Contents

I Calculus 1

1 Limit Theory 2
1.1 Function . . . 2

2 Differential Calculus 3

3 Integral Calculus 4

II Real Analysis 5

4 Measure Theory 6
4.1 Semi-algebras, Algebras and Sigma-algebras . . . 6
4.2 Measure . . . 8

5 Lebesgue Integration 9
5.1 Properties of the Integral . . . 9
5.2 Product Measures . . . 10

III Functional Analysis 11

IV Matrix Theory 12

6 Matrix Norms 13
6.1 Matrix Norms Induced by Vector Norms . . . 13

7 Matrix Decompositions 14
7.1 Spectral Decomposition . . . 14
7.2 Singular Value Decomposition . . . 15
7.2.1 Relationship to Matrix Norm . . . 16

V Convex Optimization 17

8 Convex Sets 18
8.1 Affine and Convex Sets . . . 18
8.1.1 Affine Sets . . . 18

8.1.2 Convex Sets . . . 18
8.1.3 Cones . . . 18
8.2 Some Important Examples . . . 19
8.3 Generalized Inequalities . . . 19
8.3.1 Definition of Generalized Inequalities . . . 19
8.3.2 Properties of Generalized Inequalities . . . 20

9 Convex Optimization Problems **21**
9.1 Generalized Inequality Constraints . . . 21
9.1.1 Conic Form Problems . . . 21
9.1.2 Semidefinite Programming . . . 21
9.2 Vector Optimization . . . 21

10 Unconstrained Minimization **22**
10.1 Definition of Unconstrained Minimization . . . 22
10.2 General Descent Method . . . 24
10.3 Gradient Descent Method . . . 24
10.4 Steepest Descent Method . . . 24
10.5 Newton's Method . . . 24

11 Exercises for Convex Optimization **25**
11.1 Convex Sets . . . 25

VI Probability Theory **26**

12 Random Variables **27**
12.1 Probability Space . . . 27
12.2 Random Variables . . . 27
12.3 Distributions . . . 28
12.3.1 Definition of Distributions . . . 28
12.3.2 Properties of Distributions . . . 28
12.3.3 Families of Distributions . . . 29
12.4 Expected Value . . . 31
12.5 Independence . . . 31
12.5.1 Definition of Independence . . . 31
12.5.2 Sufficient Conditions for Independence . . . 32
12.5.3 Independence, Distribution, and Expectation . . . 32
12.5.4 Sums of Independent Random Variables . . . 33
12.6 Moments . . . 33
12.7 Characteristic Functions . . . 33
12.7.1 Definition of Characteristic Functions . . . 33
12.7.2 Properties of Characteristic Functions . . . 34
12.7.3 The Inversion Formula . . . 34
12.7.4 Moments and Derivatives . . . 34

13 Convergence of Random Variables **36**
13.1 Modes of Convergence . . . 36
13.1.1 Convergence in Mean . . . 36
13.1.2 Convergence in Probability . . . 36
13.1.3 Convergence in Distribution . . . 37
13.1.4 Almost Sure Convergence . . . 39
13.1.5 Convergence in Uninform . . . 39
13.1.6 Asymptotic Notation . . . 40
13.2 Relationships of Modes . . . 40

14 Law of Large Numbers **43**
14.1 Weak Law of Large Numbers . . . 43
14.2 Strong Law of Large Numbers . . . 44
14.2.1 Borel-Cantelli Lemmas . . . 44
14.2.2 Strong Law of Large Numbers . . . 45
14.3 Uniform Law of Large Numbers . . . 45

15 Central Limit Theorems **47**
15.1 Classic Central Limit Theorem . . . 47
15.1.1 The De Moivre-Laplace Theorem . . . 47
15.1.2 Classic Central Limit Theorem . . . 49
15.1.3 Berry-Esseen Theorem . . . 49
15.2 Central Limit Theorem for independent non-identical Random Variables . . . 49
15.3 Central Limit Theorem for Dependent Random Variables . . . 50

16 Exercises for Probability Theory and Examples **51**
16.1 Measure Theory . . . 51
16.2 Laws of Large Numbers . . . 51
16.3 Central Limit Theorems . . . 51

VII Stochastic Process **54**

17 Martingales **55**
17.1 Conditional Expectation . . . 55
17.2 Martingales . . . 55
17.3 Doob's Inequality . . . 56
17.4 Uniform Integrability . . . 57
17.5 Optional Stopping Theorems . . . 57

18 Markov Chains **58**
18.1 Markov Chain . . . 58
18.2 Markov Properties . . . 59
18.3 Recurrence and Transience . . . 60
18.4 Stationary Measures . . . 61
18.5 Asymptotic Behavior . . . 61
18.6 Ergodic Theorems . . . 61

19 Brownian Motion **63**
19.1 Markov Properties . . . 64
19.2 Martingales . . . 64
19.3 Sample Paths . . . 65
19.4 Itô Stochastic Calculus . . . 67

20 Exercises for Probability Theory and Examples **69**
20.1 Martingales . . . 69
20.2 Markov Chains . . . 69
20.3 Ergodic Theorems . . . 69
20.4 Brownian Motion . . . 69
20.5 Applications to Random Walk . . . 69
20.6 Multidimensional Brownian Motion . . . 69

VIII Random Matrix Theory **70**

21 Preliminary **71**
21.1 Empirical Spectral Measure . . . 71
21.2 Stieltjes Transform . . . 72
21.3 Matrix Equivalents . . . 74
21.4 Resolvent and Perturbation Identities . . . 74

22 Wigner Matrix **76**

23 Sample Covariance Matrix **77**
23.1 Eigenvalues and Singular Values . . . 77
23.2 Laguerre Orthogonal Ensemble . . . 78
23.3 Marčenko-Pastur Theorem . . . 82
23.4 Limits of Extreme Eigenvalues . . . 84

IX Statistics Inference **86**

24 Statistical Theory **87**
24.1 Populations and Samples . . . 87
24.2 Statistics . . . 87
24.2.1 Sufficient Statistics . . . 87
24.2.2 Complete Statistics . . . 87
24.3 Estimators . . . 88

25 Point Estimation **90**
25.1 Maximum Likelihood Estimator . . . 90
25.1.1 Classical Likelihood Estimator . . . 90
25.1.2 Modified Likelihood Estimator . . . 92
25.2 Minimum-Variance Unbiased Estimator . . . 94

26 Interval Estimation **97**
26.1 Confidence Interval . . . 97
26.2 Pivot . . . 97
26.3 Likelihood Interval . . . 97
26.4 Prediction Interval . . . 97
26.5 Tolerance Interval . . . 97

27 Testing Hypotheses **98**
27.1 Testing Hypotheses . . . 98
27.2 Parametric Tests . . . 98
27.3 Specific Tests . . . 98
27.3.1 Goodness of Fit . . . 98
27.3.2 Rank statistics . . . 98

28 Bayesian Inference **99**
28.1 Bayes Estimator . . . 99
28.1.1 Single-Prior Bayes . . . 100
28.1.2 Hierarchical Bayes . . . 102
28.1.3 Empirical Bayes . . . 103
28.1.4 Bayes Prediction . . . 103

29 Nonparametric Statistics **104**
29.1 Probability Distribution . . . 104
29.1.1 Cumulative Distribution Function . . . 104
29.1.2 Probability Density Function . . . 105
29.2 Kernel Methods . . . 105
29.2.1 Positive Definite Kernels . . . 105

X Computational Statistics **109**

30 Random Generator **110**
30.1 Uniform Random Number Generation . . . 110
30.2 Non-uniform Random Number Generation . . . 110
30.2.1 Inversion Method . . . 110
30.2.2 Rejection Sampling Method . . . 111
30.3 Markov Chain Monte Carlo . . . 111
30.3.1 Metropolis-Hastings Sampling . . . 111
30.3.2 Gibbs Sampling . . . 112

31 Monte Carlo Integration **113**
31.1 Monte Carlo Integration . . . 113
31.2 Importance Sampling . . . 113

32 Bootstrap **115**
32.1 Bootstrap Principle . . . 115
32.2 Standard Error Estimation . . . 116
32.3 Bias Estimation . . . 116
32.4 Confidence Interval Estimation . . . 116
32.5 Hypothesis Testing . . . 117

32.6 Jackknife . . . 117

XI Regression Analysis 119

33 Linear Regression 120

34 Generalized Linear Model 121
34.1 Introduction . . . 121
34.2 Binary Data . . . 122
34.3 Polytomous Data . . . 123
34.4 Count Data . . . 124

35 Quantile Regression 126

36 Survival Analysis 127
36.1 General Formulation . . . 127
36.2 Estimation of Survival Function . . . 128
36.3 Proportional Hazards Model . . . 129

37 Nonparametric Regression 130
37.1 Uniform Stability of Regularized Kernel Model . . . 130
37.1.1 Introduction . . . 130
37.1.2 Some Notations and Concepts . . . 131
37.1.3 Uniform Stability of Regularized Kernel Model . . . 131

38 High Dimensional Regression Analysis 136
38.1 Lasso for Linear Regression . . . 136
38.1.1 Numerical Algorithms . . . 136
38.1.2 Selection of the Tuning Parameter . . . 137
38.2 Theory for the Lasso . . . 137
38.3 Other Lasso-Type Estimators . . . 138
38.3.1 Adaptive Lasso . . . 138
38.3.2 Elastic Net . . . 138
38.3.3 Group Lasso . . . 138
38.3.4 Fused Lasso . . . 139
38.4 Nonconvex Penalties . . . 139
38.4.1 SCAD . . . 139
38.4.2 MCP . . . 139

XII Machine Learning 140

39 Support Vector Machine 141

40 Linear Discriminant Analysis 142

41 K-Nearest Neighbor 143

42 Decision Tree 144

43 Kalman Filter **145**

XIII Statistics Applications 146

44 Missingness Data **147**
44.1 The Problem of Missing Data . . . 147
44.1.1 Missingness Mechanisms . . . 148
44.1.2 Commonly Used Methods for Missing Data . . . 148
44.2 Likelihood-Based Inference with Missing Data . . . 149
44.2.1 Ignorable Missingness Mechanism . . . 149
44.2.2 Expectation-Maximization Algorithm . . . 151
44.3 Missing Not At Random Models . . . 156
44.3.1 Normal Models for MNAR Missing Data . . . 156

45 Treatment-effects Analysis **157**
45.1 Evaluations . . . 157
45.1.1 Average Treatment Effect . . . 157
45.1.2 Mann-Whitney Statistic . . . 157
45.1.3 Distribution-type Index . . . 157

46 Graphical Lasso **158**

47 Semi-supervised Learning **160**
47.1 Assumptions . . . 160

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

Definition 1.1.2 (Closed Function)

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

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

is closed.

Chapter 2

Differential Calculus

Chapter 3

Integral Calculus

Part II

Real Analysis

Chapter 4

Measure Theory

4.1 Semi-algebras, Algebras and Sigma-algebras

Definition 4.1.1 (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 4.1.2 (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 4.1.3 (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 4.1.1

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 4.1.1

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.

4.2 Measure

Definition 4.2.1 (Measure)

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

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

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

Definition 4.2.2 (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 5

Lebesgue Integration

5.1 Properties of the Integral

Theorem 5.1.1 (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:

$$\varphi\left(\int_\Omega f \,\mathrm{d}\mu\right) \leq \int_\Omega \varphi(f) \,\mathrm{d}\mu. \tag{5.1}$$

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 5.1.2 (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,

$$\int_\Omega |f \cdot g| \,\mathrm{d}\mu \leq \|f\|_p \|g\|_q. \tag{5.2}$$

Proof. □

Theorem 5.1.3 (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,

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p. \tag{5.3}$$

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

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

Therefore,

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

By Hölder's Inequality (5.1.2),

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

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

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

When $p = \infty$,

$$a$$

□

Theorem 5.1.4 (Bounded Convergence Theorem)

Theorem 5.1.5 (Fatou's Lemma)

Theorem 5.1.6 (Monotone Convergence Theorem)

5.2 Product Measures

Theorem 5.2.1 (Fubini's Theorem)

Part III

Functional Analysis

Part IV

Matrix Theory

Chapter 6

Matrix Norms

6.1 Matrix Norms Induced by Vector Norms

Chapter 7

Matrix Decompositions

7.1 Spectral Decomposition

Definition 7.1.1 (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

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \tag{7.1}$$

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

Definition 7.1.2 (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 as real and orthonormal. Thus a real symmetric matrix $\mathbf{A}$ can be decomposed as

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}' \tag{7.2}$$

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 $\mathbf{A}$.

7.2 Singular Value Decomposition

Definition 7.2.1 (Singular Value Decomposition)

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

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' \tag{7.3}$$

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 are called singular values, which 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

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

then

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

implying that $\mathbf{u}^{(i)}$ 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 $\mathbf{u}^{(i)}$

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

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

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

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

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

□

7.2.1 Relationship to Matrix Norm

Theorem 7.2.1

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

$$\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A}) \tag{7.4}$$

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,

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

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 7.2.2

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

$$\min_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \sigma_n(\mathbf{A}) \tag{7.5}$$

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 V

Convex Optimization

Chapter 8

Convex Sets

8.1 Affine and Convex Sets

8.1.1 Affine Sets

Definition 8.1.1 (Affine Set)

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

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

8.1.2 Convex Sets

Definition 8.1.2 (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 8.1.3 (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.

8.1.3 Cones

Definition 8.1.4 (Cone)

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

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

Definition 8.1.5 (Convex Cone)

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

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

8.2 Some Important Examples

Definition 8.2.1 (Hyperplane)

A hyperplane is defined to be

$$\{x | a^\top x = b\},$$

where $a \in \mathbb{R}^n, a \neq 0, b \in \mathbb{R}$.

Definition 8.2.2 (Halfspace)

A hyperplane is defined to be

$$\{x | a^\top x \leq b\},$$

where $a \in \mathbb{R}^n, a \neq 0, b \in \mathbb{R}$.

Definition 8.2.3 ((Euclidean) Ball)

A (Euclidean) ball in $\mathbb{R}^n$ 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 8.2.4 (Ellipsoid)

A Ellipsoid in $\mathbb{R}^n$ with center x_c is defined to be

$$\mathcal{E} = \{x | (x - x_c)^\top P^{-1} (x - x_c) \leq 1\} = \{x_c + Au | \|u_2\| \leq 1\},$$

where $P \in \mathbb{S}^n_{++}$ (symmetric positive definite).

8.3 Generalized Inequalities

8.3.1 Definition of Generalized Inequalities

Definition 8.3.1 (Proper Cone)

A cone $K \subseteq \mathbb{R}^n$ is said to be a proper cone, if

- K is convex.
- K is closed.
- K is solid (nonempty interior).
- K is pointed (contains no line).

Definition 8.3.2 (Generalized Inequalities)

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

$$y - x \in K, \tag{8.1}$$

which can be denoted by

$$x \preceq_K y \text{ or } y \succeq_K x. \tag{8.2}$$

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

$$y - x \in \text{ int } K, \tag{8.3}$$

which can be denoted by

$$x \prec_K y \text{ or } y \succ_K x. \tag{8.4}$$

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

8.3.2 Properties of Generalized Inequalities

Theorem 8.3.1 (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 9

Convex Optimization Problems

9.1 Generalized Inequality Constraints

Definition 9.1.1 (With Generalized Inequality Constraints)

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

$$\begin{array}{ll} \min_x & f_0(x) \\ \text{s.t.} & f_i(x) \preceq_{K_i} 0, \quad i = 1, \ldots, m \\ & Ax = b \end{array} \tag{9.1}$$

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

9.1.1 Conic Form Problems

Definition 9.1.2 (Conic Form Problem)

A conic form problem is defined to be

$$\begin{array}{ll} \min & c^T x \\ \text{s.t.} & Fx + g \preceq_K 0 \\ & Ax = b \end{array} \tag{9.2}$$

9.1.2 Semidefinite Programming

9.2 Vector Optimization

Chapter 10

Unconstrained Minimization

10.1 Definition of Unconstrained Minimization

Definition 10.1.1 (Unconstrained Minimization Problem)

The unconstrained minimization problem is defined to be

$$\min_{\mathbf{x}} f(\mathbf{x}) \tag{10.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ 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 $\mathbf{x}^*$ is optimal. if and only if

$$\nabla f(\mathbf{x}^*) = 0 \tag{10.2}$$

Solving (10.1) is equal to finding the solution of (10.2), thus (10.1) can be solved by analytic solution of (10.2) in a few cases, but usually can be solved by an iterative algorithm, i.e.,

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

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

Remark. 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. Any closed function (Definition 1.1.2)

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

$$\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^\top \mathbf{P}\mathbf{x} + \mathbf{q}'\mathbf{x} + r \tag{10.3}$$

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

$$\mathbf{P}\mathbf{x}^* + \mathbf{q} = \mathbf{0} \tag{10.4}$$

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 (10.4) is optimal for (10.3).
3. If (10.4) does not have a solution, then (10.3) is unbounded.

Proof.

1. Obviously.
2. 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. Since (10.4) 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,
$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \mathbf{x}^\top (\mathbf{A}'\mathbf{A})\mathbf{x} - 2(\mathbf{A}'\mathbf{b})'\mathbf{x} + \mathbf{b}'\mathbf{b}$$
The optimality condition is
$$\mathbf{A}'\mathbf{A}\mathbf{x}^* = \mathbf{A}'\mathbf{b}$$
are called the normal equations of the least-squares problem.

Example (Unconstrained Geometric Programming)**.** The unconstrained geometric program in convex form
$$\min_{\mathbf{x}} f(\mathbf{x}) = \log\left(\sum_{i=1}^{m} \exp(\mathbf{a}_i'\mathbf{x} + b_i)\right)$$
The optimality condition is
$$\nabla f(x^*) = \frac{\sum_{i=1}^{m} \exp(\mathbf{a}_i'\mathbf{x}^* + b_i)\,\mathbf{a}_i}{\sum_{j=1}^{m} \exp(\mathbf{a}_j'\mathbf{x}^* + b_j)} = \mathbf{0}$$
which has no analytical solution, so we must resort to an iterative algorithm. For this problem, $\operatorname{dom} f = \mathbb{R}^n$, so any point can be chosen as the initial point $\mathbf{x}^{(0)}$.

Example (Analytic Center of Linear Inequalities)**.** Consider the optimization problem
$$\min_{\mathbf{x}} f(x) = -\sum_{i=1}^{m} \log(\mathbf{b}_i - \mathbf{a}_i^T\mathbf{x})$$
where the domain of f is the open set
$$\operatorname{dom} f = \{\mathbf{x} \mid \mathbf{a}_i'\mathbf{x} < \mathbf{b}_i, i = 1, \ldots, m\}$$

Definition 10.1.2 (Strong Convexity)

10.2 General Descent Method

10.3 Gradient Descent Method

10.4 Steepest Descent Method

10.5 Newton's Method

Method	Descent Direction	Step Length	Features
Steepest			
Steepest (MG)			
Steepest (CD)			
Steepest (BB)			
Newton			
Newton (LM)			
Newton (Mixed)			
Quasi-Newton (SR1)			
Quasi-Newton (DFP)			
Quasi-Newton (BFGS)			
Quasi-Newton (LBFGS)			

Example (Extended Rosenbrock Function)**.**

$$\min_{\mathbf{x}} f(\mathbf{x}) = \sum_{i=1}^{n} r_i^2(\mathbf{x}) \tag{10.5}$$

where n is even, and

$$r_i(\mathbf{x}) = \begin{cases} 10(x_{2k} - x_{2k-1}^2), & i = 2k-1 \\ 1 - x_{2k-1}, & i = 2k \end{cases} \tag{10.6}$$

The minimum point is $\mathbf{x}^* = (1, 1, \ldots, 1)'$, the initial point is $\mathbf{x}_0 = (-1.2, 1, \ldots, -1.2, 1)'$.

Chapter 11

Exercises for Convex Optimization

11.1 Convex Sets

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

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

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

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 $x, y \in C$.

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

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

□

Part VI

Probability Theory

Chapter 12

Random Variables

12.1 Probability Space

Definition 12.1.1 (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.

12.2 Random Variables

Definition 12.2.1 (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,

$$X^{-1}(B) \equiv \{\omega : X(\omega) \in B\} \in \mathcal{F} \quad \forall B \in \mathcal{S}. \tag{12.1}$$

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

How to prove that functions are measurable?

Theorem 12.2.1

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.

12.3 Distributions

12.3.1 Definition of Distributions

Definition 12.3.1 (Distribution)

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

$$\mu(A) = P(X \in A) = P\left(X^{-1}(A)\right), \quad \text{for} A \in \mathcal{R}. \tag{12.2}$$

Definition 12.3.2 (Distribution Function)

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

$$F(x) = P(X \leq x). \tag{12.3}$$

Definition 12.3.3 (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.

12.3.2 Properties of Distributions

Theorem 12.3.1 (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 12.3.2

If F satisfies (1), (2), and (3) in Theorem 12.3.1, then it is the distribution function of some random variable.

Proof. □

Theorem 12.3.3

A distribution function has at most countably many discontinuities

Proof. □

12.3.3 Families of Distributions

Exponential Family

Definition 12.3.4 (Exponential Family)

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

$$f\left(y \mid \theta, \phi\right) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \tag{12.4}$$

Property. The exponential family has the following properties,

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

Proof. □

Table 12.1: 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}
Gamma	$\Gamma(\alpha,\beta)$	$\log\left(\frac{\alpha}{\beta}\right)$	1	$-\log(-\theta)$	1	$-\log\left(\Gamma(\alpha)\right)+(\alpha-1)\log(y)-y$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$

12.4 Expected Value

Definition 12.4.1 (Expectation)

Theorem 12.4.1 (Bounded Convergence theorem)

Theorem 12.4.2 (Fatou's Lemma)

If $X_n \geq 0$, then

$$\liminf_{n\to\infty} EX_n \geq E\left(\liminf_{n\to\infty} X_n\right). \tag{12.5}$$

Theorem 12.4.3 (Monotone Convergence theorem)

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

$$EX_n \uparrow EX. \tag{12.6}$$

Theorem 12.4.4 (Dominated Convergence theorem)

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

$$EX_n \to EX. \tag{12.7}$$

12.5 Independence

12.5.1 Definition of Independence

Definition 12.5.1 (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 $C, D \in \mathcal{R}$

$$P(X \in C, Y \in D) = P(X \in C)P(Y \in D). \tag{12.8}$$

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 12.5.1

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 12.5.2

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 12.5.2

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

$$P\left(\cap_{i=1}^n A_i\right) = \prod_{i=1}^n P\left(A_i\right). \tag{12.9}$$

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

$$P\left(\cap_{i=1}^n \left\{X_i \in B_i\right\}\right) = \prod_{i=1}^n P\left(X_i \in B_i\right). \tag{12.10}$$

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

$$P\left(\cap_{i \in I} A_i\right) = \prod_{i \in I} P\left(A_i\right). \tag{12.11}$$

12.5.2 Sufficient Conditions for Independence

12.5.3 Independence, Distribution, and Expectation

Theorem 12.5.3

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 12.5.4

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

$$E\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n EX_i \tag{12.12}$$

12.5.4 Sums of Independent Random Variables

Theorem 12.5.5 (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
$$P(X + Y \leq z) = \int F(z - y)\,\mathrm{d}G(y). \tag{12.13}$$
2. If X and Y are independent, X with density f and Y with distribution function G, then $X + Y$ has density
$$h(x) = \int f(x - y)\,\mathrm{d}G(y). \tag{12.14}$$
Suppose Y has density g, the last formula can be written as
$$h(x) = \int f(x - y)g(y)\,\mathrm{d}y. \tag{12.15}$$
3. If X and Y are independent, integral-valued random variables, then
$$P(X + Y = n) = \sum_m P(X = m)P(Y = n - m). \tag{12.16}$$

12.6 Moments

Lemma 12.6.1

If $Y > 0$ and $p > 0$, then
$$E(Y^p) = \int_0^\infty py^{p-1}P(Y > y)\,\mathrm{d}y. \tag{12.17}$$

12.7 Characteristic Functions

12.7.1 Definition of Characteristic Functions

Definition 12.7.1 (Characteristic Function)

If X is a random variable, we define its characteristic function (ch. f) by
$$\varphi(t) = E\left(e^{itX}\right) = E\left(\cos tX\right) + iE\left(\sin tX\right). \tag{12.18}$$

Remark. Euler Equation.

12.7.2 Properties of Characteristic Functions

Theorem 12.7.1 (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. □

12.7.3 The Inversion Formula

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

Theorem 12.7.2 (The Inversion Formula)

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

$$\lim_{T\to\infty} (2\pi)^{-1} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)dt = \mu(a,b) + \frac{1}{2}\mu(\{a,b\}) \tag{12.19}$$

Proof. □

Theorem 12.7.3

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

$$f(y) = \frac{1}{2\pi} \int e^{-ity}\varphi(t)\,\mathrm{d}t. \tag{12.20}$$

Proof. □

12.7.4 Moments and Derivatives

Theorem 12.7.4

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

$$\varphi^{(n)}(t) = \int (ix)^n e^{itx} \mu(\mathrm{d}x). \tag{12.21}$$

Theorem 12.7.5

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

$$\varphi(t) = 1 + itEX - t^2 E\left(X^2\right)/2 + o\left(t^2\right). \tag{12.22}$$

Theorem 12.7.6

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

$$E|X|^2 < \infty. \tag{12.23}$$

Chapter 13

Convergence of Random Variables

13.1 Modes of Convergence

13.1.1 Convergence in Mean

Definition 13.1.1 (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

$$X_n \xrightarrow{L^r} X. \tag{13.1}$$

13.1.2 Convergence in Probability

Definition 13.1.2 (Convergence in Probability)

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

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} P\left(|X_n - X| > \varepsilon\right) = 0. \tag{13.2}$$

Convergence in probability is denoted by

$$X_n \xrightarrow{p} X. \tag{13.3}$$

13.1.3 Convergence in Distribution

Definition 13.1.3 (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

$$\lim_{n\to\infty} F_n(x) = F(x), \tag{13.4}$$

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

$$X_n \xrightarrow{d} X, \text{ or } X_n \Rightarrow X. \tag{13.5}$$

- 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 Scheff's theorem, convergence of the probability density functions implies convergence in distribution.

Theorem 13.1.1 (Portmanteau Lemma)

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

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

Proof. □

Continuous Mapping Theorem

Theorem 13.1.2 (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,

$$\begin{aligned} X_n \xrightarrow{d} X &\Rightarrow g(X_n) \xrightarrow{\text{d}} g(X), \\ X_n \xrightarrow{p} X &\Rightarrow g(X_n) \xrightarrow{p} g(X), \\ X_n \xrightarrow{a.s.} X &\Rightarrow g(X_n) \xrightarrow{a.s.} g(X). \end{aligned} \tag{13.6}$$

If in addition g is bounded, then

$$Eg(X_n) \to Eg(X). \tag{13.7}$$

Proof. □

Slutsky's Theorem

Theorem 13.1.3 (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 $X_n Y_n \xrightarrow{d} XY$.

The Delta Methods

Theorem 13.1.4 (Delta Method)

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

$$\sqrt{n}\left(X_n - \theta\right) \xrightarrow{d} N\left(0, \sigma^2\right) \tag{13.8}$$

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

$$\sqrt{n}\left[g\left(X_n\right) - g(\theta)\right] \xrightarrow{d} N\left(0, \sigma^2 \cdot \left[g'(\theta)\right]^2\right) \tag{13.9}$$

Proof. 1. Under the assumption that $g'(\theta)$ is continuous.
Since, $g'(\theta)$ exists, with the first-order Taylor Approximation, that

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

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 (13.1.2),

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

and,

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

$$\sqrt{n}\left(X_n - \theta\right) \xrightarrow{d} N\left(0, \sigma^2\right)$$

by Slutsky's Theorem (13.1.3),

$$\sqrt{n}\left[g\left(X_n\right) - g(\theta)\right] \xrightarrow{d} N\left(0, \sigma^2 \cdot \left[g'(\theta)\right]^2\right)$$

□

Theorem 13.1.5 (Second-order Delta Method)

Remark. We can approximate the moments of a function $f(\cdot)$ of a random variable X using Taylor expansions, provided that $f(\cdot)$ is sufficiently differentiable and that the moments of X are finite. Suppose $\mu = \mathrm{E}(X)$, and $\sigma^2 = \mathrm{Var}(X)$, with the Taylor expansions for the functions of random variables,

$$f(X) = f[\mu + (X - \mu)] \approx f(\mu) + f'(\mu)(X - \mu) \tag{13.10}$$

Thus,

$$\mathrm{E}[f(X)] \approx \mathrm{E}[f(\mu)], \quad \mathrm{Var}[f(X)] \approx [f'(\mu)]^2 \cdot \sigma^2 \tag{13.11}$$

Lèvy' s Continuity Theorem

Theorem 13.1.6 (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. □

Cramér-Wold Theorem

Theorem 13.1.7 (Cramér-Wold Theorem)

13.1.4 Almost Sure Convergence

Definition 13.1.4 (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

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1. \tag{13.12}$$

Almost sure convergence is denoted by

$$X_n \xrightarrow{a.s.} X. \tag{13.13}$$

Remark.

13.1.5 Convergence in Uninform

Definition 13.1.5 (Convergence in Uninform)

13.1.6 Asymptotic Notation

Definition 13.1.6

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

$$\frac{A_n}{B_n} \xrightarrow{p} 0. \tag{13.14}$$

Smaller order in probability is denoted by

$$A_n = o_p(B_n). \tag{13.15}$$

Particularly,

$$A_n = o_p(1) \iff A_n \xrightarrow{p} 0. \tag{13.16}$$

Definition 13.1.7

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

$$\forall \varepsilon > 0 \; \exists M_\varepsilon, \quad \lim_{n\to\infty} P\left(|A_n| \leq M_\varepsilon |B_n|\right) \geq 1 - \varepsilon. \tag{13.17}$$

Smaller order than or equal to in probability is denoted by

$$A_n = O_p(B_n). \tag{13.18}$$

Definition 13.1.8

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

$$\forall \varepsilon > 0 \; \exists m_\varepsilon < M_\varepsilon, \quad \lim_{n\to\infty} P\left(m_\varepsilon < \frac{|A_n|}{|B_n|} < M_\varepsilon\right) \geq 1 - \varepsilon. \tag{13.19}$$

The same order in probability is denoted by

$$A_n \asymp_p B_n. \tag{13.20}$$

13.2 Relationships of Modes

Lemma 13.2.1

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

$$Z_n \xrightarrow{p} 0. \tag{13.21}$$

Proof. □

Theorem 13.2.1

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

$$X_n \xrightarrow{d} X, \tag{13.22}$$

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

$$X_n \xrightarrow{p} c. \tag{13.23}$$

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

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

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

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

then, by the union bound,

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

So, we got

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

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

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

By the property of distribution (Theorem 12.3.1), 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 $\forall \varepsilon > 0$,

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

Therefore,

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

which means

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

□

Theorem 13.2.2

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

$$X_n \xrightarrow{p} X. \tag{13.24}$$

Proof. □

Theorem 13.2.3

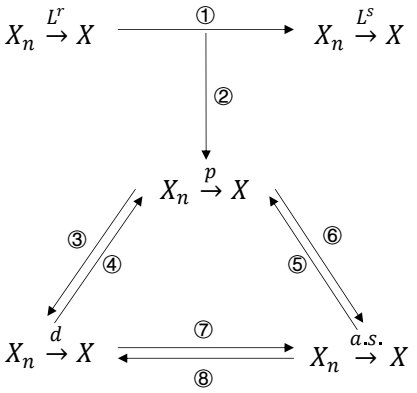
$X_n \xrightarrow{p} X$ if and only if for all subsequence $X_{n(m)}$ exists a further subsequence $X_{n(m_k)}$, such that

$$X_{n(m_k)} \xrightarrow{a.s.} X. \tag{13.25}$$

Lemma 13.2.2

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

$$Y_n \xrightarrow{a.s.} Y_\infty. \tag{13.26}$$



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

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

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

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

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

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

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

⑧ $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{d} X$

Figure 13.1: Relations of Convergence of Random Variables

Chapter 14

Law of Large Numbers

14.1 Weak Law of Large Numbers

Theorem 14.1.1 (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

$$S_n/n \xrightarrow{L^2} \mu, \quad S_n/n \xrightarrow{p} \mu. \tag{14.1}$$

Proof. □

Theorem 14.1.2 (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

$$\frac{S_n - \mu_n}{b_n} \xrightarrow{p} 0. \tag{14.2}$$

Proof. □

Theorem 14.1.3 (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

$$\frac{S_n - a_n}{b_n} \xrightarrow{p} 0. \tag{14.3}$$

Proof. □

Theorem 14.1.4 (Weak Law of Large Numbers by Feller)

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

$$\lim_{x\to 0} xP(|X_i| > x) = 0. \tag{14.4}$$

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

$$S_n/n - \mu_n \xrightarrow{p} 0. \tag{14.5}$$

Proof. □

Theorem 14.1.5 (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

$$S_n/n \xrightarrow{p} \mu. \tag{14.6}$$

Proof. □

Remark. $E|X_i| = \infty$

14.2 Strong Law of Large Numbers

14.2.1 Borel-Cantelli Lemmas

Lemma 14.2.1 (Borel-Cantelli Lemma)

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

$$P(A_n \text{ i.o. }) = 0. \tag{14.7}$$

Lemma 14.2.2 (The Second Borel-Cantelli Lemma)

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

$$P(A_n \text{ i.o. }) = 1. \tag{14.8}$$

Corollary 14.2.1

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

$$\sum_{n=1}^{\infty} P(A_n) = \infty, \tag{14.9}$$

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

Proof. □

14.2.2 Strong Law of Large Numbers

Theorem 14.2.1 (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

$$S_n/n \xrightarrow{a.s.} \mu. \tag{14.10}$$

14.3 Uniform Law of Large Numbers

Theorem 14.3.1 (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

$$\sup_{\theta \in \Theta} \left| n^{-1} \sum_{i=1}^{n} g(X_i, \theta) - Eg(X_i, \theta) \right| \xrightarrow{p} 0. \tag{14.11}$$

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

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

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

Since

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

by the Weak Law of Large Numbers (Theorem 14.1.5),

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

By analogous argument,

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

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

Chapter 15

Central Limit Theorems

15.1 Classic Central Limit Theorem

15.1.1 The De Moivre-Laplace Theorem

Lemma 15.1.1 (Stirling's Formula)

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \text{ as } n \to \infty. \tag{15.1}$$

Proof. □

Lemma 15.1.2

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

$$(1 + c_j)^{a_j} \to e^{\lambda}. \tag{15.2}$$

Proof. □

Theorem 15.1.1 (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 $m \to \infty$

$$P\left(a \leq S_m/\sqrt{m} \leq b\right) \to \int_a^b (2\pi)^{-1/2} e^{-x^2/2} \, \mathrm{d}x. \tag{15.3}$$

Proof. If n and k and integers

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

By lemma 15.1.1, we have

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

Hence,

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

Let $2k = x\sqrt{2n}$, i.e., $k = x\sqrt{\frac{n}{2}}$. By lemma 15.1.2, 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 $m \to \infty$,

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

□

15.1.2 Classic Central Limit Theorem

Theorem 15.1.2 (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

$$\frac{S_n - n\mu}{\sigma n^{\frac{1}{2}}} \xrightarrow{d} \chi, \tag{15.4}$$

where χ has the standard normal distribution.

Proof. □

Theorem 15.1.3 (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 $n \to \infty$.

15.1.3 Berry-Esseen Theorem

Theorem 15.1.4 (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),

$$|G_n(x) - \Phi(x)| \leq \frac{C}{\sqrt{n}} \frac{E\,|X_1 - \mu|^3}{\sigma^3}, \quad \forall x. \tag{15.5}$$

Corollary 15.1.1

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.

15.2 Central Limit Theorem for independent non-identical Random Variables

Theorem 15.2.1 (The Liapounov Central Limit Theorem)

15.3 Central Limit Theorem for Dependent Random Variables

Chapter 16

Exercises for Probability Theory and Examples

16.1 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 $A_i = [1, 2 - \frac{1}{i}] \in \mathcal{F}_i$,

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

□

16.2 Laws of Large Numbers

16.3 Central Limit Theorems

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

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

Proof. Let $Y_n \overset{d}{=} X_n, 1 \le n \le \infty$ with $Y_n \overset{a.s.}{\to} Y_\infty$ (Lemma 13.2.2). Since $g \ge 0$ be continuous, $g(Y_n) \overset{a.s.}{\to} g(Y_\infty)$ and $g(Y_n) \ge 0$ (Theorem 13.1.2), and the Fatou's Lemma (12.4.2) 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 13.2.2).

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

By Continuity Mapping Theorem (13.1.2), $\lim_{n\to\infty} \left|E\left(h(X_n) - h(X)\right) I_{x\in[-M,M]}\right| = 0$. Since

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

and by Exercise 16.3

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

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

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 VII

Stochastic Process

Chapter 17

Martingales

17.1 Conditional Expectation

Definition 17.1.1 (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 a 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.

17.2 Martingales

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

Definition 17.2.1 (Martingale)

A sequence $\{X_n\}$ of real-valued random variables is said to be a martingale for $\mathcal{F}_n$, 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.,

$$E(X_{n+1} \mid \mathcal{F}_n) = X_n, \quad \forall n \tag{17.1}$$

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:

$$P\left(\max_{1 \leq m \leq n} |X_m| \geq x\right) \leq x^{-2} \operatorname{var}(X_n) \tag{17.2}$$

Theorem 17.2.1 (Orthogonality of Martingale Increments)

Theorem 17.2.2 (Conditional Variance Formula)

Definition 17.2.2 (Predictable Sequence)

Definition 17.2.3 (Stopping Time)

Theorem 17.2.3 (Martingale Convergence Theorem)

17.3 Doob's Inequality

Theorem 17.3.1 (Doob's Decomposition)

Theorem 17.3.2 (Doob's Inequality)

Theorem 17.3.3 (L^p Maximum Inequality)

17.4 Uniform Integrability

17.5 Optional Stopping Theorems

Chapter 18

Markov Chains

18.1 Markov Chain

Definition 18.1.1 (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

$$P\left(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots X_0 = i_0\right) = P\left(X_{n+1} = j \mid X_n = i\right) \tag{18.1}$$

and the transition probability is

$$p(i,j) = P\left(X_{n+1} = j \mid X_n = i\right) \tag{18.2}$$

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 18.1.2 (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

$$Z_{n+1} = \begin{cases} \xi_1^{n+1} + \cdots + \xi_{Z_n}^{n+1} & Z_n > 0 \\ 0 & Z_n = 0 \end{cases} \tag{18.3}$$

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

Definition 18.1.3 (Transition Probability)

A function $p : S \times \mathcal{S} \to \mathbb{R}$ 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 18.1.4 (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

$$P\left(X_{n+1} \in B \mid \mathcal{F}_n\right) = p\left(X_n, B\right) \tag{18.4}$$

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

$$P\left(X_j \in B_j, 0 \le j \le n\right) = \int_{B_0} \mu\left(\mathrm{d}x_0\right) \int_{B_1} p\left(x_0, \mathrm{d}x_1\right) \cdots \int_{B_n} p\left(x_{n-1}, \mathrm{d}x_n\right) F \tag{18.5}$$

18.2 Markov Properties

Definition 18.2.1 (Shift Operator)

Theorem 18.2.1 (Markov Property)

Corollary 18.2.1 (Chapman-Kolmogorov Equation)

Theorem 18.2.2 (Strong Markov Property)

18.3 Recurrence and Transience

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

$$T_y^k = \inf\left\{n > T_y^{k-1} : X_n = y\right\} \tag{18.6}$$

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

$$\rho_{xy} = P_x\left(T_y < \infty\right) \tag{18.7}$$

and we have

$$P_x\left(T_y^k < \infty\right) = \rho_{xy}\rho_{yy}^{k-1} \tag{18.8}$$

Proof. □

Let

$$N(y) = \sum_{n=1}^{\infty} 1_{(X_n = y)} \tag{18.9}$$

be the number of visits to y at positive times.

Definition 18.3.1 (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 18.3.2

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 18.3.3 (Closed State Set)

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

$$x \in C, \rho_{xy} > 0 \Rightarrow y \in C. \tag{18.10}$$

Definition 18.3.4 (Irreducible State Set)

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

$$x, y \in D \Rightarrow \rho_{xy} > 0. \tag{18.11}$$

Theorem 18.3.1

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 18.3.2

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

18.4 Stationary Measures

18.5 Asymptotic Behavior

18.6 Ergodic Theorems

Definition 18.6.1 (Stationary Sequence)

Theorem 18.6.1 (Ergodic Theorem)

Example.

Chapter 19

Brownian Motion

Definition 19.0.1 (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 $A \in \mathbb{R}$,
$$P(B(s+t) - B(s) \in A) = \int_A (2\pi t)^{-1/2} \exp\left(-x^2/2t\right) \mathrm{d}x \tag{19.1}$$
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 is as follows,

Definition 19.0.2 (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.

19.1 Markov Properties

19.2 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 19.2.1 (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 $\mathcal{F}_t$

19.3 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

$$\Delta_i^n B = B\left(t_{i+1}^n\right) - B\left(t_i^n\right) \tag{19.2}$$

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

Theorem 19.3.1

$$\lim_{n\to\infty} \sum_{i=0}^{n-1} \left(\Delta_i^n B\right)^2 = T \quad \text{in} \quad L^2 \tag{19.3}$$

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 19.3.1 (Variation)

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

$$\limsup_{\Delta t \to 0} \sum_{i=0}^{n-1} |f(t_{i+1}) - f(t_i)| \tag{19.4}$$

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

$$\Delta t = \max_{i=0,\ldots,n-1} |t_{i+1} - t_i| \tag{19.5}$$

Theorem 19.3.2

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

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

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

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

By Theorem 19.3.1, 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

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

hence,

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

while

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

□

19.4 Itô Stochastic Calculus

Definition 19.4.1 (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

$$I_T(f) = I\left(1_{[0,T)} f\right) \tag{19.6}$$

which can be denoted by

$$\int_0^T f(t)\,\mathrm{d}B(t) \tag{19.7}$$

Property. The Itô Stochastic Integral has the following properties:

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

$$\int_0^t (\alpha f(r) + \beta g(r))\,\mathrm{d}B(r) = \alpha \int_0^t f(r)\,\mathrm{d}B(r) + \beta \int_0^t g(r)\,\mathrm{d}B(r) \tag{19.8}$$

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

$$E\left(\left|\int_0^t f(r)\,\mathrm{d}B(r)\right|^2\right) = E\left(\int_0^t |f(r)|^2\,\mathrm{d}r\right) \tag{19.9}$$

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

$$E\left(\int_0^t f(r)\,\mathrm{d}B(r) \mid \mathcal{F}_s\right) = \int_0^s f(r)\,\mathrm{d}B(r) \tag{19.10}$$

Proof. □

Definition 19.4.2 (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

$$\xi(T) = \xi(0) + \int_0^T a(t)\,\mathrm{d}t + \int_0^T b(t)\,\mathrm{d}B(t) \quad \text{a.s.} \tag{19.11}$$

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

$$\int_0^T |a(t)|\,\mathrm{d}t < \infty \quad \text{a.s.} \tag{19.12}$$

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

$$\mathrm{d}\xi(t) = a(t)\,\mathrm{d}t + b(t)\,\mathrm{d}B(t) \tag{19.13}$$

Remark. The class of all adapted processes $a(t)$ satisfying 19.12 for some $T > 0$ will be denoted by $\mathcal{L}_T^1$.

Theorem 19.4.1 (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

$$\begin{aligned}\mathrm{d}F(t,\xi(t)) = &\left(F'_t(t,\xi(t)) + F'_x(t,\xi(t))a(t) + \frac{1}{2}F''_{xx}(t,\xi(t))b(t)^2\right)\mathrm{d}t \\ &+ F'_x(t,\xi(t))b(t)\,\mathrm{d}B(t)\end{aligned} \tag{19.14}$$

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

$$\mathrm{d}F(t,B(t)) = \left(F'_t(t,B(t)) + \frac{1}{2}F''_{xx}(t,B(t))\right)\mathrm{d}t + F'_x(t,B(t))\,\mathrm{d}B(t) \tag{19.15}$$

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 20

Exercises for Probability Theory and Examples

20.1 Martingales

20.2 Markov Chains

20.3 Ergodic Theorems

20.4 Brownian Motion

20.5 Applications to Random Walk

20.6 Multidimensional Brownian Motion

Part VIII

Random Matrix Theory

Chapter 21

Preliminary

21.1 Empirical Spectral Measure

Definition 21.1.1 (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.,

$$\mu_{\mathbf{M}} := \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{M})} \tag{21.1}$$

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

$$\det(\mathbf{M}) = \prod_{i=1}^{n} \lambda_i = \exp\left(n \int_0^\infty \log x \mu_{\mathbf{M}}(\mathrm{d}x)\right) \tag{21.2}$$

21.2 Stieltjes Transform

Definition 21.2.1 (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

$$\mathbf{Q}_{\mathbf{M}}(z) := (\mathbf{M} - z\mathbf{I}_n)^{-1} \tag{21.3}$$

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

Definition 21.2.2 (Stieltjes Transform)

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

$$m_\mu(z) := \int \frac{1}{t-z}\mu\,(\mathrm{d}t) \tag{21.4}$$

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

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 21.2.1 (Inverse Stieltjes Transform)

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

$$\mu\left([a,b]\right) = \frac{1}{\pi}\lim_{y\downarrow 0}\int_a^b \Im\left[m_\mu(x+\imath y)\right]\mathrm{d}x \tag{21.5}$$

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

$$f(x) = \frac{1}{\pi}\lim_{y\downarrow 0}\Im\left[m_\mu(x+\imath y)\right] \tag{21.6}$$

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

$$\mu(\{x\}) = \lim_{y\downarrow 0} -\imath y m_\mu(x+\imath y) \tag{21.7}$$

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 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

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

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

$$m_{\mu_{\mathbf{M}}}(z) = \frac{1}{n}\sum_{i=1}^n \int \frac{\delta_{\lambda_i(\mathbf{M})}(t)}{t-z} = \frac{1}{n}\sum_{i=1}^n \frac{1}{\lambda_i(\mathbf{M})-z} = \frac{1}{n}\operatorname{tr}\mathbf{Q}_{\mathbf{M}}(z) \tag{21.8}$$

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

$$\frac{1}{n}\sum_{i=1}^n f\left(\lambda_i(\mathbf{M})\right) = -\frac{1}{2\pi\imath n}\oint_\Gamma f(z)\operatorname{tr}\left(\mathbf{Q}_{\mathbf{M}}(z)\right)\mathrm{d}z = -\frac{1}{2\pi\imath}\oint_\Gamma f(z)m_{\mu_{\mathbf{M}}}(z)\,\mathrm{d}z \tag{21.9}$$

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 $\lambda_i(\mathbf{M})$).

21.3 Matrix Equivalents

Definition 21.3.1 (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),

$$\frac{1}{n} \operatorname{tr} \mathbf{A}(\mathbf{Q} - \overline{\mathbf{Q}}) \to 0, \quad \mathbf{a}'(\mathbf{Q} - \overline{\mathbf{Q}})\mathbf{b} \to 0, \quad \text{as } n \to \infty \tag{21.10}$$

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 21.3.2 (Matrix Equivalents)

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

$$\mathbf{x} \leftrightarrow \mathbf{Y} \tag{21.11}$$

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.

21.4 Resolvent and Perturbation Identities

Lemma 21.4.1 (Resolvent Identity)

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

$$\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1} (\mathbf{B} - \mathbf{A}) \mathbf{B}^{-1} \tag{21.12}$$

Lemma 21.4.2 (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

$$(\mathbf{A} + \mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}}{1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}} \tag{21.13}$$

or,

$$(\mathbf{A} + \mathbf{u}\mathbf{v}')^{-1}\mathbf{u} = \frac{\mathbf{A}^{-1}\mathbf{u}}{1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}} \tag{21.14}$$

Lemma 21.4.3 (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 $k \geq 1$

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

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

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

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

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

Chapter 22

Wigner Matrix

Chapter 23

Sample Covariance Matrix

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

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

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

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

23.1 Eigenvalues and Singular Values

Theorem 23.1.1

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. □

23.2 Laguerre Orthogonal Ensemble

Definition 23.2.1 (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 means:

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

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

$$\mathbf{M} = \mathbb{X}'\mathbb{X} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \tag{23.1}$$

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 23.2.1

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

$$f\left(\mathbf{M}\right) = \frac{1}{2^{np/2}\left[\det\left(\mathbf{\Sigma}\right)\right]^{n/2}\Gamma_p\left(\frac{n}{2}\right)} \det\left(\mathbf{M}\right)^{(n-p-1)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{M}\right)\right] \tag{23.2}$$

concerning the 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. Especially, 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 density function of $\widehat{\mathbf{\Sigma}}$ can be derived from (23.2), 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,

$$f\left(\widehat{\mathbf{\Sigma}}\right) = \frac{m_n^{-n(m_n-n-1)/2+1}}{2^{m_n n/2}\Gamma_n\left(\frac{m_n}{2}\right)} \det\left(\widehat{\mathbf{\Sigma}}\right)^{(m_n-n-1)/2} \exp\left[-\frac{m_n}{2}\operatorname{tr}\left(\widehat{\mathbf{\Sigma}}\right)\right] \tag{23.3}$$

Theorem 23.2.2

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

$$p(\mathbf{\Lambda}) = \widetilde{Q}_{m_n,n}^{-1} \exp\left(-\frac{m_n}{2}\sum_{k=1}^{n}\lambda_k\right)\prod_{k=1}^{n}\lambda_k^{(m_n-n-1)/2}\prod_{i<j}|\lambda_i-\lambda_j| \tag{23.4}$$

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,

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

thus,

$$[\det(\mathbf{Q})]^{-m_n/2} = \left[\det\left(\mathbf{I}_n - \frac{2\imath}{m_n}\mathbf{P}\right)\right]^{-m_n/2} = \prod_{k=1}^{n}\left(1-\frac{2\imath}{m_n}p_k\right)^{-m_n/2} \tag{23.5}$$

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

Then, we will show that the characteristic function of (23.4) coincides with the above function. By the Wishart distribution, the probability density of the real symmetric and positive definite random matrix $\widehat{\mathbf{\Sigma}}$ is

$$\widetilde{Q}_{m_n,n}^{-1} \exp\left[-\frac{m_n}{2}\operatorname{tr}\left(\widehat{\mathbf{\Sigma}}\right)\right]\left[\det\left(\widehat{\mathbf{\Sigma}}\right)\right]^{(m_n-n-1)/2}\mathrm{d}\widehat{\mathbf{\Sigma}} \tag{23.6}$$

where $\widetilde{Q}_{m_n,n}$ is the normalization constant. Then, the characteristic function of (23.6), 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

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

where H_1 is the normalized Haar measure of $O(n)$, and the integration over $\mathbf{\Lambda}$ and $O\in O(n)$ are independent. Since the orthogonal invariance of the density of (23.6), and the characteristic function is

$$Q_{m_n,n}^{-1}\int_{(\mathbb{R}_+)^n}\exp\left[\sum_{k=1}^n\left(\imath p_k-\frac{m_n}{2}\right)\lambda_k\right]\prod_{k=1}^n\lambda_k^{(m_n-n-1)/2}\prod_{i<j}|\lambda_i-\lambda_j|\,\mathrm{d}\mathbf{\Lambda} \tag{23.7}$$

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

If we viewed (23.5) and (23.7) 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 (23.5) since this is a set of the uniqueness of both (23.5) and eqrefeq:characteristic-function-wishart analytic functions, we have

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

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,

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

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

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

The integral here is equal to $\widetilde{Q}_n$, the normalization constant of the probability measure (23.6). 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 (23.5). Thus the probability law of the Wishart matrices of $\mathbf{\Sigma}$ given by (23.6) implies that the corresponding joint probability density of eigenvalues is given by (23.4) for $\mathbf{\Sigma}$. □

Definition 23.2.2 (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

$$p(\mathbf{\Lambda}) = K_{\alpha,\beta} \exp\left(-\frac{\beta}{2}\sum_{k=1}^{n} \lambda_k\right) \prod_{k=1}^{n} \lambda_k^{\frac{\alpha\beta}{2}} \prod_{i<j} |\lambda_i - \lambda_j|^{\beta} \tag{23.8}$$

where

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

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

And Equation (23.8) can be written in the standard Boltzmann-Gibbs form, that,

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

where

$$E(\mathbf{\Lambda}) = \frac{1}{2}\sum_{k=1}^{n} (\lambda_k - \alpha \log \lambda_k) - \frac{1}{2}\sum_{i\neq j} |\lambda_i - \lambda_j| \tag{23.9}$$

Remark. For the (23.4), which can be written as (23.8) form, that,

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

where $\beta = 1$ and

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

23.3 Marčenko-Pastur Theorem

In this section, we will investigate 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.

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

Theorem 23.3.1 (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

$$\mathbf{Q}(z) \leftrightarrow \overline{\mathbf{Q}}(z), \quad \overline{\mathbf{Q}}(z) = m(z)\mathbf{I}_n$$

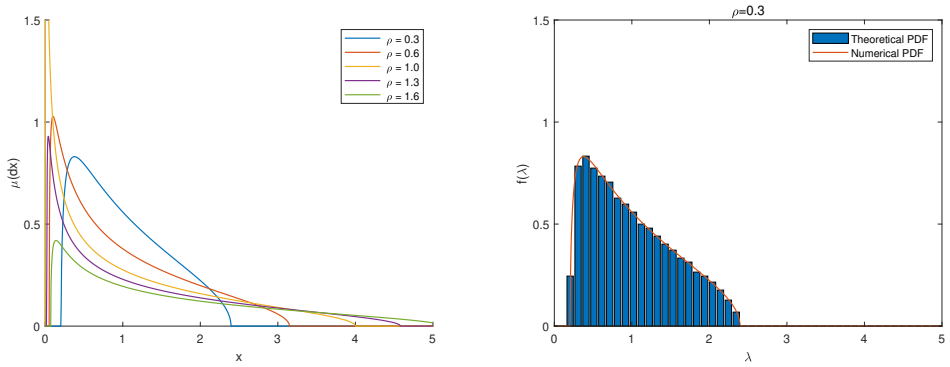
with $(z, m(z))$ the unique solution in $\mathcal{Z}\left(\mathbb{C} \backslash \left[(1-\sqrt{\rho})^2, (1+\sqrt{\rho})^2\right]\right)$ be

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

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.



(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 23.1: Illustrations of the Marčenko-Pastur Theorem

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

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 21.4.1, we have

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

Thus, we turn to prove that,

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

By Lemma 21.4.2, we have

$$\mathbf{Q}(z)\mathbf{x}_i=\frac{\mathbf{Q}_{-i}(z)\mathbf{x}_i}{1+\frac{1}{m_n}\mathbf{x}_i^\top\mathbf{Q}_{-i}(z)\mathbf{x}_i}$$

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 21.4.3, we have

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

Hence, we need to prove the approximation that

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

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 21.8, 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

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

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 21.2.1, 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 $m(z)$ is 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 $(x_i)_{1\leq i\leq n}$ are i.i.d. with finite second moment.
2. if $\mathbf{x}$ is isotropic and log-concave[1] random vector.

23.4 Limits of Extreme Eigenvalues

The weak convergence in Theorem 23.3.1 does not provide much information at the edge on the behavior of the extremal atoms, and what one can extract is that

$$\limsup_{n\to\infty}\lambda_{\min}\left(\widehat{\mathbf{\Sigma}}\right) \leq (1-\sqrt{\rho})^2 \leq (1+\sqrt{\rho})^2 \leq \liminf_{n\to\infty}\lambda_{\max}\left(\widehat{\mathbf{\Sigma}}\right) \quad \text{a.s.} \tag{23.10}$$

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

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

The weak convergence above does not provide much information at the edge on the behavior of the extremal atoms. Now, we have more exact result, that if $(X_{n,k})_{n\geq 1, 1\leq k\leq n}$ are i.i.d. with finite fourth moment then,

$$(1-\sqrt{\rho})^2 = \lim_{n\to\infty} \lambda_{\min}\left(\widehat{\mathbf{\Sigma}}\right) \leq \lim_{n\to\infty} \lambda_{\max}\left(\widehat{\mathbf{\Sigma}}\right) = (1+\sqrt{\rho})^2 \quad \text{a.s.} \tag{23.11}$$

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

Remark. The convergence of the smallest eigenvalue in the left-hand side of (23.11) holds if $(x_i)_{1\leq i\leq n}$ are i.i.d. with finite second moment.

Theorem 23.4.1

If $\bar{\rho} < 1$ (in particular $m_n > n$ for $n \gg 1$) and if the centered isotropic random vector $\mathbf{x}$ is log-concave or if $(x_i)_{1\leq i\leq n}$ are i.i.d. then

$$\liminf_{n\to\infty} \frac{E\left(\lambda_{\min}\left(\mathbf{A}_n\right)\right)}{\left(\sqrt{m_n}-\sqrt{n}\right)^2} \geq 1 \tag{23.12}$$

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

$$\lambda_{\min}\left(\widehat{\mathbf{\Sigma}}_n\right) \xrightarrow{p} (1-\sqrt{\rho})^2 \quad \text{as } n\to\infty \tag{23.13}$$

Proof. □

Theorem 23.4.2

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

$$\limsup_{n\to\infty} \frac{E\left(\lambda_{\max}\left(\mathbf{A}_n\right)\right)}{\left(\sqrt{m_n}+\sqrt{n}\right)^2} \leq 1 \tag{23.14}$$

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

$$\lambda_{\max}\left(\widehat{\mathbf{\Sigma}}_n\right) \xrightarrow{p} (1+\sqrt{\rho})^2 \quad \text{as } n\to\infty \tag{23.15}$$

Proof. □

Part IX

Statistics Inference

Chapter 24

Statistical Theory

24.1 Populations and Samples

24.2 Statistics

24.2.1 Sufficient Statistics

Definition 24.2.1 (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 24.2.1 (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. □

24.2.2 Complete Statistics

Definition 24.2.2 (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.

24.3 Estimators

Definition 24.3.1 (Estimator)

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

$$\delta : \mathbf{X} \to \mathbb{R}. \tag{24.1}$$

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

Unbiasedness

Definition 24.3.2 (Unbiasedness)

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

$$E\hat{\theta} = \theta, \quad \forall \theta \in \Theta. \tag{24.2}$$

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

Example (Nonexistence of Unbiased Estimator)**.**

Consistency

Definition 24.3.3 (Consistency)

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

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = 0, \quad \forall \varepsilon > 0, \tag{24.3}$$

that is,

$$\hat{\theta}_n \xrightarrow{p} \theta. \tag{24.4}$$

Example (Consistency of Sample Moments)**.**

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

Asymptotic Normality

Definition 24.3.4 (Asymptotic Normality)

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

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N\left(0, \sigma_\theta^2\right). \tag{24.5}$$

Efficiency

Definition 24.3.5 (Efficiency)

Robustness

Definition 24.3.6 (Robustness)

Chapter 25

Point Estimation

25.1 Maximum Likelihood Estimator

25.1.1 Classical Likelihood Estimator

Suppose that $\mathbf{x}_n = (x_1, \ldots, x_n)$, within a parametric family

$$p\left(x; \theta_0\right) \in \mathcal{P} = \{p(x; \theta) : \theta \in \Theta\} \tag{25.1}$$

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

$$\hat{\theta}_{\mathrm{MLE}} = \max_{\theta \in \Theta} L\left(\theta; \mathbf{x}_n\right). \tag{25.2}$$

In practice, it is often convenient to work with the natural logarithm of the likelihood function, called the log-likelihood:

$$\ell\left(\theta; \mathbf{x}_n\right) = \ln L\left(\theta; \mathbf{x}_n\right) \tag{25.3}$$

Since the logarithm is a monotonic function, the maximum of $\ell\left(\theta; \mathbf{x}_n\right)$ occurs at the same value of θ as does the maximum of $L\left(\theta; \mathbf{x}_n\right)$

If the data are independent and identically distributed, then we have

$$\hat{\ell}\left(\theta; \mathbf{x}_n\right) := \frac{1}{n} \sum_{i=1}^{n} \ln p\left(x_i; \theta\right) \tag{25.4}$$

this being the sample analog of the expected log-likelihood $\ell(\theta) = \mathbb{E}\left[\ln p\left(x_i \mid \theta\right)\right]$, where this expectation is taken to the true density.

$$\hat{\theta}_{\mathrm{MLE}} := \max_{\theta \in \Theta} \hat{\ell}\left(\theta; \mathbf{x}_n\right) \tag{25.5}$$

Consistency To establish consistency, the following conditions are sufficient:

1. Identification of the model: θ_0 is identified in the sense that if $\theta \neq \theta_0$ and $\theta \in \Theta$, then $p(x; \theta) \neq p\left(x; \theta_0\right)$ with respect to the dominating measure μ.
2. Compactness: the parameter space Θ of the model is compact.
3. Continuity: the function $\ln p(x \mid \theta)$ is continuous in θ for almost all values of x:

$$P\left[\ln p(x \mid \theta) \in C^0(\Theta)\right] = 1 \tag{25.6}$$

4. Dominance: there exists $D(x)$ integrable with respect to the distribution $p(x \mid \theta_0)$ such that $|\ln p(x \mid \theta)| < D(x)$ for all $\theta \in \Theta$.

Lemma 25.1.1

If θ_0 is identified and $E_{\theta_0}[|\ln p(x;\theta)|] < \infty, \forall \theta \in \Theta$, then $\ell(\theta)$ is uniquely maximized at $\theta = \theta_0$.

Proof. By the strict version of Jensen's inequality, for random variable $a = \frac{p(z|\theta)}{p(z|\theta_0)}$ with $\theta \neq \theta_0$,

$$\ell(\theta_0) - \ell(\theta) = E_{\theta_0}\left\{-\ln\left[\frac{p(z \mid \theta)}{p(z \mid \theta_0)}\right]\right\} > -\ln E_{\theta_0}\left[\frac{p(z \mid \theta)}{p(z \mid \theta_0)}\right] = -\ln\left[\int f(z \mid \theta)\,\mathrm{d}z\right] = 0 \quad (25.7)$$

□

Theorem 25.1.1 (Consistency of MLE)

Under the Assumption (1)-(4), we have

$$\hat{\theta}_{\mathrm{MLE}} \xrightarrow{p} \theta_0 \quad (25.8)$$

Proof. Suppose

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

Since $Q_0(\theta)$ is a continuous function, thus

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

is a achieved for a θ in the compact set $\theta \in \Theta \cap \Theta(\epsilon)^C$ (For open set $\Theta(\epsilon)$, $\Theta \cap \Theta(\epsilon)^C$ is a compact set). And θ_0 is the unique maximized,

$$\exists \delta > 0, \quad \ell(\theta_0) - \ell(\theta^*) = \delta$$

1. For $\theta \in \Theta \cap \Theta(\epsilon)^C$. suppose

$$A_n = \left\{\sup_{\theta \in \Theta \cap \Theta(\epsilon)^C} \left|\hat{\ell}(\theta; \mathbf{X}_n) - \ell(\theta)\right| < \frac{\delta}{2}\right\} \quad (25.9)$$

then,

$$A_n \Longrightarrow \hat{\ell}(\theta; \mathbf{X}_n) < \ell(\theta) + \frac{\delta}{2} \leq \ell(\theta^*) + \frac{\delta}{2} = \ell(\theta_0) - \frac{\delta}{2} \quad (25.10)$$

2. For $\theta \in \Theta(\epsilon)$, suppose

$$B_n = \left\{\sup_{\theta \in \Theta(\epsilon)} \left|\hat{\ell}(\theta; \mathbf{x}_n) - \ell(\theta)\right| < \frac{\delta}{2}\right\} \quad (25.11)$$

then

$$B_n \Longrightarrow \forall \theta \in \Theta(\epsilon),\ \hat{\ell}(\theta; \mathbf{x}_n) > \ell(\theta) - \frac{\delta}{2} \quad (25.12)$$

By the uniform law of large numbers, the dominance condition together with continuity establishes the uniform convergence in the probability of the log-likelihood:

$$\sup_{\theta\in\Theta} |\hat{\ell}(\theta; \mathbf{x}_n) - \ell(\theta)| \xrightarrow{p} 0 \tag{25.13}$$

Thus, we can conclude that

$$P(A_n \cap B_n) \to 1 \tag{25.14}$$

Within the definition

$$\hat{\theta}_{\text{MLE}} = \max_{\theta\in\Theta} \hat{\ell}(\theta; \mathbf{x}_n) \tag{25.15}$$

we have,

$$A_n \cap B_n \Longrightarrow \hat{\theta}_{\text{MLE}} \in \Theta(\epsilon)$$

Hence,

$$\forall \varepsilon > 0,\ P\left[\hat{\theta}_{\text{MLE}} \in \Theta(\epsilon)\right] \to 1 \Longrightarrow \hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta_0$$

□

Asymptotic Normality

Efficiency

25.1.2 Modified Likelihood Estimator

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

Marginal Likelihood

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

$$\ell_c = \ln\left(f_{Y|S_{\boldsymbol{\Lambda}}}\right) \tag{25.16}$$

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:

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

$$\ell(\boldsymbol{\theta}, \mathbf{y}) = \boldsymbol{\theta}'\mathbf{s} - b(\boldsymbol{\theta}) \tag{25.17}$$

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

$$\ell(\boldsymbol{\theta}, \mathbf{y}) = \boldsymbol{\varphi}'\mathbf{s}_1 + \boldsymbol{\Lambda}'\mathbf{s}_2 - b(\boldsymbol{\varphi}, \boldsymbol{\Lambda}) \tag{25.18}$$

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

$$\ell(\boldsymbol{\varphi} \mid \mathbf{s}_2) = \boldsymbol{\varphi}'\mathbf{s}_1 - b^*(\boldsymbol{\varphi}, \boldsymbol{\Lambda}) \tag{25.19}$$

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 $\varphi = \ln\left(\frac{\mu_2}{\mu_1}\right) = \ln(\mu_2) - \ln(\mu_1)$ is the parameter of interest and the nuisance parameter is

1. $\lambda_1 = \ln(\mu_1)$.
2.

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

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

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

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,

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

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

□

Profile Likelihood

Quasi Likelihood

25.2 Minimum-Variance Unbiased Estimator

Definition 25.2.1 (UMVU Estimators)

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

$$\mathrm{Var}_\theta \delta(\mathbf{X}) \leq \mathrm{Var}_\theta \delta'(\mathbf{X}), \quad \forall \theta \in \Theta, \tag{25.20}$$

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
$$E_\theta \delta(T(\mathbf{X})) = g(\theta), \quad \forall \theta \in \Theta, \tag{25.21}$$
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
$$E\left[\delta(\mathbf{X}) \mid T(\mathbf{X})\right]. \tag{25.22}$$

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

$$\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{1}{2\sigma^2}\sum (x_i - \xi)^2\right]. \tag{25.23}$$

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,

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

which means the UMVU estimator is

$$\hat{\theta_1} = \frac{n-1}{4n}X_{(1)} + \frac{3n+1}{4n}X_{(n)}, \quad \hat{\theta_2} = -\frac{n+1}{4n}X_{(1)} + \frac{n+1}{4n}X_{(n)}.$$

□

Chapter 26

Interval Estimation

26.1 Confidence Interval

26.2 Pivot

26.3 Likelihood Interval

26.4 Prediction Interval

26.5 Tolerance Interval

Chapter 27

Testing Hypotheses

27.1 Testing Hypotheses

27.2 Parametric Tests

27.3 Specific Tests

27.3.1 Goodness of Fit

Likelihood-Ratio Test

27.3.2 Rank statistics

Chapter 28

Bayesian Inference

28.1 Bayes Estimator

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

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

Definition 28.1.1 (Bayes Estimator)

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

$$r(\Lambda, \delta) = \int R(\theta, \delta) \, \mathrm{d}\Lambda(\theta) \tag{28.1}$$

where Λ is the probability distribution.

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

- 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 calculates 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

$$\pi(\theta \mid x, \lambda) = \frac{f(x \mid \theta) \, \pi(\theta \mid \lambda)}{m(x \mid \lambda)}, \tag{28.2}$$

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

Theorem 28.1.1

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

$$E\{L[\Theta, \delta(x)] \mid X = x\}. \tag{28.3}$$

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

Remark. Improper prior

Corollary 28.1.1

Suppose the assumptions of Theorem 28.1.1 hold.

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

$$\delta_\Lambda(x) = E[g(\Theta) \mid x]. \tag{28.4}$$

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

$$\delta_\Lambda(x) = \frac{E[w(\theta) g(\Theta) \mid x]}{E[w(\theta) \mid x]}. \tag{28.5}$$

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 28.1.2

Necessary condition for Bayes Estimator

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

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

28.1.1 Single-Prior Bayes

The Single-Prior Bayes model in a general form as

$$\begin{aligned} X \mid \theta &\sim f(x \mid \theta), \\ \Theta \mid \gamma &\sim \pi(\theta \mid \lambda), \end{aligned} \tag{28.6}$$

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

$$\int L(\theta, d(x))\, \pi(\theta \mid x)\, \mathrm{d}\theta, \tag{28.7}$$

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

$$E(\Theta \mid x) = \frac{\int \theta f(x \mid \theta)\, \pi(\theta \mid \gamma_0)\, \mathrm{d}\theta}{\int f(x \mid \theta)\, \pi(\theta \mid \gamma_0)\, \mathrm{d}\theta}. \tag{28.8}$$

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 28.1.3

28.1.2 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

$$\begin{aligned} X \mid \theta &\sim f(x \mid \theta), \\ \Theta \mid \gamma &\sim \pi(\theta \mid \lambda), \\ \Gamma &\sim \psi(\gamma), \end{aligned} \tag{28.9}$$

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 then done in practice.

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

$$\int L(\theta, d(x)) \pi(\theta \mid x) \, \mathrm{d}\theta, \tag{28.10}$$

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 written 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 28.10 can be written 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

$$\begin{aligned} X_i \mid \lambda &\overset{\text{i.i.d}}{\sim} \text{Poisson}(\lambda), \quad i = 1, 2 \ldots, n \\ \lambda \mid b &\sim \text{Gamma}(a, b), \text{ a known}, \\ \frac{1}{b} &\sim \text{Gamma}(k, \tau), \end{aligned} \tag{28.11}$$

calculate the Hierarchical Bayes estimator under squared error loss.

Theorem 28.1.4

For the Hierarchical Bayes model (28.9),

$$K[\pi(\lambda \mid x), \psi(\lambda)] < K[\pi(\theta \mid x), \pi(\theta)], \tag{28.12}$$

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

Proof. □

Remark.

28.1.3 Empirical Bayes

28.1.4 Bayes Prediction

Chapter 29

Nonparametric Statistics

29.1 Probability Distribution

29.1.1 Cumulative Distribution Function

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

Definition 29.1.1 (Empirical Cumulative Distribution Function)

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

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\left(X_i \leq x\right) \tag{29.1}$$

29.1.2 Probability Density Function

Histogram

Kernel Density Estimation

29.2 Kernel Methods

29.2.1 Positive Definite Kernels

Definition 29.2.1 (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 $\mathcal{X}$ if and only if it is

1. symmetric, that is,
$$K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x}), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X} \tag{29.2}$$
2. positive definite, that is,
$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0, \tag{29.3}$$
holds for any $x_1, \ldots, x_n \in \mathcal{X}$, given $n \in \mathbb{N}, c_1, \ldots, c_n \in \mathbb{R}$.

Construction of the Reproducing Kernel Hilbert Space

Theorem 29.2.1 (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
$$\langle f, g \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j K(\mathbf{x}_i, \mathbf{y}_j) \tag{29.4}$$
Let $\mathbf{x} \in \mathcal{X}$,
$$\langle f, K(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} a_i K(\mathbf{x}, \mathbf{x}_i) = f(\mathbf{x})$$
And, we also have
$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} a_i g(\mathbf{x}_i) = \sum_{j=1}^{n} b_j f(\mathbf{y}_j)$$

Suppose

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

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

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

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

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

(c) Positive Definiteness: $\langle f, f \rangle_{\mathcal{H}_0} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}_0} = 0$ if and only if $f = 0$.
By positive definiteness of K, we have:

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

As for, $\langle f, f \rangle_{\mathcal{H}_0} = 0$ if and only if $f = 0$, we have,
"$\Rightarrow$" If $f = 0$, that is $f = \sum_{i=1}^{m} a_i K(\cdot, \mathbf{x}_i) = 0$, we have

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

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

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

therefore, if $\|f\|_{\mathcal{H}_0} = 0$, then $f = 0$
Hence, the definition in equation (29.4) is a valid inner product, which is a valid pre-RKHS $\mathcal{H}_0$.

□

Examples of Kernels

Example (Gaussian Kernel)**.**

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2} \right), \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \tag{29.5}$$

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

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

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 X

Computational Statistics

Chapter 30

Random Generator

Random number generator is a key component in computational statistics. It is used to generate random numbers from a given probability distribution. In this chapter, we will introduce some basic concepts and algorithms of random number generation.

30.1 Uniform Random Number Generation

30.2 Non-uniform Random Number Generation

For non-uniform random number generation

30.2.1 Inversion Method

Theorem 30.2.1

Let X be a random variable with cumulative distribution function $F(x)$, then $F(x)$ is a non-decreasing function and $F(x) \in [0,1]$. Let $U \sim \mathcal{U}(0,1)$ be a random variable with uniform distribution on $(0,1)$, then

$$F^{-1}(U) \sim F. \tag{30.1}$$

Proof. Since $F(x)$ is a non-decreasing function, it is invertible. Let $Y = F^{-1}(U)$, then

$$\Pr(Y \leq y) = \Pr(F^{-1}(U) \leq y) = \Pr(U \leq F(y)) = F(y).$$

□

Example (Normal Distribution)**.** Let $X \sim \mathcal{N}(0,1)$ be a random variable with standard normal distribution, then the cumulative distribution function of X is

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t. \tag{30.2}$$

Since there is no closed form of the inverse of $F(x)$, we can use the approximation form:

$$F^{-1}(u) \approx t - \frac{a_0 - a_1 t}{1 + b_1 t + b_2 t^2}, \tag{30.3}$$

where $t = \sqrt{-2 \log u}$.

30.2.2 Rejection Sampling Method

The inversion method is a general method to generate random numbers from a given probability distribution. However, it is not always easy to find the inverse of the cumulative distribution function. In this case, we can use the rejection sampling method.

Suppose we want to generate random numbers from a probability distribution $f(x)$, which is not easy to sample from. We can find a proposal distribution $g(x)$, which is easy to sample from and satisfies

$$\exists M > 0, \quad f(x) \leq Mg(x). \tag{30.4}$$

Then the rejection sampling method is as follows:

Algorithm 1: Rejection Sampling Method

Input: Proposal distribution $g(x)$, constant M

1 Draw a sample $x \sim g(x)$ and $u \sim \mathcal{U}(0,1)$;

2 **if** $u \leq \frac{f(x)}{Mg(x)}$ **then**

3 Accept x;

4 **else**

5 Reject x and go to step 1;

Output: Sample x

Theorem 30.2.2

The rejection sampling method generates a sample x from the probability distribution $f(x)$.

Proof. Let $I = 1$ if x is accepted and $I = 0$ if x is rejected. Then the probability of accepting x given x is

$$\Pr(I = 1 \mid x) = \Pr(u \leq \frac{f(x)}{Mg(x)}) = \frac{f(x)}{Mg(x)}.$$

Thus, the probability of accepting x is

$$\Pr(x \mid I = 1) = \frac{\Pr(x, I = 1)}{\Pr(I = 1)} = \frac{\Pr(x)\Pr(I = 1 \mid x)}{\int \Pr(x)\Pr(I = 1 \mid x)\,\mathrm{d}x} = \frac{f(x)/M}{\int f(x)/M\,\mathrm{d}x} = f(x).$$

□

30.3 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. The state of the chain after several steps is then used as a sample of the desired distribution. The quality of the sample improves as a function of the number of steps.

30.3.1 Metropolis-Hastings Sampling

We want to sample from a distribution $\pi(x)$, where $x \in \mathcal{X}$.

Algorithm 2: Random Walk Metropolis-Hastings Sampling

Input: Initial state $x^{(0)}$, number of iterations N

for $i = 1, \ldots, N$ **do**
- Sample $y \sim \mathcal{N}(x^{(i-1)}, \Sigma)$;
- Compute $\alpha = \min\left\{1, \frac{\pi(y)}{\pi(x^{(i-1)})}\right\}$;
- Sample $u \sim \mathcal{U}(0, 1)$;
- **if** $u < \alpha$ **then**
  - $x^{(i)} = y$;
- **else**
  - $x^{(i)} = x^{(i-1)}$;

Output: Samples $x^{(1)}, \ldots, x^{(N)}$

Random Walk Metropolis-Hastings Sampling

Remark. It is usually difficult to sample from a high-dimensional distribution due to the rejection progressively increasing with the dimensionality.

30.3.2 Gibbs Sampling

Gibbs sampling is a special case of Metropolis-Hastings sampling. It is applicable when the joint distribution is not known explicitly or is difficult to sample from directly, but the conditional distribution of each variable is known and easier to sample from. Gibbs sampling only attempts transitions in the coordinate axis direction, using the conditional distribution of the current point to determine the next step's proposal distribution. All proposal samples are accepted without rejection, resulting in potentially higher efficiency.

Let $\mathbf{x} = (x_1, \ldots, x_d)^\top$ be a random vector with joint distribution $\pi(\mathbf{x})$, and let $\pi(x_i|\mathbf{x}_{-i})$ be the conditional distribution of x_i given $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)^\top$. The Gibbs sampling algorithm is as follows:

Algorithm 3: Gibbs Sampling

Input: Initial state $\mathbf{x}^{(0)}$, number of iterations N, burn-in period B

for $i = 1, \ldots, N$ **do**
- **for** $j = 1, \ldots, d$ **do**
  - Sample $x_j^{(i)} \sim \pi(x_j|x_1^{(i)}, \ldots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \ldots, x_d^{(i-1)})$;
- $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top$;

Output: Samples $\mathbf{x}^{(B+1)}, \ldots, \mathbf{x}^{(N)}$

Chapter 31

Monte Carlo Integration

Suppose we want to estimate the expectation of a function $h(x)$ for a probability distribution $\pi(x)$, i.e., we want to estimate

$$\mu = \mathbb{E}_\pi[h(x)] = \int h(x)\pi(x)\,\mathrm{d}x. \tag{31.1}$$

31.1 Monte Carlo Integration

If we can sample from $\pi(x)$, then we can use the Monte Carlo integration method as follows:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} h(x_i), \tag{31.2}$$

where $x_i \sim \pi(x), \quad i = 1, \ldots, N$.

31.2 Importance Sampling

Importance sampling is a variance reduction technique that can be used when sampling from a distribution $\pi(x)$ is difficult, but sampling from a distribution $g(x)$ is easy. The idea is to sample from $g(x)$ and then reweight the samples so that they are distributed according to $\pi(x)$.

If the probability distribution $\pi(x)$ is difficult to sample from, we can find a proposal distribution $g(x)$. Then the Importance sampling method is as follows:

Algorithm 4: Importance Sampling Method

Input: Proposal distribution $g(x)$, number of samples N

for $i = 1, \ldots, N$ **do**

  Draw a sample $x_i \sim g(x)$;

  Compute $w_i = \frac{\pi(x_i)}{g(x_i)}$;

Calculate $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} w_i h(x_i)$;

Output: Estimate $\hat{\mu}$

Normalized Importance Sampling If we do not know the normalization constant of $\pi(x)$, we can use the normalized importance sampling method as follows:

$$\hat{\mu} = \frac{\sum_{i=1}^{N} w_i h(x_i)}{\sum_{i=1}^{N} w_i}. \tag{31.3}$$

Chapter 32

Bootstrap

Bootstrap is a statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample, most often to derive robust estimates of standard errors and confidence intervals of a population parameter like a mean, median, proportion, odds ratio, correlation coefficient, or regression coefficient.

32.1 Bootstrap Principle

Suppose the i.i.d samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ from an unknown probability distribution F on some probability space $\mathcal{X}$. Let $\hat{\theta}_n$ be an estimator of θ based on the sample $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, that is,

$$\hat{\theta}_n = s\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\right),$$

where $s(\cdot)$ is some algorithm.

The bootstrap principle is to use the sample $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ to estimate the sampling distribution of $\hat{\theta}_n$.

Nonparametric Bootstrap In the b-th bootstrap replicate, we sample with replacement from the original sample $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ to get $\left\{\mathbf{x}_1^{*b}, \mathbf{x}_2^{*b}, \ldots, \mathbf{x}_n^{*b}\right\}$, and then compute the bootstrap estimate of $\hat{\theta}_n$ as

$$\hat{\theta}_n^{*b} = s\left(\mathbf{x}_1^{*b}, \mathbf{x}_2^{*b}, \ldots, \mathbf{x}_n^{*b}\right). \tag{32.1}$$

Parametric Bootstrap

Bayesian Bootstrap

Smooth Bootstrap

Block Bootstrap

32.2 Standard Error Estimation

The bootstrap estimate of the standard error of $\hat{\theta}_n$ is

$$\widehat{\mathrm{se}}_{\mathrm{boot}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_n^{*b} - \hat{\theta}_n^{*}\right)^2}, \tag{32.2}$$

where $\hat{\theta}_n^{*b}$ is the b-th bootstrap replicate of $\hat{\theta}_n$ and $\hat{\theta}_n^{*}$ is the bootstrap estimate of $\hat{\theta}_n$, that is,

$$\hat{\theta}_n^{*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_n^{*b}. \tag{32.3}$$

32.3 Bias Estimation

The bootstrap estimate of the bias of $\hat{\theta}_n$ is

$$\widehat{\mathrm{Bias}}_{\mathrm{boot}} = \hat{\theta}_n^{*} - \hat{\theta}_n. \tag{32.4}$$

Remark. The bias-corrected bootstrap estimate of $\hat{\theta}_n$ is

$$\hat{\theta}_n^{*} = \hat{\theta}_n - \widehat{\mathrm{Bias}}_{\mathrm{boot}} = 2\hat{\theta}_n - \hat{\theta}_n^{*}. \tag{32.5}$$

32.4 Confidence Interval Estimation

Percentile Confidence Interval The $1-\alpha$ percentile confidence interval of $\hat{\theta}_n$ is

$$\left[\hat{\theta}_n^{*}(\alpha/2), \hat{\theta}_n^{*}(1-\alpha/2)\right], \tag{32.6}$$

where $\hat{\theta}_n^{*}(\alpha/2)$ is the $\alpha/2$-th percentile of the bootstrap distribution of $\hat{\theta}_n$ and $\hat{\theta}_n^{*}(1-\alpha/2)$ is the $(1-\alpha/2)$-th percentile of the bootstrap distribution of $\hat{\theta}_n$.

Bootstrap-t Confidence Interval The $1-\alpha$ bootstrap-t confidence interval of $\hat{\theta}_n$ is

$$\left[\hat{\theta}_n - t_{n-1}^{*}(1-\alpha/2)\,\widehat{\mathrm{se}}_{\mathrm{boot}}, \hat{\theta}_n - t_{n-1}^{*}(\alpha/2)\,\widehat{\mathrm{se}}_{\mathrm{boot}}\right], \tag{32.7}$$

where $t_{n-1}^{*}(1-\alpha/2)$ is the $(1-\alpha/2)$-th percentile of the bootstrap distribution of t_{n-1} and $t_{n-1}^{*}(\alpha/2)$ is the $\alpha/2$-th percentile of the bootstrap distribution of t_{n-1}.

Bias-corrected and accelerated (BCa) Confidence Interval The $1-\alpha$ bias-corrected and accelerated (BCa) confidence interval of $\hat{\theta}_n$ is

$$\left[\hat{\theta}_n^{*}(\alpha_{\mathrm{BCa}}/2), \hat{\theta}_n^{*}(1-\alpha_{\mathrm{BCa}}/2)\right], \tag{32.8}$$

where α_{BCa} is the percentile of the bootstrap distribution of $\hat{\theta}_n$ that satisfies

$$\alpha_{\mathrm{BCa}} = \Phi\left(z_0 + \frac{z_0 + z_\alpha - \widehat{\mathrm{Bias}}_{\mathrm{boot}}}{1-\hat{\mathrm{a}}}\right), \tag{32.9}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, z_0 is the $1-\alpha/2$ percentile of the standard normal distribution, z_α is the $\alpha/2$ percentile of the standard normal distribution, and $\widehat{\mathrm{a}}$ is the percentile of the bootstrap distribution of $\hat{\theta}_n$ that satisfies

$$\widehat{\mathrm{a}} = \frac{\sum_{b=1}^{B}\left(\hat{\theta}_n^{*b} - \hat{\theta}_n^{*}\right)^3}{6\left[\sum_{b=1}^{B}\left(\hat{\theta}_n^{*b} - \hat{\theta}_n^{*}\right)^2\right]^{3/2}}. \tag{32.10}$$

32.5 Hypothesis Testing

One-sample Hypothesis Testing

Two-sample Hypothesis Testing

32.6 Jackknife

Let $\mathbf{x}_{-i}$ be the sample with x_i removed, $\mathbf{x}_{-i} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)^\top$, and denote the corresponding value of the statistic of interest as

$$\hat{\theta}_{-i} = s\left(\mathbf{x}_{-i}\right) \tag{32.11}$$

Bias of $\hat{\theta}$ For almost all reasonable and practical estimates, we have

$$\mathrm{Bias}(\hat{\theta}_n) = \mathrm{E}(\hat{\theta}_n) - \theta \to 0, \quad \text{as } n \to \infty$$

Then, it is reasonable to assume a power series of the type

$$\mathrm{E}(\hat{\theta}_n) = \theta + \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} + \ldots$$

with some coefficients $\{a_k\}$. And we have

$$\mathrm{E}(\hat{\theta}_{-i}) = \mathrm{E}(\hat{\theta}_{n-1}) = \theta + \frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \frac{a_3}{(n-1)^3} + \ldots$$

For the sake of a smaller variance, we average all such estimates and let

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i}$$

thus,

$$\mathrm{E}(\hat{\theta}_{(\cdot)}) = \theta + \frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \frac{a_3}{(n-1)^3} + \ldots$$

Thus, we have

$$(n-1)\,\mathrm{E}\left[\hat{\theta}_{(\cdot)} - \hat{\theta}_n\right] = \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} + \ldots = \mathrm{Bias}(\hat{\theta})$$

Hence, we can get the jackknife estimate bias for $\hat{\theta}$ be

$$\widehat{\mathrm{Bias}}_{\mathrm{jack}} = (n-1)\left(\hat{\theta}_{(\cdot)} - \hat{\theta}_n\right) \tag{32.12}$$

Remark. It is easy to combine the averaged Jackknife estimator $\hat{\theta}_{-i}$ with the original $\hat{\theta}$, to kill the main term in the bias of $\hat{\theta}$, thus,

$$\begin{aligned}\mathbb{E}\left[n\hat{\theta}_n - (n-1)\hat{\theta}_{(\cdot)}\right] &= [n\theta - (n-1)\theta] + [a_1 - a_1] + \left[\frac{a_2}{n} - \frac{a_2}{n-1}\right] + \dots \\ &= \theta + \frac{a_2}{n(n-1)} + \dots = \theta + \frac{a_2}{n^2} + O\left(n^{-3}\right)\end{aligned}$$

This removes the bias in the special case that the bias is $O\left(n^{-1}\right)$ and removes it to $O\left(n^{-2}\right)$ in other cases.

Variance of $\hat{\theta}$ The jackknife estimate of variance for $\hat{\theta}$ is

$$\widehat{\mathrm{Var}}_{\mathrm{jack}} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{-i} - \hat{\theta}_{(\cdot)}\right)^2, \quad \text{where } \hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i} \tag{32.13}$$

The jackknife method of estimation can fail if the statistic $\hat{\theta}_{\mathrm{jack}}$ is not smooth. Smoothness implies that relatively small changes to data values will cause only a small change in the statistic.

Example (Sample Mean)**.**

Example (Sample Correlation Coefficient)**.**

Part XI

Regression Analysis

Chapter 33

Linear Regression

Chapter 34

Generalized Linear Model

34.1 Introduction

Suppose the response Y has a distribution in the exponential family

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

with link function g, such that,

$$E\left(Y \mid \mathbf{x}\right) = \mu = g^{-1}(\eta), \quad \eta = \mathbf{x}^\top \boldsymbol{\beta} \tag{34.1}$$

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 34.1: 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: $0, 1, 2, \ldots$	$\log\left(\mu\right)$	$\exp\left(\eta\right)$

Maximum Likelihood Suppose the log-likelihood function be

$$\ell\left(\boldsymbol{\beta} \mid \mathbf{x}, y\right) = \log\left[f\left(y \mid \theta, \phi\right)\right] = \log\left[f\left(y \mid g^{-1}(\eta), \phi\right)\right] \tag{34.2}$$

where g is the canonical link function and $\eta = \mathbf{x}^\top \boldsymbol{\beta}$.

Let

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

be the score function and observed information matrix.

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

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

By the mean value theorem,

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

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

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

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

1. Newton-Raphson Method:

$$\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t + A^{-1}\left(\hat{\boldsymbol{\beta}}_t\right) U\left(\hat{\boldsymbol{\beta}}_t\right) \Leftrightarrow A\left(\hat{\boldsymbol{\beta}}_t\right)\hat{\boldsymbol{\beta}}_{t+1} = A\left(\hat{\boldsymbol{\beta}}_t\right)\hat{\boldsymbol{\beta}}_t + U\left(\hat{\boldsymbol{\beta}}_t\right) \tag{34.3}$$

   where

$$U\left(\boldsymbol{\beta}\right) = \frac{\partial \ell\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} \tag{34.4}$$

   is the score function and

$$A\left(\boldsymbol{\beta}\right) = -\frac{\partial^2 \ell\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}' \partial \boldsymbol{\beta}} \tag{34.5}$$

   is the observed information matrix.

2. Fisher Scoring Method:

$$\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t + I^{-1}\left(\hat{\boldsymbol{\beta}}_t\right) U\left(\hat{\boldsymbol{\beta}}_t\right) \Leftrightarrow I\left(\hat{\boldsymbol{\beta}}_t\right)\hat{\boldsymbol{\beta}}_{t+1} = I\left(\hat{\boldsymbol{\beta}}_t\right)\hat{\boldsymbol{\beta}}_t + U\left(\hat{\boldsymbol{\beta}}_t\right) \tag{34.6}$$

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

$$I\left(\boldsymbol{\beta}\right) = E\left[A\left(\boldsymbol{\beta}\right)\right] = -E\left[\frac{\partial^2 \ell\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}' \partial \boldsymbol{\beta}}\right] \tag{34.7}$$

   is the Fisher information matrix.

Bayesian Methods

34.2 Binary Data

Suppose

$$Y \sim b\left(m, \pi\right), \quad i = 1, 2, \ldots, n \tag{34.8}$$

with link function

$$\eta = g\left(\pi\right) = \log\left(\frac{\pi}{1-\pi}\right) = \mathbf{x}^\top \boldsymbol{\beta} \tag{34.9}$$

Remark.

The likelihood function is

$$f(\boldsymbol{\pi} \mid \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i} \tag{34.10}$$

and the log-likelihood function is

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \log\left[f(\boldsymbol{\pi} \mid \mathbf{x}, \mathbf{y})\right] = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \left\{ \log\left[\binom{m_i}{y_i}\right] + y_i \log(\pi_i) + (m_i - y_i)\log(1-\pi_i) \right\} \\ &= \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i) \right] + \sum_{i=1}^{n} \log\left[\binom{m_i}{y_i}\right] \end{aligned} \tag{34.11}$$

where

$$\pi_i = \frac{\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{1 + \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)} \tag{34.12}$$

Thus,

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

34.3 Polytomous Data

Definition 34.3.1 (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

$$\pi_j(\mathbf{x}_i) = P(Y = y_j \mid \mathbf{x} = \mathbf{x}_i) \tag{34.13}$$

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

$$r_j(\mathbf{x}_i) = P\left(Y \leq \sum_{r \leq j} y_r \mid \mathbf{x} = \mathbf{x}_i\right) \tag{34.14}$$

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

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 density 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

Nominal Scale

$$\pi_j\left(\mathbf{x}_i\right) = \frac{\exp\left[\eta_j\left(\mathbf{x}_i\right)\right]}{\sum_{j=1}^k \exp\left[\eta_j\left(\mathbf{x}_i\right)\right]} \tag{34.15}$$

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

Ordinal Scale

1. Logistic Scale:

$$\log\left[\frac{r_j\left(\mathbf{x}_i\right)}{1 - r_j\left(\mathbf{x}_i\right)}\right] = \theta_j - \mathbf{x}_i^\top \boldsymbol{\beta} \tag{34.16}$$

2. Complementary Log-Log Scale:

$$\log\left\{-\log\left[1 - r_j\left(\mathbf{x}_i\right)\right]\right\} = \theta_j - \mathbf{x}_i^\top \boldsymbol{\beta} \tag{34.17}$$

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.

$$\log\left[\frac{r_j\left(\mathbf{x}_i\right)}{1 - r_j\left(\mathbf{x}_i\right)}\right] = \varsigma_0 + \varsigma_1\left(\frac{s_j + s_{j+1}}{2}\right) - \mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_i^\top \boldsymbol{\xi}\left(c_j - \bar{c}\right) \tag{34.18}$$

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.

$$\pi_j\left(\mathbf{x}_i\right) = \frac{\exp\left[\eta_j\left(\mathbf{x}_i\right)\right]}{\sum_{j=1}^k \exp\left[\eta_j\left(\mathbf{x}_i\right)\right]} \tag{34.19}$$

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

3.

$$\sum_{j=1}^k \pi_j\left(\mathbf{x}_i\right) s_j = \mathbf{x}_i \boldsymbol{\beta} \tag{34.20}$$

34.4 Count Data

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

$$\text{Var}\left(Y\right) = \sigma^2 E\left(Y\right) \tag{34.21}$$

with link function

$$\log\left(\mu\right) = \eta = \mathbf{x}^\top \boldsymbol{\beta} \tag{34.22}$$

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

$$\ell(\boldsymbol{\beta}) \propto \sum_{i=1}^{n} (y_i \log(\mu_i) - \mu_i) \tag{34.23}$$

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

Chapter 35

Quantile Regression

Chapter 36

Survival Analysis

36.1 General Formulation

Definition 36.1.1 (Survival Function)

The survival function[a] is defined to be

$$S(t) = P(T > t) = \int_t^\infty f(u)\,\mathrm{d}u = 1 - F(t). \tag{36.1}$$

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 36.1.2 (Lifetime Distribution Function)

The lifetime distribution function is defined to be

$$F(t) = P(T \leq t) \tag{36.2}$$

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

$$f(t) = F'(t) = \frac{\mathrm{d}}{\mathrm{d}t}F(t) \tag{36.3}$$

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

Definition 36.1.3 (Hazard Function)

The Hazard function[a] is defined to be

$$\lambda(t) = \lim_{\varepsilon \to 0^+} \left[\frac{P(t \leq T < t + \varepsilon \mid T \geq t)}{\varepsilon} \right] = \frac{f(t)}{S(t)} \tag{36.4}$$

[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.
$$\lambda(t) = -\frac{\mathrm{d}\log[S(t)]}{\mathrm{d}t} \tag{36.5}$$

2.
$$S(t) = \exp\left[-\int_0^t \lambda(x)\,\mathrm{d}x\right] \tag{36.6}$$

3.
$$f(t) = \lambda(t)\exp\left[-\int_0^t \lambda(x)\,\mathrm{d}x\right] \tag{36.7}$$

Proof. □

Example (Constant Hazards)**.** Suppose

$$\lambda(t) = \lambda \tag{36.8}$$

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)**.**

$$\lambda(t) = \alpha t + \frac{\beta}{1+\gamma t} \tag{36.9}$$

36.2 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

$$\begin{aligned} f\left(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}\right) &= \prod_{i=1}^{n} \left[f\left(t_i\right)\right]^{\delta_i} \left[S\left(t_i\right)\right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left(\frac{f\left(t_i\right)}{S\left(t_i\right)}\right)^{\delta_i} S\left(t_i\right) \\ &= \prod_{i=1}^{n} \left[\lambda\left(t_i\right)\right]^{\delta_i} S\left(t_i\right) \end{aligned} \tag{36.10}$$

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

$$\begin{aligned}\hat{S}(t) &= \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right) \\ &= \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right) \cdots \left(1 - \frac{d_k}{n_k}\right) \\ &\approx \left[1 - \hat{\lambda}\left(t_1\right)\right]\left[1 - \hat{\lambda}\left(t_2\right)\right] \ldots \left[1 - \hat{\lambda}\left(t_k\right)\right]\end{aligned} \tag{36.11}$$

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

36.3 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

$$\lambda(t) = \lambda_0(t) \mathrm{e}^{\mathbf{x}^\top \boldsymbol{\beta}} \tag{36.12}$$

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

The exact likelihood function is

$$\ell\left[\beta, \lambda_0(t)\right] = \prod_{i=1}^{n} \left[\lambda_i\left(t_i\right)\right]^{\delta_i} S\left(t_i\right) \tag{36.13}$$

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

$$\ell_p(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{\mathrm{e}^{\mathbf{x}'_{(j)}\boldsymbol{\beta}}}{\sum_{l \in R\left(t_{(j)}\right)} \mathrm{e}^{\mathbf{x}'_l \boldsymbol{\beta}}} = \prod_{i=1}^{n} \left[\frac{\mathrm{e}^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{\sum_{l \in R(t_i)} \mathrm{e}^{\mathbf{x}'_l \boldsymbol{\beta}}}\right]^{\delta_i} \tag{36.14}$$

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

Chapter 37

Nonparametric Regression

37.1 Uniform Stability of Regularized Kernel Model

37.1.1 Introduction

Suppose $\mathcal{X}$ be the input space, $\mathcal{Y}$ be the output space, $\mathcal{D}$ be some (almost) completely unknown probability distribution on $\mathcal{X} \times \mathcal{Y}$. Given the n i.i.d observed data, which sampled from an unknown distribution $\mathcal{D}$, that,

$$S := \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \subseteq \mathcal{D} \tag{37.1}$$

and the goal of us is to estimate the functional relationship between $\mathcal{X}$ and $\mathcal{Y}$.

To formalize the problem, we now aim at finding a predictor function f^* among the function space $\mathcal{F} := \{f : \mathcal{X} \to \mathcal{Y}\}$ based on the observed data S, which minimizes the true risk

$$R[f] := \mathbb{E}_{\mathcal{D}}\left[L\left(y, f(\mathbf{x})\right)\right] \tag{37.2}$$

where $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is an arbitrary convex loss function, typically assumed that the smaller $L\left(y, f(\mathbf{x})\right)$ is, the better the approximation of y is. Thus, we are trying to find a predictor f^* with risk close to the optimal risk

$$R^* := \inf\left\{R[f] \mid f : \mathcal{X} \to \mathcal{Y}\right\} \tag{37.3}$$

Finding the predictor function f^* which minimizing the empirical risk

$$R_{\text{emp}}(f) = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f(\mathbf{x}_i)\right) \tag{37.4}$$

is a natural thing for us to be trying to do. However, as it is known to all, just minimizing the empirical risk is suicidal, which almost certainly leads to overfitting. Minimizing R_{emp} only makes sense if we simultaneously somehow restrict ourselves to the $\mathcal{F}$, which are of just the right level of complexity. One way to do this is by explicitly restricting the function space $\mathcal{F}$ to a "simple" space, as in structural risk minimization, which is to introduce a penalty functional $\Omega[f]$ that somehow measures the complexity of each function $f \in \mathcal{F}$, and to minimize the regularized risk

$$R_{\text{reg}}[f] = R_{\text{emp}}[f] + \Omega[f] \tag{37.5}$$

In this report, we restrict the predictor function $f \in \mathcal{F}$ among the reproducing kernel Hilbert space $\mathcal{H}$, and the regularized risk has the form

$$R_{\text{reg}}[f] = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f(\mathbf{x}_i)\right) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \tag{37.6}$$

thus, we can estimate f^* by soving the following optimization problem

$$\hat{f} = \arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f(\mathbf{x}_i)\right) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \tag{37.7}$$

where $\lambda > 0$ is a regularization parameter to reduce the danger of overfitting. Since $L\left(y, f(\mathbf{x})\right)$ is convex in f, the minimizer $\hat{f}$ is uniquely determined and a simple gradient descent algorithm can be used to find $\hat{f}$. So the main focus of this report is to answer a remaining question we want to know, whether the risk $R[\hat{f}]$ is close to the optimal risk R^*, which will influence the stability of our algorithm.

37.1.2 Some Notations and Concepts

Before getting into the formal discussion, we will introduce some notations and concepts.

- $S^{\setminus i} := S \setminus \{(\mathbf{x}_i, y_i)\}$ be the sample where the i-th observation is removed.
- $S^i := S^{\setminus i} \cup \{(\mathbf{x}, y)\}$ be the sample where the i-th observation is replaced by $(\mathbf{x}, y)$.

and let $\hat{f}_{\setminus i}$ be the estimated result based on sample $S^{\setminus i}$, $\hat{f}_i$ based on sample S^i and $\hat{f}$ based on sample S.

To quantify the stability of our algorithm, we will introduce one important concept — **Uniform Stability**.

Definition 37.1.1 (Uniform Stability)

The algorithm is uniformly β-stable with respect to the loss function $L\left(y, f(\mathbf{x})\right)$, if for all samples $S := \{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathcal{D}$ and $i \in [n]$,

$$\sup_{(x,y)\in\mathcal{D}} \left| L\left(y, \hat{f}(\mathbf{x})\right) - L\left(y, \hat{f}_{\setminus i}(\mathbf{x})\right)\right| \leq \beta \tag{37.8}$$

i.e. the algorithm is "stable" to remove a single sample at all points.

37.1.3 Uniform Stability of Regularized Kernel Model

Firstly, we will provide an auxiliary lemma.

Lemma 37.1.1 (Convex Functions and Derivatives)

For any differentiable convex function $f : \mathbb{R} \to \mathbb{R}$ and any $a, b \in \mathbb{R}$, we have

$$\left[f'(a) - f'(b)\right](a - b) \geq 0 \tag{37.9}$$

Proof. Due to the convexity of f we know that $f(a) + (b-a)f'(a) \leq f(b)$ and, likewise, $f(b) + (a-b)f'(b) \leq f(a)$. Summing up both inequalities and subtracting the terms in $f(a)$ and $f(b)$ proves (37.9). □

Then, we will show that the algorithm we studied in this paper satisfied the definition 37.1.1, and the corresponding value of β can be calculated.

Theorem 37.1.1 (Algorithmic Stability of Risk Minimizers [2, 6])

The algorithm that minimizes the regularized empirical risk in (37.6) has stability

$$\beta = \frac{2C^2\kappa^2}{n\lambda} \tag{37.10}$$

where κ is bound on $\|k(x,\cdot)\| = \sqrt{k(x,x)}$, $\|\cdot\|$ is the RKHS norm induced by k, and C is a bound on the Lipschitz constant of the loss function $L\left(y, f(\mathbf{x})\right)$, which can be viewed as a function of f.

Remark. We can see that the stability of the algorithm depends on the regularization constant via $\frac{1}{\lambda n}$, hence we may be able to afford to choose weaker regularization if the sample size n increases.

Proof. To distinguish between different training sets, we use $R_{\text{reg}}[f,S]$ and $R_{\text{reg}}[f,S^{\setminus i}]$ (and likewise $R_{\text{emp}}[f,S]$) during the remainder of the proof.

Since $\hat{f}$ minimizes $R_{\text{reg}}[f,S]$, that is, the **functional derivative** [7] of $R_{\text{reg}}[f,S]$ at $\hat{f}$ vanishes, and so does $R_{\text{reg}}[f,S^{\setminus i}]$ at $\hat{f}_{\setminus i}$,

$$\begin{aligned}\partial_f R_{\text{reg}}\left[\hat{f},S\right] =& \partial_f R_{\text{emp}}\left[\hat{f},S\right] + \lambda\hat{f} = 0\\ \partial_f R_{\text{reg}}\left[\hat{f}_{\setminus i},S^{\setminus i}\right] =& \partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i},S^{\setminus i}\right] + \lambda\hat{f}_{\setminus i} = 0\end{aligned} \tag{37.11}$$

Next, we construct an auxiliary risk function $\tilde{R}[f]$ by

$$\tilde{R}[f] := \left\langle \partial_f R_{\text{emp}}\left[\hat{f},S\right] - \partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i},S^{\setminus i}\right], f - \hat{f}_{\setminus i}\right\rangle + \frac{\lambda}{2}\left\|f - \hat{f}_{\setminus i}\right\|_{\mathcal{H}}^2 \tag{37.12}$$

$\tilde{R}[f]$ is a convex function in f (the first term is linear, the second quadratic).

Additionally, by construction, we have

$$\tilde{R}[\hat{f}_{\setminus i}] = 0 \tag{37.13}$$

Furthermore, taking the functional derivative of $\tilde{R}[f]$, that,

$$\partial_f \tilde{R}[f] = \partial_f R_{\text{emp}}\left[\hat{f},S\right] - \partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i},S^{\setminus i}\right] + \lambda\left(f - \hat{f}_{\setminus i}\right) = \partial_f R_{\text{emp}}\left[\hat{f},S\right] + \lambda f \tag{37.14}$$

the functional derivative of $\tilde{R}[f]$ (37.14) vanishes at $f = \hat{f}$ due to (37.11), thus the minimum of $\tilde{R}[f]$ is obtained for $f = \hat{f}$. Therefore, combined with $\tilde{R}[\hat{f}_{\setminus i}] = 0$, we can conclude that $\tilde{R}[\hat{f}] \leq 0$.

To obtain bounds on $\left\|\hat{f} - \hat{f}_{\setminus i}\right\|$, we have to get rid of some of the first terms in $\tilde{R}[f]$, since

$$\begin{aligned}& n\left\langle \partial_f R_{\text{emp}}\left[\hat{f},S\right] - \partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i},S^{\setminus i}\right], \hat{f} - \hat{f}_{\setminus i}\right\rangle\\ =& \sum_{j\neq i}\left[L'\left(y_j,\hat{f}(\mathbf{x}_j)\right) - L'\left(y_j,\hat{f}_{\setminus i}(\mathbf{x}_j)\right)\right]\left[\hat{f}(\mathbf{x}_j) - \hat{f}_{\setminus i}(\mathbf{x}_j)\right]\\ &+ L'\left(y_i,\hat{f}(\mathbf{x}_i)\right)\left[\hat{f}(\mathbf{x}_i) - \hat{f}_{\setminus i}(\mathbf{x}_i)\right]\\ \geq& L'\left(y_i,\hat{f}(\mathbf{x}_i)\right)\left[\hat{f}(\mathbf{x}_i) - \hat{f}_{\setminus i}(\mathbf{x}_i)\right]\end{aligned} \tag{37.15}$$

The first equation is since the functional derivative $\partial_f(f) = k(\mathbf{x}, .)$ and then collecting the common terms between $R_{\text{emp}}\left[\hat{f}, S\right]$ and $R_{\text{emp}}\left[\hat{f}_{\setminus i}, S^{\setminus i}\right]$. And, as for the last inequation, we use lemma 37.1.1 applied to the loss function $L\left(y, f(\mathbf{x})\right)$ which is a convex function of $f(\mathbf{x})$.

Combine the above result with the fact $\tilde{R}[\hat{f}] \leq 0$, we have

$$\left\langle \partial_f R_{\text{emp}}\left[\hat{f}, S\right] - \partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i}, S^{\setminus i}\right], \hat{f} - \hat{f}_{\setminus i} \right\rangle + \frac{\lambda}{2}\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}}^2 \leq 0 \tag{37.16}$$

thus,

$$L'\left(y_i, \hat{f}(\mathbf{x}_i)\right)\left[\hat{f}(\mathbf{x}_i) - \hat{f}_{\setminus i}(\mathbf{x}_i)\right] + \frac{n\lambda}{2}\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}}^2 \leq 0 \tag{37.17}$$

and by the convexity of loss function $L\left(y, f(\mathbf{x})\right)$,

$$\begin{aligned} \frac{n\lambda}{2}\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}}^2 \leq & L'\left(y_i, \hat{f}(\mathbf{x}_i)\right)\left[\hat{f}_{\setminus i}(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)\right] \\ \leq & L\left(y_i, \hat{f}(\mathbf{x}_i)\right) - L\left(y_i, \hat{f}_{\setminus i}(\mathbf{x}_i)\right) \\ \leq & \left|L\left(y_i, \hat{f}(\mathbf{x}_i)\right) - L\left(y_i, \hat{f}_{\setminus i}(\mathbf{x}_i)\right)\right| \end{aligned} \tag{37.18}$$

By the Cauchy-Schwarz inequality, we can see that, for any $f, f' \in \mathcal{H}$ and any $\mathbf{x} \in \mathcal{X}$,

$$|f(\mathbf{x}) - f'(\mathbf{x})| = |\langle f - f', k(\mathbf{x}, \cdot)\rangle| \leq \|f - f'\|_{\mathcal{H}} \|k(\mathbf{x}, \cdot)\|_{\mathcal{H}} \leq \kappa \|f - f'\|_{\mathcal{H}} \tag{37.19}$$

and since $L\left(y, f(\mathbf{x})\right)$ is Lipschitz continuous at $\mathbf{x}_i$, we have

$$\left|L\left(y, \hat{f}(\mathbf{x}_i)\right) - L\left(y, \hat{f}_{\setminus i}(\mathbf{x}_i)\right)\right| \leq C\left|\hat{f}(\mathbf{x}_i) - \hat{f}_{\setminus i}(\mathbf{x}_i)\right| \leq C\kappa\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}} \tag{37.20}$$

Combine equation (37.18) and (37.20), we get

$$\frac{n\lambda}{2}\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}}^2 \leq C\kappa\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}} \tag{37.21}$$

thus,

$$\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}} \leq \frac{2C\kappa}{n\lambda} \tag{37.22}$$

Therefore, by the equation (37.20) for every $\mathbf{x}$, we have

$$\left|L\left(y, \hat{f}(\mathbf{x})\right) - L\left(y, \hat{f}_{\setminus i}(\mathbf{x})\right)\right| \leq C\kappa\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}} \leq \frac{2C^2\kappa^2}{n\lambda} \tag{37.23}$$

□

Within the uniform stability of our algorithm, we will also prove that the β-stable algorithm exhibits uniform convergence of the empirical risk $R_{\text{emp}}[f]$ to the true risk $R[f]$.

Theorem 37.1.2 (McDiarmid's Bound [4])

Suppose $\xi_1, \ldots, \xi_n$ be i.i.d real value random variables and assume that there exists a function $g : \mathbb{R}^n \to \mathbb{R}$ with the property that for all $i \in [n]$ and $c_i > 0$,

$$\sup_{\xi_1,\ldots,\xi_n,\xi_i' \in \xi} |g(\xi_1, \ldots, \xi_n) - g(\xi_1, \ldots, \xi_{i-1}, \xi_i', \xi_{i+1}, \ldots, \xi_n)| \leq c_i \tag{37.24}$$

where ξ_i' is drawn from the same distribution as ξ_i. Then

$$\mathbb{P}\{|g(\xi_1, \ldots, \xi_n) - \mathbb{E}[g(\xi_1, \ldots, \xi_n)]| > \varepsilon\} \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right) \tag{37.25}$$

Theorem 37.1.3 (Bousquet and Elisseeff [1, 5])

Assume that we have a β-stable algorithm with the additional requirement that the loss function $L(y, f(\mathbf{x})) \leq M$ for all $(\mathbf{x}, y) \in \mathcal{D}$ and for all samples $S := \{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathcal{D}$. Then, for any $n \geq 1$

$$\mathbb{P}\left\{\left|R_{\text{emp}}[\hat{f}, S] - R[\hat{f}]\right| > \varepsilon + 2\beta\right\} \leq 2\exp\left(-\frac{n\varepsilon^2}{2(n\beta + M)^2}\right) \tag{37.26}$$

Proof. Within the i.i.d assumption, we have

$$\mathbb{E}_{S\sim\mathcal{D}}\left[R_{\text{emp}}[\hat{f}]\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S\sim\mathcal{D}}\left[L\left(y_i, \hat{f}(\mathbf{x}_i)\right)\right] = \mathbb{E}_{S\sim\mathcal{D}}\left[L\left(y_i, \hat{f}(\mathbf{x}_i)\right)\right] \tag{37.27}$$

If we replace $(\mathbf{x}_i, y_i)$ by $(\mathbf{x}, y)$, we can get

$$\mathbb{E}_{S\sim\mathcal{D}}\left[R_{\text{emp}}[\hat{f}]\right] = \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}_i(\mathbf{x})\right)\right] \tag{37.28}$$

and with the observation that

$$\mathbb{E}_{\mathcal{D}}\left[R[\hat{f}]\right] = \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}(\mathbf{x})\right)\right] \tag{37.29}$$

In order to bound on the expected difference between $R_{\text{emp}}[\hat{f}, S]$ and $R[\hat{f}]$, which leads to

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[R_{\text{emp}}[\hat{f}, S] - R[\hat{f}]\right] =& \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}_i(\mathbf{x})\right)\right] - \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}(\mathbf{x})\right)\right] \\ =& \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}_i(\mathbf{x})\right) - L\left(y, \hat{f}(\mathbf{x})\right)\right] \leq 2\beta\end{aligned} \tag{37.30}$$

By the triangle inequality, we have

$$\left|R[\hat{f}] - R[\hat{f}_i]\right| \leq \left|R[\hat{f}] - R[\hat{f}_{\setminus i}]\right| + \left|R[\hat{f}_{\setminus i}] - R[\hat{f}_i]\right| \leq 2\beta \tag{37.31}$$

Also, we have

$$\begin{aligned}\left|R_{\text{emp}}[\hat{f}, S] - R_{\text{emp}}\left(\hat{f}_i, S^i\right)\right| \leq& \frac{1}{n}\sum_{j\neq i}\left|L\left(y_j, \hat{f}(\mathbf{x}_j)\right) - L\left(y_j, \hat{f}_i(\mathbf{x}_j)\right)\right| \\ &+ \frac{1}{n}\left|L\left(y_i, \hat{f}(\mathbf{x}_i)\right) - L\left(y_i, \hat{f}_i(\mathbf{x}_i)\right)\right| \\ \leq& \frac{n-1}{n}2\beta + \frac{2M}{n} \leq 2\beta + \frac{2M}{n}\end{aligned} \tag{37.32}$$

and,

$$\begin{aligned}\left|\left[R_{\text{emp}}[\hat{f},S]-R[\hat{f}]\right]-\left[R_{\text{emp}}\left(\hat{f}_i,S^i\right)-R[\hat{f}_i]\right]\right| \leq & \left|R_{\text{emp}}[\hat{f},S]-R_{\text{emp}}\left(\hat{f}_i,S^i\right)\right| \\ & +\left|R[\hat{f}]-R[\hat{f}_i]\right| \\ \leq & 4\beta+\frac{2M}{n}\end{aligned} \tag{37.33}$$

Thus, by the Theorem 37.1.2, we have $c_i = 4\beta + \frac{2M}{n}$, that,

$$\begin{aligned}\mathbb{P}\left\{|R_{\text{emp}}[\hat{f},S]-R[\hat{f}]-2\beta|>\varepsilon\right\} \leq & \mathbb{P}\left\{|R_{\text{emp}}[\hat{f},S]-R[\hat{f}]|>\varepsilon+2\beta\right\} \\ \leq & 2\exp\left(-\frac{n\varepsilon^2}{2(2n\beta+M)^2}\right)\end{aligned} \tag{37.34}$$

□

Within the above two theorems, we can directly get the following practical consequence.

Corollary 37.1.1 (Uniform Convergence Bounds for RKHS)

The algorithm minimizing the regularized risk $R_{\text{reg}}[f]$, as in (37.6), and with the assumptions of Theorem 37.1.1 and 37.1.3, we obtain

$$\mathbb{P}\left\{\left|R_{\text{emp}}[\hat{f}]-R[\hat{f}]\right|>\varepsilon+2\beta\right\} \leq 2\exp\left(-\frac{n}{2}\left(\frac{\varepsilon}{M}\right)^2\left(1+\frac{4}{\lambda M}(C\kappa)^2\right)^{-2}\right) \tag{37.35}$$

where

$$\beta = \frac{2C^2\kappa^2}{n\lambda}$$

Remark. For practical considerations, (37.35) may be very useful, even if the rates are not optimal, since the bound is predictive even for small sample sizes and moderate regularization strength. Still, we expect that the constants

Comments

The idea of the discussion content was inspired by [3, Section 3] review, and the report is organized following [6, Chapter 12] structure, and so are the main proof ideas.

Chapter 38

High Dimensional Regression Analysis

38.1 Lasso for Linear Regression

If $p > n$, then the least squares estimator is not unique. In this case, we can use the lasso to select a unique estimator. The lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{38.1}$$

where $\lambda > 0$ is a tuning parameter. In addition, the optimization problem (38.1) can be rewritten as

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_1 \leq R}{\arg\min} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2,$$

with a one-to-one correspondence between λ and R.

38.1.1 Numerical Algorithms

The lasso estimator can be computed by the following numerical algorithms.

Cyclic Coordinate Descent

Cyclic coordinate descent is an iterative algorithm. At each iteration, we update one coordinate of $\boldsymbol{\beta}$ while fixing all other coordinates. The cyclic coordinate descent algorithm for the lasso is given in Algorithm 5.

Algorithm 5: Cyclic Coordinate Descent for the Lasso Estimator

Input: Data $\{(\mathbf{X}_i, y_i)\}_{i=1}^n$, tuning parameter $\lambda > 0$, initial value $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^p$, and tolerance $\epsilon > 0$.

$\mathbf{r} \leftarrow \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(0)}$;

while $\left\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\right\|_\infty > \epsilon$ **do**

  for $j = 1, \ldots, p$ **do**

    $\beta_j^{(t+1)} \leftarrow \frac{S_\lambda\left(\mathbf{X}_j^\top \mathbf{r}/n\right)}{\mathbf{X}_j^\top \mathbf{X}_j/n}$;

    $\mathbf{r} \leftarrow \mathbf{r} + \left(\beta_j^{(t+1)} - \beta_j^{(t)}\right)\mathbf{X}_j$;

Output: Estimate $\hat{\boldsymbol{\beta}}_{\text{lasso}}$.

38.1.2 Selection of the Tuning Parameter

K-Fold Cross-Validation

38.2 Theory for the Lasso

Let $\boldsymbol{\beta}^0$ be the true parameter vector, and let $\hat{\boldsymbol{\beta}}$ be the lasso estimator. We assume that the design matrix $\mathbf{X}$ is orthogonal, i.e., $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$, and that the noise vector $\boldsymbol{\varepsilon}$ is a random vector with mean zero and covariance matrix $\sigma^2\mathbf{I}_n$. We further denote $S_0 := \left\{j : \beta_j^0 \neq 0\right\}$ as the active set, and $s_0 := |S_0|$ as the cardinality of the active set. We also denote $\phi_0 := \min_{j\in S_0}\left|\beta_j^0\right|$ as the minimum absolute value of the nonzero coefficients.

In this section, we will give two properties of the lasso estimator:

1. Prediction Error:

$$\left\|\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)\right\|_2^2/n \leq 2\sigma^2 s_0\left(\frac{\lambda}{\phi_0}\right)^2,$$

Lemma 38.2.1 (Basic Inequality)

$$\left\|\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)\right\|_2^2/n + \lambda\left\|\hat{\boldsymbol{\beta}}\right\|_1 \leq 2\boldsymbol{\varepsilon}^\top\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)/n + \lambda\left\|\boldsymbol{\beta}^0\right\|_1.$$

Proof. □

Corollary 38.2.1

Assume that $\hat{\sigma}_j^2 = 1$ for all j and that the compatibility condition holds for S_0, with $\hat{\boldsymbol{\Sigma}}$ normalized in this way. For some $t > 0$, let the regularization parameter by

$$\lambda := 4\hat{\sigma}\sqrt{\frac{t^2 + 2\log p}{n}},$$

where $\hat{\sigma}^2$ is an estimator of the noise variance σ^2. Then with probability at least $1 - \alpha$, where

$$\alpha := 2\exp\left[-t^2/2\right] + \Pr(\hat{\sigma} \leq \sigma),$$

we have

$$\left\|\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)\right\|_2^2/n + \lambda\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_1 \leq 4\lambda^2 s_0/\phi_0^2.$$

Proof. By Lemma 6.2, for

$$\mathcal{F} = \left\{ \max_{1 \le j \le p} 2|\boldsymbol{\varepsilon}^\top \mathbf{X}^{(j)}|/n \le \lambda_0 \right\},$$

we have for all $t > 0$,

$$\Pr(\mathcal{F}) = 1 - 2\exp\left[-t^2/2\right], \quad \text{where} \quad \lambda_0 = 2\sigma\sqrt{\frac{t^2 + 2\log p}{n}}.$$

As in the proof of corollary 6.2, if $\hat{\sigma} > \sigma$, then we have $\lambda \ge 2\lambda_0$, and

$$\Pr(\mathcal{F} \cap \{\hat{\sigma} > \sigma\}) = 1 - \Pr(\mathcal{F}^c \cup \{\hat{\sigma} \le \sigma\}) \ge 1 - \Pr(\mathcal{F}^c) - \Pr(\hat{\sigma} \le \sigma) = 1 - \alpha,$$

where $\alpha = 2\exp\left[-t^2/2\right] + \Pr(\hat{\sigma} \le \sigma)$. And since the compatibility condition holds, according to Theorem 6.1, we have

$$\left\| \mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right) \right\|_2^2 / n + \lambda \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_1 \le 4\lambda^2 s_0 / \phi_0^2,$$

with probability at least $1 - \alpha$. □

38.3 Other Lasso-Type Estimators

38.3.1 Adaptive Lasso

The adaptive lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\text{adaptive}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{\hat{\sigma}_j} \right\},$$

where $\lambda > 0$ is a tuning parameter, and $\hat{\sigma}_j$ is an estimator of σ_j.

38.3.2 Elastic Net

The elastic net estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\text{elastic}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\},$$

where $\lambda_1, \lambda_2 > 0$ are tuning parameters.

38.3.3 Group Lasso

The group lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\text{group}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g} \|\boldsymbol{\beta}_g\|_2 \right\},$$

where $\lambda > 0$ is a tuning parameter, and $\boldsymbol{\beta}_g$ is the subvector of $\boldsymbol{\beta}$ corresponding to the gth group.

38.3.4 Fused Lasso

The fused lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\text{fused}} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j| \right\},$$

38.4 Nonconvex Penalties

The main drawback of the lasso estimator is that it is a biased estimator. To reduce the bias, we can use nonconvex penalties, such as the smoothly clipped absolute deviation (SCAD) penalty and the minimax concave penalty (MCP).

38.4.1 SCAD

The SCAD penalty is defined as

$$\psi_{\text{SCAD}}(\beta) := \begin{cases} \lambda\beta, & \text{if} \quad |\beta| \le \lambda, \\ \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda}, & \text{if} \quad \lambda < |\beta| \le a\lambda, \\ \frac{1}{(a-1)\lambda^2}\left[(a+1)\lambda^2 - 2a\lambda|\beta| + \beta^2\right], & \text{if} \quad a\lambda < |\beta|, \end{cases}$$

where $a > 2$ is a constant.

38.4.2 MCP

The MCP penalty is defined as

$$\psi_{\text{MCP}}(\beta) := \begin{cases} \lambda\beta, & \text{if} \quad |\beta| \le \lambda, \\ \frac{|\beta|^2}{2(a-1)}, & \text{if} \quad \lambda < |\beta| \le a\lambda, \\ \frac{a\lambda|\beta| - \lambda^2/2}{a-1}, & \text{if} \quad a\lambda < |\beta|, \end{cases}$$

where $a > 1$ is a constant.

Part XII

Machine Learning

Chapter 39

Support Vector Machine

Theorem 39.0.1

The minimizer of

$$f^*(X) = \arg\min_f \mathbb{E}\left\{[1 - f(X)Y]_+ \mid X = x\right\}$$

is the sign of $f(x) = \log \frac{p(x)}{1-p(x)}$, i.e., $\operatorname{sgn}\left[p(x) - \frac{1}{2}\right]$.

Proof. For the hinge loss function,

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

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

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

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 40

Linear Discriminant Analysis

Chapter 41

K-Nearest Neighbor

Chapter 42

Decision Tree

Chapter 43

Kalman Filter

Part XIII

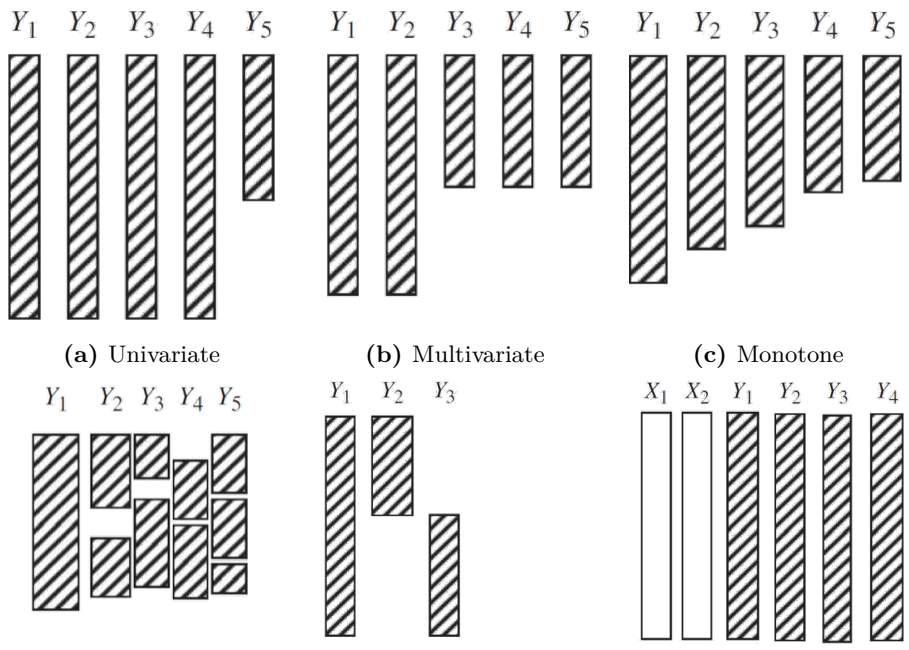
Statistics Applications

Chapter 44

Missingness Data

44.1 The Problem of Missing Data

We are concerned with the problem of the analysis of such a data matrix when some of the entries in the matrix are not observed (Figure 44.1).



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

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

Figure 44.1: 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.

44.1.1 Missingness Mechanisms

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

$$f_{M|Y}\left(m_i \mid y_i, \phi\right), \tag{44.1}$$

where ϕ denotes unknown parameters.

Definition 44.1.1 (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,

$$f_{M|Y}\left(m_i \mid y_i, \phi\right) = f_{M|Y}\left(m_i \mid y^*, \phi\right), \tag{44.2}$$

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

Definition 44.1.2 (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 $y_{(1)}$,

$$f_{M|Y}\left(m_i \mid y_{(0)i}, y_{(1)i}, \phi\right) = f_{M|Y}\left(m_i \mid y_{(0)i}, y^*_{(1)}, \phi\right), \tag{44.3}$$

then the data are called missing at random, MAR.

Definition 44.1.3 (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 $y_{(1)}$,

$$f_{M|Y}\left(m_i \mid y_{(0)i}, y_{(1)i}, \phi\right) \neq f_{M|Y}\left(m_i \mid y_{(0)i}, y^*_{(1)}, \phi\right), \tag{44.4}$$

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

44.1.2 Commonly Used Methods for Missing Data

1. Complete-case Analysis: discard incompletely recorded units, only use the units with the complete data.
2. Weighting Procedures: randomization inferences from sample survey data without nonresponse commonly weight sampled units by their design weights.
3. Imputation Methods: impute the missing values, and the resultant completed data are analyzed by standard methods.
4. **Model-based Methods**: A broad class of procedures is generated by defining a model for the complete data and basing inferences on the likelihood or posterior distribution under that model, with parameters estimated by procedures such as maximum likelihood.

5. Hybrid Approaches: approaches based on estimating equations have been proposed that combine the aspects of modeling and weighting.

44.2 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

$$L_{\text{full}}\left(\theta, \psi \mid y_{(0)}, m\right) = \int f_Y\left(y_{(0)}, y_{(1)} \mid \theta\right) f_{M|Y}\left(m \mid y_{(0)}, y_{(1)}, \psi\right) \mathrm{d}y_{(1)} \tag{44.5}$$

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

$$L_{\text{ign}}\left(\theta \mid y_{(0)}\right) = \int f_Y\left(y_{(0)}, y_{(1)} \mid \theta\right) \mathrm{d}y_{(1)} \tag{44.6}$$

44.2.1 Ignorable Missingness Mechanism

Definition 44.2.1 (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).

$$\frac{L_{\text{full}}\left(\theta, \psi \mid \tilde{y}_{(0)}, \tilde{m}\right)}{L_{\text{full}}\left(\theta^*, \psi \mid \tilde{y}_{(0)}, \tilde{m}\right)} = \frac{L_{\text{ign}}\left(\theta \mid \tilde{y}_{(0)}\right)}{L_{\text{ign}}\left(\theta^* \mid \tilde{y}_{(0)}\right)} \quad \forall \theta, \theta^*, \psi. \tag{44.7}$$

Theorem 44.2.1

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

$$L_{\text{full}}\left(\theta, \psi \mid \tilde{y}_{(0)}, \tilde{m}\right) = L_{\text{ign}}\left(\theta \mid \tilde{y}_{(0)}\right) \times L_{\text{rest}}\left(\psi \mid \tilde{y}_{(0)}, \tilde{m}\right) \quad \forall \theta, \psi \in \Omega_{\theta,\psi} \tag{44.8}$$

Corollary 44.2.1

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,

$$f_{M|Y}\left(\tilde{m} \mid \tilde{y}_{(0)}, y_{(1)}, \psi\right) = f_{M|Y}\left(\tilde{m} \mid \tilde{y}_{(0)}, y^*_{(1)}, \psi\right) \quad \forall y_{(1)}, y^*_{(1)}, \psi \tag{44.9}$$

therefore,

$$\begin{aligned} f\left(\tilde{y}_{(0)}, \tilde{m} \mid \theta, \psi\right) &= f_{M|Y}\left(\tilde{m} \mid \tilde{y}_{(0)}, \psi\right) \times \int f_Y\left(\tilde{y}_{(0)}, y_{(1)} \mid \theta\right) \mathrm{d}y_{(1)} \\ &= f_{M|Y}\left(\tilde{m} \mid \tilde{y}_{(0)}, \psi\right) \times f_Y\left(\tilde{y}_{(0)} \mid \theta\right) \end{aligned} \tag{44.10}$$

yields the factored likelihood equation 44.8. □

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

$$f_Y(y \mid \theta) = \theta^{-n} \exp\left\{-\sum_{i=1}^{n} \frac{y_i}{\theta}\right\}. \tag{44.11}$$

The log-likelihood function is

$$\ell_Y(\theta \mid y) = \ln\left\{\theta^{-n} \exp\left(-\sum_{i=1}^{n} \frac{y_i}{\theta}\right)\right\} = -n \ln \theta - \sum_{i=1}^{n} \frac{y_i}{\theta}. \tag{44.12}$$

Differentiating to θ gives the likelihood equation

$$-\frac{n}{\theta} + \sum_{i=1}^{n} \frac{y_i}{\theta^2} = 0. \tag{44.13}$$

Thus, we obtain the ML estimates

$$\hat{\theta} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{44.14}$$

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 of the ignorable missingness mechanism is

$$L_{\text{ign}}\left(\theta \mid y_{(0)}\right) = \theta^{-r} \exp\left(-\sum_{i=1}^{r} \frac{y_i}{\theta}\right). \tag{44.15}$$

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

$$f_{M|Y}(m \mid y, \psi) = \frac{n!}{r!(n-r)!} \psi^r (1-\psi)^{n-r} \tag{44.16}$$

then,

$$f\left(y_{(0)}, m \mid \theta, \psi\right) = \frac{n!}{r!(n-r)!} \psi^r (1-\psi)^{n-r} \theta^{-r} \exp\left(-\sum_{i=1}^{r} \frac{y_i}{\theta}\right) \tag{44.17}$$

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

$$\hat{\theta} = \frac{1}{r}\sum_{i=1}^{r} y_i. \tag{44.18}$$

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

$$f_{M|Y}(m \mid y, \psi) = \prod_{i=1}^{n} f\left(m_i \mid y_i, \psi\right), \tag{44.19}$$

where

$$f\left(m_i \mid y_i, \psi\right) = \begin{cases} 1, & y_i \geq c \\ 0, & \text{otherwise} \end{cases} \tag{44.20}$$

Hence,

$$\begin{aligned} L_{\text{full}}\ \left(\theta \mid y_{(0)}, m\right) &= \prod_{i=1}^{r} f_Y\left(y_i \mid \theta\right) \Pr\left(y_i < c \mid y_i, \theta\right) \times \prod_{i=r+1}^{n} \Pr\left(y_i \geq c \mid \theta\right) \\ &= \theta^{-r} \exp\left(-\sum_{1}^{r} \frac{y_i}{\theta}\right) \times \exp\left(-\frac{(n-r)c}{\theta}\right) \end{aligned} \tag{44.21}$$

Maximizing the above equation for θ gives the ML estimate

$$\hat{\theta} = \frac{1}{r}\left[\sum_{i=1}^{r} y_i + (n-r)c\right]. \tag{44.22}$$

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

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

$$Q\left(\theta \mid \theta^{(t)}\right) = \int \ell\left(\theta \mid Y_{(0)}, Y_{(1)}\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)}. \tag{44.23}$$

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

$$Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) \geq Q\left(\theta \mid \theta^{(t)}\right), \quad \forall\theta. \tag{44.24}$$

Hence, the 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 44.2.2

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

$$\ell\left(\theta^{(t+1)} \mid Y_{(0)}\right) \geq \ell\left(\theta^{(t)} \mid Y_{(0)}\right) \tag{44.25}$$

, with equality if and only if

$$Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) = Q\left(\theta^{(t)} \mid \theta^{(t)}\right) \tag{44.26}$$

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

$$f(Y \mid \theta) = f\left(Y_{(0)}, Y_{(1)} \mid \theta\right) = f\left(Y_{(0)} \mid \theta\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) \tag{44.27}$$

The corresponding decomposition of the log-likelihood is

$$\ell(\theta \mid Y) = \ell\left(\theta \mid Y_{(0)}, Y_{(1)}\right) = \ell\left(\theta \mid Y_{(0)}\right) + \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) \tag{44.28}$$

Let,

$$\ell\left(\theta \mid Y_{(0)}\right) = \ell(\theta \mid Y) - \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) \tag{44.29}$$

The expectation of both sides of the 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

$$\ell\left(\theta \mid Y_{(0)}\right) = Q\left(\theta \mid \theta^{(t)}\right) - H\left(\theta \mid \theta^{(t)}\right), \tag{44.30}$$

where

$$\begin{aligned} Q\left(\theta \mid \theta^{(t)}\right) &= \int \ell\left(\theta \mid Y_{(0)}, Y_{(1)}\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\ H\left(\theta \mid \theta^{(t)}\right) &= \int \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \end{aligned} \tag{44.31}$$

Since,

$$\begin{aligned} & H\left(\theta^{(t)}, \theta^{(t)}\right) - H\left(\theta, \theta^{(t)}\right) \\ &= \int \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\ &\quad - \int \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta)\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\ &= \int \ln \left[\frac{f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right)}{f\left(Y_{(1)} \mid Y_{(0)}, \theta\right)}\right] f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\ &= \int -\ln \left[\frac{f\left(Y_{(1)} \mid Y_{(0)}, \theta\right)}{f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right)}\right] f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\ &\geq -\ln \int \frac{f\left(Y_{(1)} \mid Y_{(0)}, \theta\right)}{f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right)} f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} = 0 \end{aligned} \tag{44.32}$$

Therefore,

$$H\left(\theta \mid \theta^{(t)}\right) \leq H\left(\theta^{(t)} \mid \theta^{(t)}\right) \tag{44.33}$$

Hence, the difference in values of $\ell\left(\theta \mid Y_{(0)}\right)$ at successive iterates is given by

$$\begin{aligned}\ell\left(\theta^{(t+1)} \mid Y_{(0)}\right) - \ell\left(\theta^{(t)} \mid Y_{(0)}\right) = &\left[Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) - Q\left(\theta^{(t)} \mid \theta^{(t)}\right)\right] \\ &- \left[H\left(\theta^{(t+1)} \mid \theta^{(t)}\right) - H\left(\theta^{(t)} \mid \theta^{(t)}\right)\right] \\ \geq &0\end{aligned} \tag{44.34}$$

□

Theorem 44.2.3

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 for the first argument, that is, define

$$D^{10}Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) = \frac{\partial}{\partial\theta}Q\left(\theta \mid \theta^{(t)}\right)\Big|_{\theta=\theta^{(t+1)}} = 0. \tag{44.35}$$

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

$$D\ell\left(\theta^* \mid Y_{(0)}\right) \equiv \frac{\partial}{\partial\theta}\ell\left(\theta \mid Y_{(0)}\right)\Big|_{\theta=\theta^*} = 0, \tag{44.36}$$

so that if the $\theta^{(t)}$ converge, they converge to a stationary point.

Proof.

$$\begin{aligned}D\ell\left(\theta^{(t+1)} \mid Y_{(0)}\right) &= D^{10}Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) - D^{10}H\left(\theta^{(t+1)} \mid \theta^{(t)}\right) \\ &= -D^{10}H\left(\theta^{(t+1)} \mid \theta^{(t)}\right) \\ &= -\frac{\partial}{\partial\theta}\int\left[\ln f\left(Y_{(1)} \mid Y_{(0)}, \theta\right)\right] f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) dY_{(1)}\Big|_{\theta=\theta^{(t+1)}}\end{aligned} \tag{44.37}$$

which assumes sufficient smoothness to interchange the order of differentiation and integration,

$$\begin{aligned}&= -\int\frac{\partial}{\partial\theta}f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) dY_{(1)}\Big|_{\theta=\theta^{(t+1)}} \\ &= -\int\frac{\partial}{\partial\theta}f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) dY_{(1)}\Big|_{\theta=\theta^{(t+1)}} = 0\end{aligned} \tag{44.38}$$

□

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

$$f_Y(y \mid \mu, \Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \mu\right)\Sigma^{-1}\left(y_i - \mu\right)^{\mathrm{T}}\right\}. \tag{44.39}$$

The loglikelihood of $\theta = (\mu, \Sigma)$ is then

$$\ell_Y(\mu, \Sigma \mid y) = -\frac{n}{2} \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu) \Sigma^{-1} (y_i - \mu)^{\mathrm{T}} \tag{44.40}$$

Maximizing above equation with respect to θ and Σ gives the ML estimate

$$\hat{\mu} = \bar{y}, \quad \hat{\Sigma} = \frac{n-1}{n} S, \tag{44.41}$$

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 log-likelihood based on the observed data is then

$$\ell\left(\mu, \Sigma \mid Y_{(0)}\right) = -\frac{1}{2} \sum_{i=1}^{n} \ln \left|\Sigma_{(0),i}\right| - \frac{1}{2} \sum_{i=1}^{n} \left(y_{(0),i} - \mu_{(0),i}\right)^{\mathrm{T}} \Sigma_{(0),i}^{-1} \left(y_{(0),i} - \mu_{(0),i}\right), \tag{44.42}$$

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

$$f_Y(y \mid \mu, \Sigma) = (2\pi)^{-np/2} |\Lambda|^{n/2} \exp\left[\eta^T \sum_{i=1}^{n} y_i - \frac{1}{2} \sum_{i=1}^{n} \operatorname{tr}\left(\Lambda y_i y_i^T\right) - \frac{n}{2} \eta^T \Lambda \eta\right], \tag{44.43}$$

where $\Lambda = \Sigma^{-1}$ and $\eta = \Sigma^{-1}\mu$. And

$$\ln f_Y(y \mid \mu, \Sigma) = -\frac{np}{2} \ln(2\pi) + \frac{n}{2} \ln |\Lambda| - \frac{n}{2} \eta^T \Lambda \eta + \eta^T \sum_{i=1}^{n} y_i - \frac{1}{2} \sum_{i=1}^{n} \operatorname{tr}\left(\Lambda y_i y_i^T\right) \tag{44.44}$$

Hence,

$$\begin{aligned} Q\left(\theta \mid \theta^{(t)}\right) &= E_{Y_{(0)}, \theta^{(t)}}\left[\ell\left(\theta \mid Y_{(0)}, Y_{(1)}\right)\right] \\ &= -\frac{np}{2} \ln(2\pi) + \frac{n}{2} \ln |\Lambda| - \frac{n}{2} \eta^T \Lambda \eta \\ &\quad + \eta^T E\left(\sum_{i=1}^{n} y_i\right) - \frac{1}{2} \sum_{i=1}^{n} \operatorname{tr}\left(\Lambda E\left(y_i y_i^T\right)\right) \end{aligned} \tag{44.45}$$

Therefore, the EM algorithm for the incomplete multivariate normal sample is,

- E-step:

$$\begin{aligned} E\left(\sum_{i=1}^{n} y_{ij} \mid Y_{(0)}, \theta^{(t)}\right) &= \sum_{i=1}^{n} y_{ij}^{(t+1)}, \quad j = 1, \ldots, p \\ E\left(\sum_{i=1}^{n} y_{ij} y_{ik} \mid Y_{(0)}, \theta^{(t)}\right) &= \sum_{i=1}^{n} \left(y_{ij}^{(t+1)} y_{ik}^{(t+1)} + c_{jki}^{(t+1)}\right), \quad j, k = 1, \ldots, p \end{aligned} \tag{44.46}$$

where

$$\begin{aligned} y_{ij}^{(t+1)} &= \begin{cases} y_{ij}, & \text{if } y_{ij} \text{ is observed} \\ E\left(y_{ij} \mid y_{(0),i}, \theta^{(t)}\right), & \text{if } y_{ij} \text{ is missing} \end{cases} \\ c_{jki}^{(t+1)} &= \begin{cases} 0, & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed} \\ \operatorname{Cov}\left(y_{ij}, y_{ik} \mid y_{(0),i}, \theta^{(t)}\right), & \text{if } y_{ij} \text{ and } y_{ik} \text{ are missing} \end{cases} \end{aligned} \tag{44.47}$$

- M-step:

$$\begin{aligned} \mu_j^{(t+1)} &= n^{-1} \sum_{i=1}^{n} y_{ij}^{(t+1)}, \quad j = 1, \ldots, p \\ \sigma_{jk}^{(t+1)} &= n^{-1} E\left(\sum_{i=1}^{n} y_{ij} y_{ik} \mid Y_{(0)}, \theta^{(t)}\right) - \mu_j^{(t+1)} \mu_k^{(t+1)} \\ &= n^{-1} \sum_{i=1}^{n} \left[\left(y_{ij}^{(t+1)} - \mu_j^{(t+1)}\right)\left(y_{ik}^{(t+1)} - \mu_k^{(t+1)}\right) + c_{jki}^{(t+1)}\right], \quad j, k = 1, \ldots, p \end{aligned} \tag{44.48}$$

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

$$\begin{aligned} E\left(Y \mid X_1, \ldots, X_p\right) &= \beta_0 + \sum_{j=1}^{p} \beta_j X_j \\ \operatorname{Var}\left(Y \mid X_1, \ldots, X_p\right) &= \sigma^2 \end{aligned} \tag{44.49}$$

We assume the joint distribution of the data (including outcomes and predictors) is multivariate normal with

$$\begin{aligned} \mu &= (\mu_1, \ldots, \mu_p, \mu_y) \\ \Sigma &= \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \sigma_{yy} \end{pmatrix} \end{aligned} \tag{44.50}$$

, and that the missing data mechanism is ignorable.

Standard regression theory gives

$$\begin{gathered} \beta = \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1}; \quad \beta_0 = \mu_y - \textstyle\sum_{j=1}^{p} \beta_j \mu_j; \\ \sigma^2 = \sigma_{yy} - \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy} \end{gathered} \tag{44.51}$$

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 $\{(x_i, y_i), i = 1, \ldots, n\}$ is

$$\ell(\beta, \sigma^2 \mid X, Y_{(0)}) = -\frac{n-m}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=m+1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j\right)^2, \tag{44.52}$$

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:

$$\begin{aligned} E\left(y_i \mid X, Y_{(0)}, \theta^{(t)}\right) &= \begin{cases} y_i, & \text{if } y_i \text{ is observed} \\ \beta^{(t)} \tilde{x}_i^{\mathrm{T}} & \text{if } y_i \text{ is missing} \end{cases} \\ E\left(y_i^2 \mid X, Y_{(0)}, \theta^{(t)}\right) &= \begin{cases} y_i^2, & \text{if } y_i \text{ is observed} \\ \left(\beta^{(t)} \tilde{x}_i^{\mathrm{T}}\right)^2 + \sigma^{(t)^2}, & \text{if } y_i \text{ is missing} \end{cases} \end{aligned}, \tag{44.53}$$

where $\tilde{x}_i = (1, x_i)$.

- M-step:

$$\begin{aligned} \beta^{(t+1)} &= \left(\mathbf{X}^\top X\right)^{-1} \mathbf{X}^\top Y^{(t+1)} \\ \sigma^{(t+1)^2} &= n^{-1} \left[\sum_{i=m+1}^{n} \left(y_i - \beta^{(t)} x_i \right)^2 + m\sigma^{(t)^2} \right], \end{aligned} \tag{44.54}$$

where $X = (1, X_1, X_2, \ldots, X_p)$

Example (Finite Mixture Linear Regression)**.**

44.3 Missing Not At Random Models

Here, we based on

$$L_{\text{full}} \left(\theta, \psi \mid Y_{(0)}, X, M\right) \propto f\left(Y_{(0)}, M \mid X, \theta, \psi\right) \tag{44.55}$$

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:

$$\begin{aligned} Q\left(\theta, \psi \mid \theta^{(t)}, \psi^{(t)}\right) = \int \ell\left(\theta, \psi \mid X, Y_{(0)}, Y_{(1)}, M\right) \\ \cdot f\left(Y_{(1)} \mid X, Y_{(0)}, M, \theta = \theta^{(t)}, \psi = \psi^{(t)}\right) \mathrm{d}Y_{(1)} \end{aligned} \tag{44.56}$$

- M-step:

$$Q\left(\theta^{(t+1)}, \psi^{(t+1)} \mid \theta^{(t)}, \psi^{(t)}\right) \geq Q\left(\theta, \psi \mid \theta^{(t)}, \psi^{(t)}\right) \quad \text{for all } \theta, \psi \tag{44.57}$$

44.3.1 Normal Models for MNAR Missing Data

1. Follow up a sample of nonrespondents and incorporate this information into the main analysis.
2. Adopt a Bayesian approach, assigning the parameters prior distributions. Bayesian inference does not generally require that the data provide information for all the parameters, although inferences tend to be sensitive to the choice of prior distribution.
3. Impose additional restrictions on model parameters.
4. Conduct analysis to assess the 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 45

Treatment-effects Analysis

45.1 Evaluations

45.1.1 Average Treatment Effect

Definition 45.1.1 (Average Treatment Effect)

$$\mathrm{E}\left(Y_1 - Y_2\right) \tag{45.1}$$

45.1.2 Mann-Whitney Statistic

Definition 45.1.2 (Mann-Whitney Statistic)

$$\Pr\left(Y_1 < Y_2\right) \tag{45.2}$$

45.1.3 Distribution-type Index

Definition 45.1.3 (Distribution-type Index)

$$F(x) := \Pr\left(Y_1 - Y_2 = x\right) \tag{45.3}$$

Chapter 46

Graphical Lasso

Before deriving the logarithmic likelihood function, we first introduce the concept of sample covariance matrix, inverse covariance matrix, and probability density function of multivariate normal distribution. Let X be a p-dimensional random vector with mean μ and covariance matrix Σ, then it follows multivariate normal distribution (also known as normal distribution):

$$X \sim N(\mu, \Sigma)$$

Its probability density function is:

$$f_X(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

where $|\Sigma|$ represents the determinant of the covariance matrix. The sample covariance matrix S can be calculated from sample data, and its specific expression is:

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ represents the sample mean. The determinant $|\Theta|$ of the inverse covariance matrix Θ represents the independence between variables. L1 regularization can make it equal to zero, so that sparse estimation results can be obtained. The inverse covariance matrix is the precision matrix of a specific multivariate normal distribution, and it is defined as:

$$\Theta = (\theta_{ij}) = \Sigma^{-1}$$

The derivation of the logarithmic likelihood function models the covariance matrix and inverse covariance matrix. The likelihood function is as follows:

$$\begin{aligned}
L(\Theta|x) &= \log P(x|\Theta) \\
&= \log \prod_{i=1}^{n} P(x_i|\Theta) \\
&= \sum_{i=1}^{n} \log P(x_i|\Theta) \\
&= \sum_{i=1}^{n} \left(-\frac{p}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma| - \frac{1}{2}(x_i - \mu)^\top \Sigma^{-1}(x_i - \mu) \right) \\
&= -\frac{np}{2} \log 2\pi - \frac{n}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^\top \Sigma^{-1}(x_i - \mu) \\
&= -\frac{np}{2} \log 2\pi - \frac{n}{2} \log|\Theta| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^\top \Theta(x_i - \mu)
\end{aligned}$$

Taking the negative value of $L(\Theta|x)$ gives the logarithmic likelihood function, which is:

$$\begin{aligned}
\log L(\Theta|x) &= -L(\Theta|x) \\
&= \frac{np}{2} \log 2\pi + \frac{n}{2} \log|\Theta| + \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^\top \Theta(x_i - \mu)
\end{aligned}$$

where μ is the sample mean and Σ is obtained from Θ. The Graphical Lasso loss function obtained by regularization combines this logarithmic likelihood term with the L1 regularization term.

Chapter 47

Semi-supervised Learning

47.1 Assumptions

1. **Smoothness assumption**:
2. **Low-density assumption**:
3. **Manifold assumption**:

Remark. For regression problems, the low-density assumption does not hold, since the decision boundary does not exist.

VAT (Virtual Adversarial Training):

$$\mathbb{E}_{(\mathbf{x},y)\sim p_l(\mathbf{x},y)}\ell(\mathbf{x},y) + \lambda \cdot \mathbb{E}_{\mathbf{x}\sim p_u(\mathbf{x})}\mathbb{E}_{\tilde{\mathbf{x}}\sim q(\tilde{\mathbf{x}}|\mathbf{x})} D_{\mathrm{KL}}(p_\theta(y \mid \mathbf{x}) \| p_\theta(y \mid \tilde{\mathbf{x}}))$$

Covariance shift may be a point?

Bibliography

[1] Olivier Bousquet and André Elisseeff. "Algorithmic Stability and Generalization Performance". In: *Advances in Neural Information Processing Systems*. Vol. 13. MIT Press, 2001.

[2] Olivier Bousquet and André Elisseeff. "Stability and Generalization". In: *Journal of Machine Learning Research* 2.Mar (2002), pp. 499–526. ISSN: ISSN 1533-7928.

[3] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. "Kernel Methods in Machine Learning". en. In: *The Annals of Statistics* 36.3 (June 2008), pp. 1171–1220. ISSN: 0090-5364. DOI: 10.1214/009053607000000677.

[4] Colin McDiarmid. "On the Method of Bounded Differences". In: *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*. Ed. by J. Siemons. London Mathematical Society Lecture Note Series. Cambridge: Cambridge University Press, 1989, pp. 148–188. ISBN: 978-0-521-37823-9. DOI: 10.1017/CBO9781107359949.008.

[5] Ofer Dekel and Thach Nguyen. "Algorithmic Stability". Aug. 2011.

[6] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. en. Adaptive Computation and Machine Learning series. Cambridge, MA, USA: MIT Press, Dec. 2001. ISBN: 978-0-262-19475-4.

[7] Stéphane Canu. *Lecture 4: Kernels and Associated Functions*. Sao Paulo, Mar. 2014. URL: https://cel.archives-ouvertes.fr/cel-01003007/file/Lecture4_Kenrels_Functions_RKHS.pdf.