} \mid y_{(0),i}, \theta^{(t)}\right), & \text{if } y_{ij} \text{ and } y_{ik} \text{ are missing} \end{cases} \end{aligned} \tag{36.47}$$

- M-step:

$$\begin{aligned} \mu_j^{(t+1)} &= n^{-1} \sum_{i=1}^{n} y_{ij}^{(t+1)}, \quad j = 1, \ldots, p \\ \sigma_{jk}^{(t+1)} &= n^{-1} E\left(\sum_{i=1}^{n} y_{ij} y_{ik} \mid Y_{(0)}, \theta^{(t)}\right) - \mu_j^{(t+1)} \mu_k^{(t+1)} \\ &= n^{-1} \sum_{i=1}^{n} \left[\left(y_{ij}^{(t+1)} - \mu_j^{(t+1)}\right)\left(y_{ik}^{(t+1)} - \mu_k^{(t+1)}\right) + c_{jki}^{(t+1)}\right], \quad j, k = 1, \ldots, p \end{aligned} \tag{36.48}$$

Example (Missing Outcomes in Multiple Linear Regression)**.** Suppose a scalar outcome variable Y is regressed on p predictor variables $X_1, \ldots, X_p$, $y_i, i = 1, \ldots, m$ are missing, where

$$\begin{aligned} E\left(Y \mid X_1, \ldots, X_p\right) &= \beta_0 + \sum_{j=1}^{p} \beta_j X_j \\ \operatorname{Var}\left(Y \mid X_1, \ldots, X_p\right) &= \sigma^2 \end{aligned} \tag{36.49}$$

We assume the joint distribution of the data (including outcomes and predictors) is multivariate normal with

$$\begin{aligned} \mu &= (\mu_1, \ldots, \mu_p, \mu_y) \\ \Sigma &= \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \sigma_{yy} \end{pmatrix} \end{aligned} \tag{36.50}$$

, and that the missing data mechanism is ignorable.

Standard regression theory gives

$$\begin{gathered} \beta = \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1}; \quad \beta_0 = \mu_y - \textstyle\sum_{j=1}^{p} \beta_j \mu_j; \\ \sigma^2 = \sigma_{yy} - \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy} \end{gathered} \tag{36.51}$$

The loglikelihood based on the observed data of $\theta = \left(\beta, \sigma^2\right)$, where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, given observed data $\left\{\left(x_i, y_i\right), i = 1, \ldots, n\right\}$ is

$$\ell(\beta, \sigma^2 \mid X, Y_{(0)}) = -\frac{n-m}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=m+1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j\right)^2 \tag{36.52}$$

, 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:

$$\begin{aligned} E\left(y_i \mid X, Y_{(0)}, \theta^{(t)}\right) &= \begin{cases} y_i, & \text{if } y_i \text{ is observed} \\ \beta^{(t)} \tilde{x}_i^{\mathrm{T}} & \text{if } y_i \text{ is missing} \end{cases} \\ E\left(y_i^2 \mid X, Y_{(0)}, \theta^{(t)}\right) &= \begin{cases} y_i^2, & \text{if } y_i \text{ is observed} \\ \left(\beta^{(t)} \tilde{x}_i^{\mathrm{T}}\right)^2 + \sigma^{(t)^2}, & \text{if } y_i \text{ is missing} \end{cases} \end{aligned} \tag{36.53}$$

, where $\tilde{x}_i = (1, x_i)$.

- M-step:

$$\begin{aligned} \beta^{(t+1)} &= \left(X^{\mathrm{T}} X\right)^{-1} X^{\mathrm{T}} Y^{(t+1)} \\ \sigma^{(t+1)^2} &= n^{-1} \left[\sum_{i=m+1}^{n} \left(y_i - \beta^{(t)} x_i \right)^2 + m\sigma^{(t)^2} \right] \end{aligned} \tag{36.54}$$

, where $X = (1, X_1, X_2, \ldots, X_p)$

36.3 Missing Not At Random Models

Here, we based on

$$L_{\text{full}} \left(\theta, \psi \mid Y_{(0)}, X, M\right) \propto f\left(Y_{(0)}, M \mid X, \theta, \psi\right) \tag{36.55}$$

regarded as a function of the parameters θ, ψ for fixed observed data $Y_{(0)}$ and missingness pattern M; here $f\left(Y_{(0)}, M \mid X, \theta, \psi\right)$ 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:

$$\begin{aligned} Q\left(\theta, \psi \mid \theta^{(t)}, \psi^{(t)}\right) = \int & \ell\left(\theta, \psi \mid X, Y_{(0)}, Y_{(1)}, M\right) \\ & \cdot f\left(Y_{(1)} \mid X, Y_{(0)}, M, \theta = \theta^{(t)}, \psi = \psi^{(t)}\right) \mathrm{d}Y_{(1)} \end{aligned} \tag{36.56}$$

- M-step:

$$Q\left(\theta^{(t+1)}, \psi^{(t+1)} \mid \theta^{(t)}, \psi^{(t)}\right) \geq Q\left(\theta, \psi \mid \theta^{(t)}, \psi^{(t)}\right) \quad \text{for all } \theta, \psi \tag{36.57}$$

36.3.1 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 onmodel 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 37

Treatment-effects Analysis

37.1 Evaluations

37.1.1 Average Treatment Effect

Definition 37.1.1 (Average Treatment Effect)

$$\mathrm{E}\left(Y_1 - Y_2\right) \tag{37.1}$$

37.1.2 Mann-Whitney Statistic

Definition 37.1.2 (Mann-Whitney Statistic)

$$\Pr\left(Y_1 < Y_2\right) \tag{37.2}$$

37.1.3 Distribution-type Index

Definition 37.1.3 (Distribution-type Index)

$$F(x) := \Pr\left(Y_1 - Y_2 = x\right) \tag{37.3}$$

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.