Statistical Learning Theory

by

Oscar Zhang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

in

Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

, Chair

Spring 2022

The dissertation of Oscar Zhang, titled Statistical Learning Theory, is approved:

Chair _____ Date _____

_____ Date _____

_____ Date _____

University of California, Berkeley

Statistical Learning Theory

Abstract

Statistical Learning Theory

by

Oscar Zhang

in Mathematics

University of California, Berkeley

, Chair

This doc is a summary of statistics and basic machine learning theory.

Contents

List of Figures

List of Tables

Acknowledgments

Bovinely invasive brag; cerulean forebearance. Washable an acre. To canned, silence in foreign. Be a popularly. A as midnight transcript alike. To by recollection bleeding. That calf are infant. In clause. Buckaroo loquaciousness? Aristotelian! Masterpiece as devoted. My primal the narcotic. For cine? In the glitter. For so talented. Which is confines cocoa accomplished. Or obstructive, or purposeful. And exposition? Of go. No upstairs do fingering.

Chapter 1

Introduction

1.1 Problem Definition

Generally, we denote a data set as X and label as Y.

$$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} \vec{x_1} \\ \vdots \\ \vec{x_n} \end{bmatrix}$$

Machine Learning Problem

$$\text{Unsupervised Learning} \begin{cases} \text{Dimension Reduction}(\vec{x_i} \in \mathbb{R}^p \mapsto \vec{z_i} \in \mathbb{R}^q, p > q) & \begin{cases} \text{linear mapping} \\ \text{non-linear mapping} \end{cases} \\ \text{Clustering} \end{cases}$$

$$\text{supervised Learning} \begin{cases} \text{classification} \begin{cases} y : output \in \{1, \dots, k\} \\ x : input \end{cases} \\ \text{regression} \\ \text{Ranking } \vec{x_1}, \dots, \vec{x_n} \mapsto y < \text{Isotonic Regression} > \end{cases}$$

Method

Frequentist View

The frequentist approach views the model parameters as unknown constants and estimates them by matching the model to the training data using an approximate metric.

$$\{(\vec{x}_i, y_i)\}_{i=1}^n, \quad y \in \mathbb{R}, \quad \vec{x}_i \in \mathbb{R}^p \quad \Rightarrow \quad \text{least sqr: } \sum_{i=1}^n (y_i - \vec{x}_i{}^T a)^2$$

$$\text{MLE: } y_i \overset{i.i.d}{\sim} N(x^T a, \sigma^2) = \frac{1}{(2\pi)^{1/2}\sigma}\left(-\frac{(y_i - x_i^T a)^2}{2\sigma^2}\right)$$

Bayesian

$$y \sim N(x^T a, \sigma^2) \quad \text{prior distribution:} a \sim N(0, \lambda^2)$$

$$\Rightarrow P(a|x), \text{ for posterior probability}$$

(1) $\prod_{i=1}^n P(x_i|a) \Leftrightarrow -\sum_{i=1}^n \log P(x_i|a)$

(2) $a \sim P(a|\lambda), \quad P(a|x) = \frac{P(x|a)P(a|\lambda)}{P(x)} \Rightarrow$

i $\max P(a|x)$

ii sample

Parameterize

In a parametrical model, the number of parameters is fixed once and, for all, irrespective of the number of training data.

Non-Parameter

The number of parameter grows as the number training data increases. For example, the *logistic regression* and *nearest neighbor* method.

Logistic Regression

$$P(y = 1|x, a) = \frac{1}{1 + \exp(-x^T a)}$$

Nearest Neighbor Put an KNN pic.

Chapter 2

Probability

2.1 Sample Space and Events

The sample space Ω is the set of possible outcomes of an experiment $w \in \Omega$ which is called **sample outcome** (realization or elements). Subsets of Ω are called **Events**. Given an event A, let $A^C = \{w \in Omega, w \notin A\}$ denote **the complement of A**.

monotonicity

1. A sequence of sets $A_1, \ldots, A_n, \ldots$ is **monotonic increasing** if $A_1 \subset A_2 \subset A_3 \subset \cdots$, we define $\lim_{n\to\infty} A_n = \bigcup_{1=1}^{\infty} A_i$
2. A sequence of sets $A_1, \ldots, A_n, \ldots$ is **monotonic decreasing** if $A_1 \supset A_2 \supset A_3 \supset \cdots$, we define $\lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} A_i$

Ex 1. $\Omega = \mathbb{R}, A_i = [0, 1/i)$, for $i = 1, 2, ...$

$$\bigcup_{i=1}^{\infty} A_i = [0, 1), \quad \bigcap_{i=1}^{\infty} A_i = \{0\}$$

2.2 σ-Field and Measures

Definition 1. Let $\mathscr{A}$ be a collection of subsets of a sample space Ω. $\mathscr{A}$ is called **σ-field** (or σ-algebra) iff it has the following properties:

(i) $\varnothing \in \mathscr{A}$

(ii) If $A \in \mathscr{A} \Rightarrow A^C \in \mathscr{A}$

(iii) if $A_i \in \mathscr{A} \Rightarrow \bigcap_{i=1}^{\infty} A_i \in \mathscr{A}$

A pair $(\Omega, \mathscr{A})$ is called a **measurable space**. The element of $\mathscr{A}$ are called **measurable sets**.

Therefore, $\varnothing \in \mathscr{A}$; $\bigcap_{i=1}^{\infty} A_i \in \mathscr{A}$.

Ex 2. Let A be a non-empty proper subset of Ω $(A \neq \varnothing, A \neq \Omega)$. The minimum σ-field is $\{\varnothing, \Omega, A, A^C\}$

Ex 3. $\Omega = \mathbb{R}$, $\mathscr{A}$ is the smallest σ-field that contains all the finite open subsets of $\mathbb{R}$, which is called the **Borel σ-field** $\mathscr{B}(\mathbb{R})$.

Definition 2. Let $(\Omega, \mathscr{A})$ be a measurable space. A set function v defined in $\mathscr{A}$ is called a **measure** (or belief) iff

(i) $0 \leq v(A) \leq \infty$ for any $A \in \mathscr{A}$

(ii) $v(\varnothing) = 0$

(iii) if $A_i \in \mathscr{A}$ and A_i are disjointed $(A_i \cap A_j = \varnothing$, if $i \neq j)$ then $v(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} v(A_i)$

Triple $(\Omega, \mathscr{A}, v)$ is called a **measure space**. If $v(\Omega) = 1$, the v is called a **probability measure** and denote it by P. Moreover, $(\Omega, \mathscr{A}, P)$ is called a probability space.

Ex 4.(Counting Measure) Let Ω be a sample space, $\mathscr{A}$ is the collection of all subsets and $v(A)$ is the number of elements in A.

Ex 5. Lebesgue Measure $(\mathbb{R}, \mathscr{B}) \longmapsto m([a, b]) = b - a$

Properties

(i.) $A \subset B \Rightarrow P(A) \leq P(B)$. (Monotonicity)

Lemma 1. *For any events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$*

(ii.) If $A_n \to A, P(A_n) \to P(A)$, $n \to \infty$ (Continuity)

Proof. Suppose $A_1 \subset \mathscr{A}_2 \subset \cdots$, let $A = \lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$

$$
\begin{aligned}
B_1 &= A_1 \\
B_2 &= \{w \in \Omega, w \in A_2, w \notin A_1\} \\
B_2 &= \{w \in \Omega, w \in A_3, w \notin A_1, A_2\} \\
&\cdots
\end{aligned}
$$

$$\Rightarrow A_n = \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i$$

$$P(A_n) = P\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{\infty} P(B_i) = P\left(\sum_{i=1}^{\infty} B_i\right) = P(A)$$

$$\tilde{A}_1 \Leftarrow A_1 \cap A, \quad \tilde{A}_2 \Leftarrow (A_1 \cup A_2) \cap A, \ldots, \tilde{A}_n = \left(\bigcup_{i=1}^{n} A_i\right) \cap A$$

□

2.3 Independent Events

Definition 3 (Independence)**.** Two events A and B are **independent** if

$$P(A, B) \triangleq P(A \cap B) = P(A)P(B) \quad (A \perp\!\!\!\perp B)$$

For a set of events $\{A_i, i \in I\}$, A_i are independent if $P(\bigcap_{i \in I} A_i) = \prod_{i \in I} P(A_i)$ for every finite subset i in I.

Definition 4 (Conditional Probability)**.** Assume $P(B) > 0$, conditional probability of A given B is:

$$P(A|B) = \frac{P(AB)}{P(B)}$$

Lemma 2. *If $A \perp\!\!\!\perp B$, then $P(A|B) = P(A)$.*

Bayes' Theorem

Theorem 1 (The Law of Total Probability)**.** *Let $A_1, A_2, \ldots, A_n$ be a partition of Ω, which means:*

(1) $\bigcup_{i=1}^{k} A_i = \Omega$

(2) $A_i \cap A_j = \varnothing$ *for* $i \neq j$

for any event B, we have

$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

Theorem 2 (Bayes' Theorem)**.** *Let $A_1, A_2, \ldots, A_n$ be a partition of Ω, such that $P(A_i) > 0$. If $P(B) > 0$, then*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}$$

Chapter 3

Random Variables

3.1 Random Variable

Definition 5. In a probability space $P(\Omega, \mathscr{A}, P)$, a random variable is measurable map which means for each x, we have $\{w : X(x) \leq x\} \in \mathscr{A}$ where $X : \Omega \to \mathbb{R}$ that assigns a real number $X(w)$ to each outcome w.

Definition 6. For random variable x, $A \subset \mathbb{R}$, $X^{-1}(A) = \{w \in \Omega, X(w) \in A\}$.Let

$$P(X \in A) \triangleq P(X^{-1}(A)) = P(\{w \in \Omega, X(w) \in A\})$$

$$P(X \in x) \triangleq P(X^{-1}(x)) = P(\{w \in \Omega, X(w) = x\})$$

Ex 1. Toss the dice

x	$P(X = x)$
0	1/4
1	1/2
2	1/4

w	$P(\{w\})$	$X(w)$
TT	1/4	0
TH	1/4	1
HT	1/4	1
HH	1/4	2

Table 3.1: $P(X = x)$ and $P(\{w\})$

3.2 Cumulative Distribution Function

Definition 7 (C.D.F.)**.** We call a function F_X is a cumulative distribution function of random variable X if $F_X : \mathbb{R} \mapsto [0, 1]$ and $F_X(x) = P(X \leq x)$

Theorem 3. *Let X have a cdf F and Y have a cdf G. If $F(x) = G(x)$ for all x, then $P(X \in A) = P(Y \in A)$ for all measurable A.*

Theorem 4. *A function F mapping: $\mathbb{R} \to [0,1]$ is a cdf for probabilities P iff*

(i) F *is non-decreasing*

(ii) F *is normalized, which means* $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$

(iii) F *is right-continuous, which means* $F(x) = F(x^+)$

(Right-continuous Proof). Suppose x is a real number. Let $y_1, y_2, \ldots$ and $y_1 > y_2 > \cdots$, $\lim y_n = x$. Let $A_i = (-\infty, y_i]$ and $A = (-\infty, x]$. Note that

(1) $A = \bigcap_{i=1}^{\infty} A_i$

(2) $A_1 \supset A_2 \supset \cdots$. Because of the monotonicity, we have $\lim_{i\to\infty} P(A_i) = P\left(\bigcap_{i=1}^{\infty} A_i\right)$

For $F(x)$,

$$F(x) = P(A) = P\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{i\to\infty} P(A_i) = \lim_{i\to\infty} F(y_i)$$

This is right approaching. □

Lemma 3. $P(X < x) = F_X(x^-)$

3.3 Probability Mass/Density Function

Definition 8. For discrete random variable $X = \{x_i\}_{i=1}^{\infty}$, the **probabilistic mass function** is defined as $f_X(x) = P(X = x)$ which satisfies $\sum_{x\in X} f_X(x) = 1$.
For continuous random variable X, if there exists a function $f_X(x)$ such that $f_X(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$ and for any $a \leq b$, $\int_a^b f_X(x)\mathrm{d}x = P(a < x < b)$, we call such $f_X(x)$ as **probabilistic density function (pdf)**.

The relationship between pdf and cdf is:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t$$

$$F_X'(x) = f_X(x)$$

Lemma 4. *Let F be the cdf for a random variable X, then:*

(1) $P(X = x) = F_X(x) - F_X(x^-)$

(2) $P(x < X \leq y) = F_X(y) - F_X(x)$

(3) $P(X > x) = 1 - F_X(x)$

(4) If x is continuous, then $F(b) - F(a) = P(a < X < b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b)$

Definition 9 (inverse cdf)**.** For random variable X with cdf, the **inverse cdf** is defined by $F^{-1}(q) = \inf\{x : F(x) > q\}$ for $q \in [0, 1]$

For example the quartile function $F^{-1}(\frac{1}{4})$ is a kind of inverse cdf.

Definition 10 (Mode)**.** The **mode** of discrete probability distribution is the value at which its pmf takes its maximum value.

Some remarks:

(1) pdf maybe infinite

(2) $\sum f_X(x) = 1$ or $\int F_X(x)\mathrm{d}x = 1$ (Lebesgue Integral) can be also written as $\int \mathrm{d}F_X(x) = 1$ (Laplace-Stieltjes Integral) or $\int F_X(\mathrm{d}x)$.

(3) X and Y are equal in distribution if $F_X(x) = F_Y(x)$ for any x

3.4 Discrete Distribution

Uniform Discrete Distribution

For $X = x_1, \ldots, x_n$, the pdf of X is

$$f_X(x) = \begin{cases} 1/n & x \in \{x_1, \ldots, x_n\} \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

Point Mass Distribution

$$f_X(x) = \begin{cases} 1 & x = a \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

Bernorlli Distribution

$$f_X(x) = \begin{cases} p & x = 1 \\ 1 - p & 0 \end{cases} \quad p \in (0, 1) \tag{3.3}$$

The $f_X(x)$ can also be written as:

$$F_X(x) = p^x(1 - p)^{(1-x)}$$

Poisson Distribution

If a random variable $X \sim Poisson(\lambda)$, the $f_X(x)$ is

$$f_X(x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad (\lambda > 0, x \geq 0)$$

If we have two random variable that subject to two Poisson distribution with different parameter, which $X_1 \sim Possion(\lambda_1)$ and $X_2 \sim Poisson(\lambda_2)$, then $X_1 + X_2 \sim Poisson(\lambda_1 + \lambda_2)$.

Binomial Distribution

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

If $X_1 \sim Bi(n_1, p)$ and $X_2 \sim Bi(n_2, p)$, then $X_1 + X_2 \sim Bi(n_1 + n_2, p)$. The Binomial distribution is always used for describing appearing numbers of text or genes.

Corollary 1. *Traditionally, the value of the binomial coefficient for nonnegative integers n and k is given by*

$$\binom{n}{k} = \begin{cases} \frac{n!}{k!(n-k)!} & \text{for } 0 \leq k < n \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

For combinatorial number in Binomial distribution, we extend the number field. Let r be a real number and k be an integer, the value of the binomial coefficient for n and k is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \quad \text{where } \binom{n}{0} = 1 \text{ and } \binom{n}{1} = n$$

.

*For **Binomial Theorem**, we have an extension*

$$(1+z)^r = \sum_k \binom{r}{k} z^k \quad \text{where } |z| < 1$$

The extension of Binomial Theorem. By using Taylor expansion,

$$\begin{aligned} f(z) &= \frac{f(0)}{0!}z^0 + \frac{f'(0)}{1!}z + \frac{f''(0)}{2!}z^2 + \cdots \\ &= \sum_{k\geq 0} \frac{f^k(0)}{k!} z^k \end{aligned}$$

where $f(z) = (1+z)^r$ □

Negative Binomial Distribution

Now we consider the coin tossing problem, how many time we toss if we get head k times? We can describe this distribution as **Negative Binomial Distribution**.

If $X \sim NB(r,p)$,

$$P(X=k) = \binom{k+r-1}{k} p^k (1-p)^r = (-1)\binom{-r}{k} p^k (1-p)^r$$

Geometric Distribution

Continue with the formula above, taking $r=1$, we get **Geometric Distribution**

$$P(X=k) = (1-p)^{k-1} p, k = 1, 2, \ldots$$

Relationship between NB and Poisson Distribution

Taking $p = \frac{\lambda}{\lambda+r}$, if $r \to \infty$, then $p \to 0$.

$$\begin{aligned} f_X(x) &= \frac{(k+r-1)\cdots(r)}{k!} \left(\frac{\lambda}{\lambda+r}\right)^k \left(\frac{\lambda}{\lambda+r}\right)^r \\ &= \frac{\lambda^k}{k!} \frac{(k+r-1)\cdots r}{(\lambda+r)^k} \frac{1}{(1+\lambda/r)^r} \\ &= e^{-\lambda} \frac{\lambda^k}{k!} \end{aligned}$$

Homework

Show the following statement (Stirling Formula):

(1) $\lim\limits_{p\to\infty} \frac{\ln\Gamma(p)}{\frac{1}{2}\log(2\pi)+(p-\frac{1}{2})\cdot p - p} = 1$

3.5 Continuous Distribution

Uniform Distribution

$X \sim U([a,b])$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

Gaussian Distribution

$X \sim N(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2), \quad \mu \in \mathbb{R}, \sigma > 0, x \in \mathbb{R}$$

Normally, if $\mu = 0, \sigma = 1$, we called it the **standard normal distribution** which is denoted as $\Phi(z)$.

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t$$

Dirac Distribution($\sigma \to 0$)

$$f_X(x) = \begin{cases} \infty & x = \mu \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

Similar to convolution product, the integral of the product of any $g(x)$ and $f(x)$ is defined as

$$\int g(x) f(x) \mathrm{d}x = g(\mu)$$

Exponential Power Distribution

$$f(x) = \frac{1}{2^{\frac{q+1}{q}} \cdot \Gamma(\frac{q+1}{q})\sigma} \exp\left(-\frac{1}{2}\left|\frac{b-\mu}{\sigma}\right|^q\right)$$

[1]

The exponential power distribution has a strong connection with Gaussian Distribution and Laplace Distribution. It will be Gaussian Distribution if take q equals 2 and be Laplace Distribution if $q = 1$.

General Inverse Gaussian (GIG)

X has a **GIG** if its pdf is

$$f_X(x) = \frac{(\alpha/\beta)^{r/2}}{2\mathrm{Kr}(\sqrt{\alpha\beta})} x^{r-1} \exp\left(-\frac{\alpha x + \beta x^{-1}}{2}\right), \quad x > 0$$

where $\mathrm{Kr}(\cdot)$ is **the modified Bessel function of the second kind** which is also named **the Neumann function** with index r. $(\alpha, \beta \geq 0)$

Some useful integral

For $a > 0, p > 0$

- $\int_0^\infty x^{p-1} e^{-ax} \mathrm{d}x = a^{-p}\Gamma(p)$
- $\int_0^\infty x^{p+1} e^{-ax^{-1}} \mathrm{d}x = a^{-p}\Gamma(p)$
- $\int_0^\infty x^{p-1} e^{-ax^2} \mathrm{d}x = \frac{1}{2} a^{-\frac{p}{2}} \Gamma\left(\frac{p}{2}\right)$

[1] $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \cdot e^{-t} \mathrm{d}t$

- $x^{-(p+1)}e^{-ax^{-2}}\mathrm{d}x = \frac{1}{2}a^{-\frac{p}{2}}\Gamma\left(\frac{p}{2}\right)$

More generally, for $a > 0, p > 0$ and $q > 0$

- $\int_0^\infty x^{p-1}e^{-ax^q}\mathrm{d}x = \frac{1}{q}a^{-\frac{p}{q}}\Gamma\left(\frac{p}{q}\right)$
- $\int_0^\infty x^{-(p+1)}e^{-ax^{-q}}\mathrm{d}x = \frac{1}{q}a^{-\frac{p}{q}}\Gamma\left(\frac{p}{q}\right)$

Properties of Bessel:

(1) $\mathrm{K}_r(u) = \mathrm{K}_{-r}(u)$

(2) $\mathrm{K}_{r+1}(u) = \frac{r}{u}\mathrm{K}_r(u) + \mathrm{K}_{r-1}(u)$

(3) $\mathrm{K}_{1/2}(u) = \mathrm{K}_{-1/2}(u) = \sqrt{\frac{\pi}{2u}}\exp(-u)$

(4) $u \to 0$

$$\begin{cases} \mathrm{K}_r(u) \sim \frac{1}{2}\Gamma(r)\left(\frac{u}{2}\right)^{-r} & r > 0 \\ \mathrm{K}_0(u) = \ln(u) \end{cases}$$

By using these properties, we can get three special cases for GIG: Gamma distribution, Inverse Gamma distribution, and Inverse Gaussian distribution.

Gamma Distribution

For **Gamma Distribution**, $\beta = 0$ and $r > 0, \alpha > 0$, which can also be written as $x \sim Ga(r, \frac{\alpha}{2})$

$$f_X(x) = \frac{\alpha^r}{2^r\Gamma(r)}x^{r-1}\exp\left(-\frac{\alpha x}{2}\right)$$

If $r = 1$, then the **Gamma Distribution** degenerates to the **exponential distribution**.

If $x_i \sim Ga(r_i, \alpha/2)$, then $\sum_{i=1}^n x_i \sim Ga\left(\sum_{i=1}^n r_i, \alpha/2\right)$.

Inverse Gamma

For $IG(\tau, \beta/2), \alpha = 0, r < 0, \beta > 0$

$$f_X(x) = \frac{\beta^\tau}{2^\tau\Gamma(\tau)}x^{-(\tau+1)}\exp\left(-\frac{\beta}{x}\right), \quad \text{where } \tau = -r$$

Inverse Gaussian

Taking $r = -\frac{1}{2}$,

$$f_X(x) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} \exp\left(\sqrt{2\rho}\right) x^{-\frac{3}{2}} \exp\left(-\frac{\alpha x + \beta x^{-1}}{2}\right)$$

If $B(t)$ is a Brown process, $C(t) = B(t) + \alpha t, \quad r \in \mathbb{R}$. We have

$$C(t), t \geq 0$$

is a Gaussian process. Then the inverse GP $T(t)$ is defined as

$$T(t) = \inf\{s > 0, C(s) = \sigma t\}$$

χ^2 Distribution

χ^2 with p degree of freedom. If random variables $x \sim \chi^2_p$, we have

$$f_X(x) = \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2}} x^{p/2-1} \exp\left(-\frac{x}{2}\right) \quad x > 0$$

For a sequence of random variables that normally distributed $(z_1, \ldots, z_p \overset{iid}{\sim} N(0,1))$, the sum of squares of these random variables is χ^2 distributed, which means

$$\sum_{i=1}^{p} z_i^2 \sim \chi^2_p.$$

If we have a vector that each dimension is standard normally distributed, then the square of the length of it is Chi-square distributed.

Beta Distribution

$\alpha > 0, \beta > 0$. If $X \sim Beta(\alpha, \beta)$, then the pdf of X is

$$f_X(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1$$

The front part of $F_X(x)$ which is irrelevant to x is the reciprocal of Beta function $Be(\alpha, \beta)$. The definition of $Be(p, q)$ is

$$Be(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

Student-t Distribution

If X is t distributed denoted by $X \sim t_v$, the pdf of X is

$$f_X(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \cdot \frac{1}{\left(1+\frac{x-\mu}{v\sigma^2}\right)^{\frac{v+1}{2}}} \cdot \frac{1}{\sqrt{v\pi}/\sigma}$$

Theorem 5. [2]
If $v = 1 \Rightarrow$ Cauchy distribution.
If $v \to \infty \Rightarrow$ Gaussian distribution $N(v, \sigma)$.

The t distribution can be also regarded as the scale mixture of normal distribution which can be denoted as $S_t(\mu, \sigma, v)$. There is

$$\int_0^\infty N(x|\,\mu, (\lambda\tau)^{-1}) \underbrace{Ga\left(\tau\middle|\frac{v}{2}, \frac{v}{2}\right)}_{weight} \mathrm{d}\tau$$

$$= S_t\left(x|\mu, \lambda^{-1}, v\right)$$

Ex 2. Suppose $x \sim Bernoulli(\theta)$, $0 < \theta < 1$. We give θ a prior distribution $Beta(\alpha, \beta)$. Then

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta|\alpha, \beta) \\ &\propto C \cdot \theta^x (1-\theta)^{1-x} \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &\propto C \cdot \theta^{x+\alpha-1} (1-\theta)^{\beta-x} \end{aligned}$$

where $C = \frac{\Gamma(\alpha+\beta+1)}{\Gamma(x+\alpha)\Gamma(\beta-x+1)}$. We can easily find that the posterior distribution is **conjugate** with prior distribution. After that, we can use MAP to find the mode of θ.

Ex 3. Suppose $X \sim N(0, \lambda) = \frac{1}{\sqrt{2\pi}} \lambda^{-1/2} \exp\left(-\frac{x^2}{2\lambda}\right)$, $\lambda > 0$. Let λ has a prior distribution $Ga\left(x|r, \frac{\alpha}{2}\right)$. Then we can compute $p(\lambda|x)$

$$\begin{aligned} p(\lambda|x) &= C \cdot \frac{1}{\sqrt{2\pi}} \lambda^{-1/2} \exp\left(-\frac{x^2}{2\lambda}\right) \lambda^{r-1} \exp(-\frac{\alpha\lambda}{2}) \\ &\propto \lambda^{r-\frac{1}{2}-1} \exp\left(-\frac{1}{2}\left(\frac{x^2}{\lambda} + \alpha\lambda\right)\right) \end{aligned}$$

The posterior is a GIG. Therefore, the posterior and prior are generally conjugated. Let us see another kind of prior. What if $\lambda \sim Ig(\tau, \frac{\beta}{2})$ (Inverse Gamma)?

$$p(\lambda|x) = \frac{1}{\sqrt{2\pi}} \lambda^{-1/2} \exp\left(-\frac{x^2}{2\lambda}\right) \lambda^{-(\tau+1)} \exp\left(-\frac{\beta}{2\lambda}\right) \Rightarrow \text{Inverse Gamma}$$

[2] The proof is included as a homework in chapter 3

Scale Mixture Distribution

Scale Mixture of Normals

$$
\begin{aligned}
&\int_0^\infty N\left(x|\mu, \frac{\sigma^2}{r}\right) Ga\left(r|\frac{v}{2}, \frac{v}{2}\right) \mathrm{d}r \\
&= \int_0^\infty \frac{r^{1/2}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x-\mu}{2\sigma^2} r\right) \frac{\left(\frac{v}{2}\right)^{v/2}}{\Gamma\left(\frac{v}{2}\right)} r^{v/2-1} \exp\left(-\frac{rv}{2}\right) \mathrm{d}r \\
&\propto \int_0^\infty r^{\frac{v+1}{2}-1} \exp\left(-\frac{r}{2}\left(\frac{(x-\mu)^2}{\sigma^2} + v\right)\right) \mathrm{d}r \\
&\Rightarrow \text{Student-t Distribution}
\end{aligned}
$$

Laplace Distribution

The Laplace distribution can also be written as a mixture of Gaussian distribution and Exponential Distribution.

$$
\begin{aligned}
f(x) &= \frac{1}{2\sigma} \exp\left(-\frac{|x-\mu|}{\sigma}\right) \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi r}} \exp\left(-\frac{1}{2r}(x-\mu)^2\right) \frac{1}{2\sigma^2} \exp\left|-\frac{r}{2\sigma^2}\right| \mathrm{d}r \\
&= \frac{1}{\sqrt{2\pi}} \int_0^\infty r^{-1/2} \exp\left(-\frac{1}{2}\left(\frac{(x-\mu)^2}{r} + \frac{r}{\sigma^2}\right)\right) \mathrm{d}r \\
&= \frac{2K_{\frac{1}{2}}\left(\frac{|x-\mu|}{\sigma}\right)}{\left(\frac{1}{|\sigma(x-\mu)|}\right)^{1/2}}
\end{aligned}
$$

where $K_{\frac{1}{2}}(\mu) = \sqrt{\frac{\mu}{2\pi}} \exp(-\mu)$

Negative Binomial Distribution

The negative binomial distribution can be regarded as a gamma poisson mixture.

$$
\begin{aligned}
f_X(x) &= \int_0^\infty f_{Po(\lambda)}(k) f_{Ga}\left(r, \frac{1-p}{p}\right) \mathrm{d}\lambda \\
&= \int_0^\infty \frac{\lambda^k}{k!} e^{-\lambda} \frac{\lambda^{r-1} \exp\left(-\frac{1-p}{p}\lambda\right)}{\Gamma(r)\left(\frac{p}{1-p}\right)^r} \\
&= \frac{p^{-r}(1-p)^r}{k!\Gamma(r)} \int_0^\infty \lambda^{r+k-1} \exp\left(-\frac{\lambda}{p}\right) \mathrm{d}\lambda \\
&\propto C\Gamma(r+k)p^{r+k} \\
&= \frac{\Gamma(r+k)}{k!\Gamma(r)} p^k (1-p)^r
\end{aligned}
$$

3.6 Homework

(1) Given the approximation of t distribution when $v = 1$ and $v \to \infty$.

(2) Compute the Gamma-Poisson Mixture: $\sum_{k=0}^{\infty} Ga(x|k, \beta)Po(k|\lambda)$

Chapter 4

Statistic Inference (I)

4.1 Jeffrey Prior

Now we have $p(x|\theta)$, how we set the prior of parameter θ? One method is set it depends on the model.

Definition 11 (Fisher Information)**.**

$$\begin{aligned} I(\theta) &= \mathbb{E}\left[\left(\frac{\mathrm{d}\log(f(x,\theta))}{\mathrm{d}\theta}\right)^2\right] \\ &= \int\left(\frac{\mathrm{d}\log f(x,\theta)}{\mathrm{d}\theta}\right)^2 f(x,\theta)\mathrm{d}x \end{aligned}$$

Lemma 5. *Under certain condition, which means the integral and differential is changeable,*

$$I(\theta) = -\mathbb{E}\left[\frac{\mathrm{d}^2\log f(x,\theta)}{\mathrm{d}\theta^2}\right]$$

Proof.

$$\begin{aligned} \frac{\mathrm{d}^2\log f}{\mathrm{d}\theta^2} &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\frac{\mathrm{d}f}{\mathrm{d}e}}{f}\right) \\ &= \frac{\frac{\mathrm{d}^2 f}{\mathrm{d}\theta^2}}{f} - \left(\frac{\frac{\mathrm{d}f}{\mathrm{d}\theta}}{f}\right)^2 \\ &= \frac{\frac{\mathrm{d}^2 f}{\mathrm{d}\theta^2}}{f} - \left(\frac{\mathrm{d}\log f}{\mathrm{d}\theta}\right)^2 \end{aligned}$$

Take integral for both sides:

$$\begin{aligned}\int \frac{\mathrm{d}^2 \log f}{\mathrm{d}\theta^2} f \mathrm{d}x &= \int \frac{\mathrm{d}^2 f}{\mathrm{d}\theta^2} \mathrm{d}x - I(\theta) \\ &= \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \int f \mathrm{d}x - I(\theta) \\ &= \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \cdot 1 - I(\theta) \\ &= -I(\theta)\end{aligned}$$

□

Now we have the Fisher Information, we define the prior of parameter as

$$p(\theta) \propto \sqrt{I(\theta)}$$

which is called **Jeffrey Prior**. Why we can set prior like this? Suppose we have a one-to-one map $\varphi(\theta)$ of θ, then

$$\begin{aligned}p(\varphi) &= p(\theta) \left| \frac{\mathrm{d}\theta}{\mathrm{d}\varphi} \right| && \text{(Jaccobin)} \\ &\propto \sqrt{I(\theta) \left(\frac{\mathrm{d}\theta}{\mathrm{d}\varphi} \right)^2} \\ &= \sqrt{\mathbb{E}\left[\left(\frac{\mathrm{d} \log f}{\mathrm{d}\theta} \right)^2 \right] \left(\frac{\mathrm{d}\theta}{\mathrm{d}\varphi} \right)^2} \\ &= \sqrt{\mathbb{E}\left[\left(\frac{\mathrm{d} \log f}{\mathrm{d}\theta} \cdot \frac{\mathrm{d}\theta}{\mathrm{d}\varphi} \right)^2 \right]} \\ &= \sqrt{\mathbb{E}\left[\left(\frac{\mathrm{d} \log f}{\mathrm{d}\varphi} \right)^2 \right]} \\ &= \sqrt{I(\varphi)}\end{aligned}$$

We call this **prior invariance**.

Ex 1. Suppose $x \sim N(\mu, \sigma^2)$ with σ fixed, compute the $I(\mu)$.

$$\log f \propto -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2$$

$$\sqrt{I(\mu)} = \sqrt{\mathbb{E}\left[\left(\frac{x-\mu}{\sigma} \right)^2 \right]} = \sqrt{\frac{\mathbb{E}(x-\mu)^2}{\sigma^4}} = \sqrt{\frac{1}{\sigma^2}} \propto 1$$

This prior propers to a constant and the integral of that does not equal to 1. Therefore, we define it as **improper prior**.

Ex 2. Suppose $x \sim N(\mu, \sigma^2)$ with μ fixed, compute the $I(\sigma)$.

Let $\tau = \frac{1}{\sigma^2}$

$$f(\tau) = \frac{\tau^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{\tau(x-\mu)^2}{2}\right)$$

Compute $\log f$

$$\log f = \frac{1}{2}\ln\tau - \frac{\tau}{2}(x-\mu)^2$$

Derivative of τ

$$\frac{\mathrm{d}\log f}{\mathrm{d}\tau} = \frac{1}{2\tau} - \frac{1}{2}(x-\mu)^2$$

Compute expect of the derivative:

$$\mathbb{E}\left[\frac{1}{4}\left(\frac{1}{\tau} - (x-\mu)^2\right)^2\right] = \mathbb{E}\left[\frac{1}{4\tau^2} - \frac{(x-\mu)^2}{2\tau} + \frac{(x-\mu)^4}{4}\right] = \frac{1}{4\tau^2} - \frac{1}{2\tau^2} + \mathbb{E}\left[\frac{(x-\mu)^4}{4}\right]$$

where

$$\mathbb{E}\left[\frac{(x-\mu)^4}{4}\right] = \frac{1}{4}\int (x-\mu)^4 N(x|\mu, \tau^{-1/2})\mathrm{d}x = \frac{3}{4\tau^2}$$

In terms of the integral above, we have

$$\begin{aligned}
&\int (x-\mu)^2 \frac{\tau^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{\tau(x-\mu)^2}{2}\right)\mathrm{d}x = \frac{1}{\tau} \\
\Rightarrow &\int (x-\mu)^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\tau(x-\mu)^2}{2}\right)\mathrm{d}x = \tau^{-\frac{3}{2}} \\
\Rightarrow &-\int \frac{(x-\mu)^4}{2} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\tau(x-\mu)^2}{2}\right)\mathrm{d}x = -\frac{3}{2}\tau^{-\frac{5}{2}} \\
\Rightarrow &\int (x-\mu)^4 \frac{\tau^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{\tau(x-\mu)^2}{2}\right)\mathrm{d}x = 3\tau^{-2}
\end{aligned}$$

Therefore,

$$\frac{\mathrm{d}\log f}{\mathrm{d}\tau} = \frac{1}{4\tau^2} - \frac{1}{2\tau^2} + \frac{3}{4\tau^2} = \frac{1}{2\tau^2} \propto \frac{1}{\tau^2}$$

Ex 3. For Poisson distribution with parameter λ,

$$\begin{aligned}
&f(n|\lambda) = e^{-\lambda}\frac{\lambda^n}{n!} \\
&\mathbb{E}(I(\lambda)) = \mathbb{E}\left[(\frac{n}{\lambda} - 1)^2\right] = \mathbb{E}\left[1 + \frac{n^2}{\lambda^2} - \frac{2n}{\lambda}\right] = -1 + \mathbb{E}\left[\frac{n^2}{\lambda^2}\right]
\end{aligned}$$

For $\mathbb{E}\left[\frac{n^2}{\lambda^2}\right]$, we use the property

$$\sum_{n=0}^{\infty} n\frac{e^{-\lambda}\lambda^n}{n!} = \lambda$$

Therefore,

$$\mathbb{E}\left[\frac{n^2}{\lambda^2}\right] = 1 + \frac{1}{\lambda}$$
$$\Rightarrow \mathbb{E}\left[I(\lambda)\right] \propto \sqrt{\frac{1}{\lambda}}$$

Ex 4. Now we have the model $x = \theta + \varepsilon$ and $\varepsilon \sim N(0, \tau^{1/2})$ How to estimate the parameter θ?

Case 1 τ fixed, and let $p(\theta) \sim N(\theta|0, \lambda^1/2)$

$$p(\theta|x) \propto p(x|\theta)p(\theta) = \frac{1}{\sqrt{2\pi\lambda}} exp\left(-\frac{\theta^2}{2\lambda}\right) \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{(x-\theta)^2}{2\tau}\right)$$

For the exponential part,

$$\exp\left(-\frac{\theta^2}{2\lambda}\right) \cdot \exp\left(-\frac{(x-\theta)^2}{2\tau}\right)$$
$$= \exp\left(\left(\frac{1}{\tau\lambda} + \frac{1}{\tau}\right)\theta^2 - \frac{2\theta x}{\tau}\right)$$

Therefore, the prior and the posterior are conjugated.

$$\Rightarrow p(\theta|x) \propto N\left(\theta \middle| \frac{\lambda x}{\lambda+\tau}, \left(\frac{\lambda\tau}{\lambda+\tau}\right)^{1/2}\right)$$

Case 2 Suppose θ, τ are all parameters. Method 1 is to let $\theta \perp\!\!\!\perp \tau \Rightarrow p(\theta, \tau) = p(\theta)p(\tau)$

$$p(\theta, \tau|x) = p(x|\theta, \tau) \cdot p(\theta, \tau)$$
$$= p(x|\theta, \tau)p(\theta)p(\tau)$$

Provide two different priors:

$$p(\theta) \sim N(0, \lambda^{1/2}) \quad \Gamma(\tau) \sim Ga(0, \frac{\beta}{2})$$

Compute the posterior.

$$p(\theta, \tau|x) = \frac{1}{\sqrt{2\pi}} \tau^{1/2} \exp\left(-\frac{\tau(x-\theta)^2}{2}\right) \frac{\lambda^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{\lambda\theta^2}{2}\right) \left(\frac{\beta}{2}\right)^{\alpha} \exp\left(-\frac{\beta\tau}{2}\right) \tau^{\alpha-1}/\Gamma(\alpha)$$

Let α replace $\alpha/2$

$$\mathscr{L} = \tau^{\frac{\alpha+1}{2}-1} \exp\left(-\frac{\tau}{2}\left((x-\theta)^2 + \beta\right)\right) \exp\left(-\frac{\lambda\theta^2}{2}\right) \lambda^{1/2}$$
$$\ln \mathscr{L} = \left(\frac{\alpha+1}{2} - 1\right) \ln\tau - \frac{\tau}{2}\left((x-\theta)^2 + \beta\right) + \frac{1}{2}\ln\lambda - \frac{\lambda\theta^2}{2}$$

Let $Q(\theta, \tau) = -2\ln\mathscr{L} = -(\alpha-1)\ln\tau - \tau((x-\theta)^2 + \beta) - \ln\lambda + \lambda\theta^2$

$$\min Q \Rightarrow \begin{cases} \frac{\partial Q}{\partial \theta} = -2\tau(x-\theta) + 2\lambda\theta = 0 \\ \frac{\partial Q}{\partial \tau} = \frac{1-\alpha}{\tau} + (x-\theta)^2 + \beta = 0 \end{cases} \tag{4.1}$$

Check the result, the prior and posterior in non-conjugate under the hypothesis. So, this is not a good solution.

The second way is to suppose θ is depend on τ where $p(\theta|\tau) \sim N\left(0, (\lambda\tau)^{1/2}\right)$ and $\tau \sim Ga(\alpha, \beta)$. In other words, $p(\theta, \tau) = p(\theta|\tau)p(\tau)$ Again, compute the posterior.

$$\begin{aligned} \mathscr{L} &= \tau^{\frac{\alpha+1}{2}-1} \exp\left(-\frac{\tau}{2}(x-\theta)^2 + \beta\right) \exp\left(-\frac{\lambda\tau\theta^2}{2}\right) (\tau\lambda)^{1/2} \\ &= \tau^{\alpha/2} \exp\left[-\frac{\tau}{2}\left((x-\tau)^2 + \beta + \lambda\theta^2\right)\right] \\ \ln \mathscr{L} &= \frac{\alpha}{2}\ln\tau - \frac{\tau}{2}\left((x-\theta)^2 + \beta + \lambda\theta^2\right) = Q \end{aligned}$$

The optimal condition:

$$\begin{cases} \frac{\partial Q}{\partial \theta} = \tau\left(-(x-\theta) + \lambda\theta\right) = 0 \\ \frac{\partial Q}{\partial \tau} = \frac{\alpha}{2}\frac{1}{\tau} - \frac{1}{2}\left((x-\theta)^2 + \beta + \lambda\theta^2\right) = 0 \end{cases} \tag{4.2}$$

This time we can a conjugate result which means when using MAP algorithm we can guarantee the convergence.

4.2 Moment Generation Function

Definition 12. The **moment-generating function** of a random variable X is

$$\psi_X(t) = \mathbb{E}\left(e^{tx}\right) = \int e^{tx} \mathrm{d}F_X(x)$$

According to the definition of moment generation function, we have the following attributes:

- $\psi'_X(0) = \int x \mathrm{d}F_X(x)$
- The exchangeabilty of deriviative and integral $\psi_X^{(}k)(0) = \int \mathrm{d}F_X(x) = \mathbb{E}(x^k)$
- Laplace Transformation: $\mathscr{L}(t) = \mathbb{E}(e^{-tx}) = \int e^{-tx}\mathrm{d}F_X(x)$. Considering any measure μ:

$$\mathscr{L}(\mu, t) = \int \exp(-tx) \cdot \mu(\mathrm{d}x)$$

Definition 13 (completely monotone)**.** A function $g : (0, \infty) \mapsto \mathbb{R}$ is **completely monotone** function if the f is of class C^∞ which means ∞ deriviative and $(-1)^n g^{(n)} \geq 0$ for all $n \in N \cup \{0\}$ and $\lambda > 0$.

Theorem 6 (Bernstein Theorem)**.** *Let $g : (0, \infty) \mapsto \mathbb{R}$ be a c.m. function. Then it is the Laplace Transform of a unique measure M on $[0, \infty)$, i.e, for all $\lambda > 0$*

$$g(\lambda) = \int_0^\infty \exp(-\lambda t) \cdot (\mathrm{d}t) = \mathscr{L}(\mu, t)$$

Inversely, whenever $\mathscr{L}(\mu, \lambda) < \infty$ for every $\lambda > 0$, $\lambda \mapsto \mathscr{L}(\mu, \lambda)$ is a c.m. function.

Proof. First of all, we have a corollary.

$$g(0+) = 1 \quad g(+\infty) = 0$$

We can also regard $\mu(\mathrm{d}t) = F(\mathrm{d}t)$ as a probability measure. Then the original statement equals to:

$$g(\lambda) = \int \exp(-\lambda t) \cdot F(\mathrm{d}t)$$

According to Taylor expansion: for any $a > 0$ and $n \in N$

$$\begin{aligned} g(\lambda) &= \sum_{k=0}^{n-1} \frac{g^{(k)}a}{k!}(\lambda - a)^k + \int_a^\lambda \frac{g^{(n)}(s)}{(n-1)!}(\lambda - s)^{n-1}\mathrm{d}s \in (a, \lambda) \\ &= \underbrace{\sum_{k=0}^{n-1} \frac{(-1)^k g^{(k)}a}{k!}(a - \lambda)^k}_{\alpha} + \underbrace{\int_\lambda^a \frac{(-1)^n g^{(n)}(s)}{(n-1)!}(s - \lambda)^{n-1}\mathrm{d}s}_{\beta} \end{aligned}$$

For $a > \lambda$: $(\alpha) \geq 0$

$$\begin{aligned} &\lim_{a\to\infty} \int_\lambda^a \frac{(-1)^n g^{(n)}(s)}{(n-1)!}(s - \lambda)^{n-1}\mathrm{d}s \\ =&\int_\lambda^\infty \frac{(-1)^n g^{(n)}(s)}{(n-1)!}(s - \lambda)^{n-1}\mathrm{d}s \\ \leq&\varphi(\lambda) \end{aligned}$$

Let

$$\rho_k(\lambda) = \lim_{a\to\infty} \frac{(-1)^k g^{(k)}(a)}{k!}(a - \lambda)^k$$

Obviously, the $\rho_k(\lambda)$ is independent to λ, since

$$\rho_k(\nu) = \lim_{a\to\infty} \frac{(-1)^k g(k)a}{k!} \cdot \underbrace{\frac{(a - \nu)^k}{(a - \lambda)^k}}_{=1}(a - \lambda)^k$$

. And

$$g(+\infty) = 0 \Rightarrow \rho_k = 0$$

So,

$$g(\lambda) = \sum_{k=0}^{n-1} \rho_k + \int_\lambda^\infty \frac{(-1)^n g^{(n)}(s)}{(n-1)!}(s-\lambda)^(n-1)\mathrm{d}s$$
$$\Rightarrow g(\lambda) = \int_\lambda^\infty \frac{(-1)^n g^{(n)}(s)}{(n-1)!}(s-\lambda)^{(n-1)}\mathrm{d}s$$

On the other hand, from $g(0+) = 1$, we can move forward.

$$\Rightarrow 1 = \lim_{\lambda\to 0} g(\lambda) = \int_0^\infty \underbrace{\left(\frac{(-1)^n g^{(n)}(s)}{(n-1)!}s^{n-1}\right)}_{(\gamma)}\mathrm{d}s \Rightarrow \gamma \text{ can be regarded as a p.d.f}$$

$$\Leftrightarrow g(\lambda) = \int_0^\infty \left(1-\frac{\lambda}{s}\right)_+^{n-1} \frac{(-1)^n g^{(n)} s}{(n-1)!}\mathrm{d}s, \quad \text{where } (a)_+ := \max(a,0)$$

Let $s = \frac{n}{t}$, $\mathrm{d}s = |s^{-2}n|\mathrm{d}t$. $g(\lambda)$ can be rewritten as

$$\begin{aligned} g(\lambda) &= \int_0^\infty \left(1-\frac{\lambda t}{n}\right)_+^{n-1} \frac{(-1)^n g^{(n)}(\frac{n}{t})}{(n-1)!}\left(\frac{n}{t}\right)^{n-1} t^{-2}n\mathrm{d}t \\ &= \int_0^\infty \left(1-\frac{\lambda t}{n}\right)_+^{n-1} \frac{(-1)^n g^{(n)}(\frac{n}{t})(\frac{n}{t})^{n+1}}{n!}\mathrm{d}t \end{aligned}$$

Therefore,

$$\lim_{n\to\infty} g(\lambda) = \int_0^\infty \exp(-\lambda t) f(t)\mathrm{d}t$$

□

Corollary 2 (Mixture of Bartlett-Fejer Kernels)**.** *Let $g(t)$ be a function that is symmetric about the origin, integrable, convex and twice differentiable on $(0,+\infty)$ and $g(0+) = 1$ and $g(+\infty) = 0$. Then*

$$g(t) = \int_0^\infty \frac{1}{s}\left(1-\frac{t}{s}\right)_+ s g''(s)\mathrm{d}s, \quad t > 0$$

4.3 Homework

(1) Compute the expectation in **Ex 2.**

(2) Compute the integrals:

- $m_0 = \int_{-\infty}^\infty \Phi(x)N(x|\mu,\sigma^2)\mathrm{d}x$
- $m_0 = \int_{-\infty}^\infty \Phi(x)N(x|\mu,\sigma^2)x\mathrm{d}x$
- $m_0 = \int_{-\infty}^\infty \Phi(x)N(x|\mu,\sigma^2)(x-m_0)^2\mathrm{d}x$

where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t$

(3) $f(x, \theta) = \theta^x (1-\theta)^{1-x}$

- Compute $\pi(\theta)$ or $\mathbb{E}[I(\theta)]$
- If $\theta = \sin^2 \alpha$, compute $\pi(\alpha)$

(4) In **Ex 4.** case 2, compute the posterior given the following prior:

- Uniformative $p(\tau) \propto 1$
- $\pi \propto \frac{1}{\tau^2}$. Actually we can get a Gamma posterior.

Chapter 5

Multivariate Distribution

5.1 Bivariate Distribution

Definition 14 (Joint Mass Function)**.** Given a pair of discrete r.v. X and Y. Define the joint mass function by

$$f_{(X,Y)}(x,y) = P(X = x, Y = y)$$

Definition 15 (Probability Density Function)**.** In the continuous case, we call a function $f(x,y)$ a p.d.f for (X,Y) if:

(1) $f(x,y) \geq 0$ for all x, y

(2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{d}x\mathrm{d}y = 1$

(3) For any set $A \subset \mathbb{R} \times \mathbb{R}$, $P((x,y) \in A) = \int\int_A f(x,y)\mathrm{d}x\mathrm{d}y$. The cdf $F_{XY}(x,y) = P(X \leq x, Y \leq y)$

Definition 16 (Marginal Distribution)**.** If (X,Y), $f(x,y)$, $f_X(x) = P(X = x) = \sum\limits_y P(X = x, Y = y) = \sum\limits_y f(x,y)$. For continuous distribution, $f_X(x) = P(X = x) = \int_y f(x,y)\mathrm{d}y$

Definition 17 (Indepedent R.V.)**.** X, Y are independent if for every A and B

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$$

. We denote that $X \perp\!\!\!\perp Y$.

Definition 18 (Conditional Distribution)**.** $f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X=x,Y=y)}{P(Y=y)} = \frac{F_{XY}(x,y)}{F_Y(y)}$ $f_Y(y) > 0$

5.2 Multinomial Distribution

Let $X = (X_1, ..., X_n)$ where X_i are r.v. We call X a random vector and $f(x_1, ..., x_n)$ as p.d.f (p.m.f). Similarly, we can also deduce the marginal distribution and independence:

- Marginal distribution: $f(x_i) = \sum_{x_1,...,x_{i-1},x_{i+1},...,x_n} f(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$
- Independence: $f(x_1, ..., x_n) = \prod_{i=1}^{n} f_{x_i}(x_i)$

Definition 19 (iid)**.** If $X_1, ..., X_n$ are independent and each has the same marginal distribution with c.d.f, we say that $X_1, ..., X_n$ are **i.i.d** or **independent and identically distributed**, denoted as $X_i \sim F(\theta)$.

Definition 20 (exchangeable)**.** Let $f(x_1, ..., x_n)$ be the joint density of $X_1, ..., X_n$. If $f(x_1, ..., x_n) = f(x_{\pi_1}, ..., x_{\pi_n})$ and $\{x_{\pi_1}, ..., x_{\pi_n}\}$ is permutation of $\{1, ..., n\}$, then $X_1, ..., X_n$ are **exchangeable**.

Ex 1. Let $P(X_1 = x_1, ..., X_{10} = x_{10}|\theta) = \prod_{i=1}^{10} \theta^{x_i}(1-\theta)^{1-x_i}$. If $\theta \sim f(\theta)$,

$$\begin{aligned} f(x_1, ..., x_{10}) &= \int f(x_1, ..., x_{10}|\theta) f(\theta) \mathrm{d}\theta \\ &= \int_0^1 \theta^{\Sigma x_i}(1-\theta)^{1-\Sigma x_i} f(\theta) \mathrm{d}\theta \end{aligned}$$

This indicate that if $f(\mathbf{X})$ is exchangeable, then $\{X_i\}$ might be exchangeable. The following theorem provide a principle to judge that.

Theorem 7 (De Finetti)**.** *Let $X_i \subset X$ for all $i \in \{1, 2, ...\}$. Suppose that for any n, $X_1, ..., X_n$ are exchangeable:*

$$f(x_1, ..., x_n) = f(x_{\pi_1}, ..., x_{\pi_n})$$

for all parameters π_i of $\{1, ..., n\}$. Then we have:

$$f(x_1, ..., x_n) = \int \left[\sum_{i=1}^{n} f(x_i|\theta) \right] p(\theta) \mathrm{d}\theta$$

$p(\theta)$ are parameter θ, some prior distribution $p(\theta)$ on θ and some sample model $p(x|\theta)$.

Inversely, we have another conclusion.

Theorem 8. *If $\theta \sim P(\theta)$ and $X_1, ..., X_n$ are conditionally i.i.d given θ, then marginally $X_1, ..., X_n$ are exchangeable (LDA).*

Proof.

$$\begin{aligned} f(x_1,..,x_n) &= \int f(x_1,...,x_n|\theta)P(\theta)\mathrm{d}\theta \\ &= \int \prod_{i=1}^{n} f(x_i|\theta)P(\theta)\mathrm{d}\theta = f(x_{\pi_1},...,x_{\pi_n}) \end{aligned}$$

□

5.3 Transformation

One to One Map

Theorem 9 (Law of transformation)**.** *Let $X \sim$ pdf f_X/cdf F_X and $Y = g(X)$ be a function of X. In the discrete case, the pmf of Y*

$$f_Y(y) = P(Y \le y) = P(g(X) \le y) = P(x \in g^{-1}(y))$$

In continuous case, we have a similar conclusion.

(1) For each y, find set $A_y = \{x : g(x) \le y\}$

(2) Find cdf

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(g(x) \le y) \\ &= \int_{A_y} f_X(x)\mathrm{d}x \end{aligned}$$

(3) $f_Y(y) = F_Y'(y)$

Ex 2. Suppose $P(X = -1) = P(X = 1) = \frac{1}{4}$ and $P(X = 0) = \frac{1}{2}$. Let $Y = X^2$, we have

$$P(Y = 0) = P(X = 0) = \frac{1}{2}$$

$$P(Y = 1) = P(X = 1) + P(X = -1) = \frac{1}{2}$$

Ex 3. Let $f_X(x) = e^{-x}$ for $x > 0$, $Y = g(X) = \log X$.

$$\begin{aligned} &F_X(x) = \int_0^{\infty} f_X(u)\mathrm{d}u = 1 - e^{-x} \\ &A_y = \{x : x \le e^y\} \\ &F_Y(y) = P(Y \le y) = P(\log x \le y) = P(X \le e^y) = F_X(e^y) = 1 - e^{-e^y} \\ &f_Y(y) = e^y \cdot e^{-e^y} \end{aligned}$$

Ex 4. Let $X \sim U(-1,3)$ and $Y + X^2$.

$$\begin{cases} \frac{1}{4} & x \in (-1,3) \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

Therefore, Y can only take values in $[0,9)$. Consider (1) $0 \le y \le 1$; (2) $1 \le y < 9$. For case (1), $A_y = \left[-\sqrt{y}, \sqrt{y}\right]$

$$F_y = \int_{A_y} F_X(x)\mathrm{d}x = \int_{-\sqrt{y}}^{\sqrt{y}} F_X(x)\mathrm{d}x = \frac{1}{2}\sqrt{y}$$

For case (2), $A_y = \left[-1, \sqrt{y}\right]$

$$F_y = \int_{A_y} \frac{1}{4}\mathrm{d}x = \frac{1}{4}(1+\sqrt{y})$$

In conclusion,

$$\begin{cases} \frac{1}{4\sqrt{y}} & 0 < y < 1 \\ \frac{1}{8\sqrt{y}} & 1 \le y < 9 \\ 0 & \text{otherwise} \end{cases}$$

Multivariate Mapping

Theorem 10 (Law of Transformation)**.** *Given a transformation $Z = g(X,Y)$, the pdf of Z can be computed by the following method.*

Step (1) For each z, find $A_z = \{(x,y) : g(x,y) \le z\}$

Step (2) Find CDF $F_Z(z) = P(Z \le z) = \int\int_{Z_z} f_{XY}(x,y)\mathrm{d}x\mathrm{d}y$

Step (3) $f_Z(z) = F_Z'(z)$

Ex 5. Let $X_1, X_2 \overset{iid}{\sim} U(0,1)$ and $Y = X_1 + X_2$. The joint pdf is given as

$$f_{X_1,X_2}(x_1,x_2) = \begin{cases} 1 & 0 < x_1 < 1, 0 < x_2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

The transformation g is defined as the sum of X_1 and X_2 which is $g_{X_1X_2}(x_1,x_2) = x_1 + x_2$. Then we can compute the CDF and PDF of g.

$$\begin{aligned} F_Y(y) &= P(\{x_1,x_2\} : x_1 + x_2 \le y) \\ &= \int\int_{A_Y} f(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2 \end{aligned}$$

$$= \begin{cases} 0 & y < 0 \\ \frac{1}{2}y^2 & 0 < y < 1 \\ 1 - \frac{(2-y)^2}{2} & 1 \le y < 2 \\ 1 & y > 2 \end{cases}$$

$$f_Y(y) = \begin{cases} y & 0 \le y < 1 \\ 2 - y & 1 \le y < 2 \\ 0 & \text{otherwise} \end{cases}$$

Theorem 11. *Let X have a CDF $F_X(x)$ and $Y = g(x)$ and let $\mathscr{X} = \{x : f_X(x) > 0\}$ and $\mathscr{Y} = \{y : y = g(x) \text{ for some } x \in \mathscr{X}\}$.*

(1) If g is a strictly increasing function on $\mathscr{X}$,

$$F_Y(y) = F_X(g^{-1}(y)) \text{ for } y \in \mathscr{Y}$$

(2) If g is a strictly decreasing function on $\mathscr{X}$ and X is a continuous r.v.

$$F_Y(y) = 1 - F_X(g^{-1}(y)) = \int_{A_Y} \mathrm{d}(F_X(x)) \text{ for } y \in \mathscr{Y}$$

Proof. If decreasing: $\{x \in \mathscr{X}, g(x) \le y\} = \{x \in \mathscr{X}, g^{-1}(g(x)) \ge g^{-1}(y)\}$

$$\begin{aligned} F_Y(y) &= \{x \in \mathscr{X}, x \ge g^{-1}(y)\} \\ &= \int_{\{x \in \mathscr{X}, x \le g^{-1}(y)\}} f_X(x)\mathrm{d}x \\ &= \int_{g^{-1}(y)}^{\infty} f_X(x)\mathrm{d}x = 1 - F_X(g^{-1}(y)) \end{aligned}$$

□

Theorem 12. *Let $X \sim f_X(x)$ and $Y = g(X)$. The continuous g is a strictly monotonic function. $\mathscr{X} = \{x : f_X(x) > 0\}$ and $\mathscr{Y} = \{y : y = g(x) \text{ for some } x \in \mathscr{X}\}$. Then*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y) & \text{increasing} \\ -f_X(g^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y) & \text{decreasing} \end{cases}$$

Now, we have another problem. Suppose we have a continuous distribution F_X. How to sample x from F_X. We have $\nu \sim U(0,1)$.

Theorem 13 (Probability Integral Transform)**.** *Let X has continuous cdf $F_X(x)$ and $Y = F_X(x)$. Then Y is uniformly distributed on $U(0,1)$, which is $P(Y \le y) = y$ $0 < y < 1$.*

Proof.

$$\begin{aligned} P(Y \le y) &= P(F_X(x0 \le y)) \\ &= P(F_X^{-1}(F(x)) \le F^{-1}(y)) \\ &= P(X \le F^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) = y \end{aligned}$$

□

Jaccobian

Definition 21 (Jaccobian)**.** Let X be an $m \times 1$ random vector having a density function $f(x)$, which is positive on a set $\mathscr{X} \subset \mathbb{R}^m$. Suppose the transformation $X = \mathbf{Y}(X) = (Y_1(X), ..., Y_n(X))^T$ is $1-1$ of some y, where $\mathscr{Y}$ denies the image of X under y, so that the inverse transformation $X = \mathbf{X}(Y)$ exists for $Y \in \mathscr{X}$. Assuming that the partial derivative $\partial x_i/\partial y_j$ $(i, j = 1, 2, ..., m)$ and continuous on $\mathscr{Y}$. It is well known that the density function of random vector $Y = \mathbf{Y}(X)$ is

$$f_Y(y) = f_X(X(y))|J(x \to y)| \quad y \in \mathscr{Y}$$

where $J(x \to y)$ is the **Jaccobian** of transformation.

$$J(x \to y) = \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}$$

Definition 22 (Exterior Product)**.**

$$\mathrm{d}x_i \cdot \mathrm{d}x_j = -\mathrm{d}x_j \cdot \mathrm{d}x_i$$

Definition 23 (Wedge Product)**.**

$$\mathrm{d}x_i \wedge \mathrm{d}x_j = -\mathrm{d}x_j \wedge \mathrm{d}x_i$$

Theorem 14. *If* $\mathrm{d}y = (\mathrm{d}y_1, ..., \mathrm{d}y_m)^T$ *is an* $m \times 1$ *vector of differentials and if* $\mathrm{d}x = (\mathrm{d}x_1, ..., \mathrm{d}x_m)^T = B \cdot \mathrm{d}y$*, where* B *is an* $m \times m$ *nonsingular matrix, then* $\bigwedge_{i=1}^{m} \mathrm{d}x_i = \det(B) \cdot \bigwedge_{i=1}^{m} \mathrm{d}y_i$

Proof.

$$\begin{pmatrix} \mathrm{d}x_1 \\ \mathrm{d}x_2 \end{pmatrix} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \cdot \begin{pmatrix} \mathrm{d}y_1 \\ \mathrm{d}y_2 \end{pmatrix} = \begin{pmatrix} B_{11}\mathrm{d}y_1 + B_{12}\mathrm{d}y_2 \\ B_{21}\mathrm{d}y_1 + B_{22}\mathrm{d}y_2 \end{pmatrix}$$

$$\begin{aligned} dx_1 \wedge \mathrm{d}x_2 &= (B_{11}\mathrm{d}y_1 + B_{12}\mathrm{d}y_2)(B_{21}\mathrm{d}y_1 + B_{22}\mathrm{d}y_2) \\ &= (B_{11} \cdot B_{12} - B_{12} \cdot B_{21})\mathrm{d}y_1 \wedge \mathrm{d}y_2 \end{aligned}$$

$$(\dots)\mathrm{d}x = (\dots)B\mathrm{d}y \Leftarrow \begin{pmatrix} B_{11} - \vec{b_1}b_{m\times m}^{-1}\vec{a_1} & 0 \\ \vec{a_1} & b_{m\times m} \end{pmatrix} = \begin{pmatrix} I_{m\times m} & -\vec{b_1}b_{m\times m}^{-1} \\ 0 & 1 \end{pmatrix}\begin{pmatrix} B_{11} & \vec{b_1} \\ \vec{a_1} & b_{m\times m} \end{pmatrix}$$

□

Ex 6. Carton Compute the Jaccobian of projection from rectangular coordinates $x_1, \dots, x_m$ to polar coordinate $r, \theta_1, \dots, \theta_{m-1}$, where

$$\begin{aligned} x_1 &= r\sin\theta_1\cdots\sin\theta_{m-1} \\ x_2 &= r\sin\theta_1\cdots\sin\theta_{m-2}\cos\theta_{m-1} \\ x_3 &= r\sin\theta_1\cdots\cos\theta_{m-2} \\ &\vdots \\ x_{m-1} &= r\sin\theta_1\cos\theta_2 \\ x_m &= r\cos\theta_1 \quad (r>0, 0<\theta_i\le\pi(i=1,\cdots,m-2), 0<\theta_{m-1}\le 2\pi) \end{aligned}$$

$$\Rightarrow J(\vec{x}\to r\theta_1\cdots\theta_{m-1}) = r^{m-1}\sin^{m-2}\theta_1\sin^{m-3}\theta_2\cdots\sin\theta_{m-1}$$

Proof.

$$\begin{aligned} x_1^2 &= r^2\sin^2\theta_1\cdots\sin^2\theta_{m-1} \\ x_2^2 &= r^2\sin^2\theta_1\cdots\sin^2\theta_{m-2}\cos^2\theta_{m-1} \\ &\vdots \\ x_m^2 &= r^2\cos^2\theta_1 \end{aligned}$$

Sum up these terms, we get

$$\sum_{i=1}^{m} x_i^2 = r^r$$

For derivatives,

$$\begin{aligned} 2x_1\mathrm{d}x_1 &= 2r^2\sin^2\theta_1\cdots\sin^2\theta_{m-2}\sin\theta_{m-1}\cos\theta_m\mathrm{d}\theta_{m-1} \\ &\quad + \text{terms including } \mathrm{d}r, \mathrm{d}\theta_1, \dots \mathrm{d}\theta_{m-1} \\ 2x_1\mathrm{d}x_1 + 2x_2\mathrm{d}x_2 &= 2r^2\sin^2\theta_1\cdots\sin\theta_{m-2}\cos\theta_{m-2}\mathrm{d}\theta_{m-2} \\ &\quad + \text{terms including other } \mathrm{d}(\cdot) \\ &\vdots \\ \dots &\Rightarrow \sum_{i=1}^{m} 2x_i\mathrm{d}x_i = 2r\mathrm{d}r \end{aligned}$$

Take wedge from both sides,

$$2^m\prod_{i=1}^{m} x_i \overset{m}{\underset{i=1}{\wedge}} \mathrm{d}x_i = 2^m r^{2m-1}\sin^{2m-3}\theta_1\sin^{2m-5}\theta_2\dots\overset{m}{\underset{i=1}{\wedge}}\mathrm{d}\theta_i\wedge\mathrm{d}r$$

□

For any matrix $X = (x_{ij})$, we have some basic law of differential.

Theorem 15. *For any matrix $X = (x_{ij}), X \in \mathbb{R}^{n\times m}$,*

1. $\mathrm{d}X = (\mathrm{d}x_{ij})$
2. $\mathrm{d}(XY) = X\mathrm{d}Y + Y\mathrm{d}X$
3. $(\mathrm{d}X) = \wedge_{i=1}^{m} \wedge_{j=1}^{n} \mathrm{d}x_{ij}$
4. *If X is a symmetric $m \times m$ matrix, the symbol*

$$(\mathrm{d}X) = \bigwedge_{1\le i\le j\le m} \mathrm{d}X_{ij}$$

5. *If X is a anti-symmetric matrix, the symbol*

$$(\mathrm{d}X) = \bigwedge_{1\le i\le j\le m} \mathrm{d}X_{ij}$$

6. *If X is upper-triangluar matrix, the symbol*

$$(\mathrm{d}X) = \bigwedge_{1\le i\le j\le m} \mathrm{d}X_{ij}$$

Theorem 16. *X and Y are two $n \times m$ matrixes. Given $X = BYC$, from which $B^{n\times n}$ and $C^{m\times m}$ are non-singular. We have*

$$\begin{aligned}(\mathrm{d}X) &= (\det B)^m (\det C)^n (\mathrm{d}Y)\\ J(X \to Y) &= (\det B)^m (\det C)^n\end{aligned}$$

Proof. [1]

$$\begin{aligned}vec(X) &= vec(BYC)\\ &= (C^T \otimes B) vec(Y)\\ (\mathrm{d}X) &= \det(C^T \otimes B)(\mathrm{d}Y)\\ &= (\det C)^n (\det B)^m (\mathrm{d}Y)\end{aligned}$$

where $\otimes$ is Kronecker product.[2]

By SVD decomposition, we can denote C as Σ and B as Π.

$$\Rightarrow \Sigma^n \Pi^m (\mathrm{d}Y)$$

□

[1] $vec(\cdot)$ means to stretch a matrix to a vector row by row

[2] Kronecker Product: Given two matrics $A^{p\times q}$ and $B^{m\times n}$, the Kronecker product is definded as: $A \otimes B = (a_{ij}B)^{pm,qn}$

Theorem 17. *If $X = BYB^T$, where X and Y are two $m \times m$ symmetric matrixes, and B is a non-singular matrix, then*

$$\begin{aligned}(\mathrm{d}X) &= (\det B)^{m+1}(\mathrm{d}Y)\\ J(X \to Y) &= (\det B)^{m+1}\end{aligned}$$

Proof.

$$(\mathrm{d}X) = (B\mathrm{d}YB^T) = \rho(B)(\mathrm{d}Y)$$

where $\rho(B)$ is a polynomial of elements of B.

For $\rho(\cdot)$, we want to prove $\rho(B_1B_2) = \rho(B_1)\rho(B_2)$

$$\begin{aligned}(\mathrm{d}X) &= (B_1B_2\mathrm{d}YB_2^TB_2)\\ &= \rho(B_1)(B_2\mathrm{d}YB_2^T)\\ &= \rho(B_1)\rho(B_2)\mathrm{d}Y\end{aligned}$$

So, $\rho(B) = (\det B)^k$ for some k. □

5.4 Random Vector

Definition 24 (Random Vector)**.** $X = (x_1, ..., x_m)^T$. The mean of a random vector X can be written as

$$\bar{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} = \begin{pmatrix} \mathbb{E}(x_1) \\ \mathbb{E}(x_2) \\ \vdots \\ \mathbb{E}(x_m) \end{pmatrix}$$

The covariance matrix can be defined as:

$$\mathrm{Cov}(X) = \begin{pmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1, x_2) & \cdots & \mathrm{Cov}(x_1, x_m) \\ \vdots & \mathrm{Var}(x_2) & & \mathrm{Cov}(x_2, x_n) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(x_1, x_n) & \mathrm{Cov}(x_2, x_n) & \vdots & \mathrm{Var}(x_n) \end{pmatrix}$$

where the expectation of x is defined as:

$$\mathbb{E}(x) = \begin{cases} \sum_x xp(X = x) & \text{if x is discrete random vector} \\ \int xf_x(x)\mathrm{d}x & \text{if x is continuous random vector} \end{cases}$$

and the covairance of Z and Y is defined as:

$$\mathrm{Cov}(Z, Y) = \mathbb{E}(ZY) - \mathbb{E}(Z) \cdot \mathbb{E}(Y) = \int (Z - \mathbb{E}(Z))(Y - \mathbb{E}(Y))\mathrm{d}F$$

Lemma 6. *If $\bar{a}$ is a vector and $\boldsymbol{x}$ is a random vector with mean μ and covariance matrix Σ, we have*

1. $\mathbb{E}(\bar{a}^T \boldsymbol{x}) = \bar{a}^T \mu$
2. $\mathrm{Var}(\bar{a}^T \boldsymbol{x}) = \bar{a} \Sigma \bar{a}^T$

If A is a matrix, we have

1. $\mathbb{E}(Ax) = A\mu$
2. $\mathrm{Cov}(Ax) = A\Sigma A^T$

Ex 7. The Multinomial Distribution A discrete random vector $\bar{x} = (x_1, ..., x_k, x_{k+1})$ has multivariate distribution of dimension k with parameters $\theta = (\theta_1, \theta_2, ..., \theta_k, \theta_{k+1})^T$ and n. $(0 \le \theta_i \le 1, \sum \theta_i < 1, n = 1, 2, ...)$. If its p.m.f is

$$f_{nk}(\bar{x}|\theta, n) = \frac{n!}{\prod_{i=1}^k x_i! \left(n - \sum_{i=1}^k x_i\right)!} \sum_{i=1}^k \theta_i^{x_i} \left(1 - \sum_{i=1}^k \theta_i\right)^{n - \sum_{i=1}^k x_i}$$

The above form indicate that if we restrict the sum of each sub-element to n and the sum of θ_i to 1, the probability mass function can be represented by the first k components.

The mean vector and covariance matrix are:

$$\mathbb{E}(X_i) = n\theta_i$$

$$\mathrm{Var}(X_i) = n\theta_i(1 - \theta_i)$$

$$\mathrm{Cov}(X_i X_j) = -n\theta_i\theta_j$$

Theorem 18. *The marginal distribution of $\boldsymbol{x}^{(m)} = (x_1, ..., x_m)$ where $m < k$ is the multimonial distribution with p.m.f*

$$f_{n,m}(\boldsymbol{x}^{(m)}|(\theta_1, \theta_2, ..., \theta_m), n) \tag{5.2}$$

The conditional distribution of $\boldsymbol{x}^{(m)}$ giving the remaining x_i is also multinomial distribution with p.m.f

$$f_{n-s,m-1}\left(\boldsymbol{x}^{(m)}|\left(\frac{\theta_1}{\sum_{j=1}^m \theta_j}, \frac{\theta_2}{\sum_{j=1}^m \theta_j}, ..., \frac{\theta_m}{\sum_{j=1}^m \theta_j}\right), n - s\right) \tag{5.3}$$

where $s = \sum_{j=m+1}^k x_j$.

This theorem tells us that the parameter of remaining part of θ_i are independent to the other part of θ which is only normalized. The conditional distribution only depends on $n - s$.

5.5 Dirichlet Distribution

After introducing the multinomial distrbution, we know that it can be used to do multi class classification. The problem now is how to set the prior distribution of multinomial distribution? The Dirichlet distribution is the general form of Beta distribution. It is the conjugate prior of the categorical distribution and multinomial distribution.

Definition 25 (Dirichlet Distribution)**.** A continuous random vector $\mathbf{x} = (x_1, ..., x_k)$ has a Dirichlet distribution of dimension k, with parameters $\alpha = (\alpha_1, ..., \alpha_{k+1})$ $(\alpha_i > 0, i = 1, ..., k+1)$ if its p.d.f $\text{Dir}(\mathbf{x}|\alpha)$

$$\text{Dir}(\mathbf{x}|\alpha) = \frac{\Gamma(\sum_{i=1}^{k+1} \alpha_i)}{\prod_{i=1}^{k+1} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{k} x_i\right)^{\alpha_{k+1} - 1} \tag{5.4}$$

The mean and covariance of Dirichlet distribution are given by

$$\mathbb{E}(X_i) = \frac{\alpha_i}{\sum_{j=1}^{k+1} \alpha_j}$$

$$\text{Var}(X_i) = \frac{\mathbb{E}(X_i) - (1 - \mathbb{E}(X_i))}{1 + \sum_{j=1}^{k+1} \alpha_j}$$

$$\text{Cov}(X_i, X_j) = \frac{\mathbb{E}(X_i) - (\mathbb{E}(X_j))}{1 + \sum_{j=1}^{k+1} \alpha_j}$$

Theorem 19. *The marginal distribution of $\boldsymbol{x}^{(m)} = (x_1, ..., x_m)$ where $m < k$ is the Dirichlet distribution*

$$\text{Dir}\left(\boldsymbol{x}^{(m)} \middle| \alpha_1, \alpha_2, ..., \alpha_m, \sum_{j=m+1}^{k+1} \alpha_j\right) \tag{5.5}$$

The conditional distribution given $x_{m+1}, ..., x_k$ of $y_i = \frac{x_i}{1 - \sum_{j=m+1}^{k+1} x_j}$ is also Dirichlet:

$$\text{Dir}\left((y_1, ..., y_m)|\alpha_1, ..., \alpha_m, \alpha_{k+1}\right) \tag{5.6}$$

Theorem 20 (Transformation)**.** *Given a random vector $\boldsymbol{x} = (x_1, ..., x_k)$ follows Dirichlet distribution and a random vector $\boldsymbol{z} = (z_1, ..., z_t)$ where*

$$\begin{aligned} z_1 &= x_1 + ... + x_{i1} & \beta_1 &= \alpha_1 + ... + \alpha_{i1} \\ z_2 &= x_{i1+1} + ... + x_{i2} & \beta_2 &= \alpha_{i1+1} + ... + \alpha_{i2} \\ &\vdots & &\vdots \\ z_t &= x_{it+1} + ... + x_k & \beta_t &= \alpha_{it+1} + ... + \alpha_{k+1} \end{aligned}$$

Then,

$$\boldsymbol{Z} \sim \text{Dir}(\boldsymbol{z}|\beta)$$

5.6 Multi Gaussian

Definition 26. Suppose $\mathbf{x}$ is a $p \times 1$ vector follows the Multi-Gaussian distribution and its covariance matrix $\Sigma_{p\times p}$ is positive definite which means $\Sigma \succ 0$ if its p.d.f is $\mathcal{N}_p(\mu, \Sigma)$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right) \tag{5.7}$$

where $\mathbb{E}(X) = \mu$ and $\mathrm{Cov}(X) = \Sigma$.

Theorem 21. *If $\boldsymbol{x} \sim \mathcal{N}_p(\mu, \Sigma)$ and X can be represented as a block matrix $\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \end{pmatrix}$ which $\boldsymbol{x}^{(1)} \in \mathbb{R}^p$ and $\boldsymbol{x}^{(2)} \in \mathbb{R}^{q-p}$. The mean and covariance matrix can be represented as $\mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Besides, we also define $x_{2\cdot1} = \boldsymbol{x}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{x}^{(1)}$. Therefore, we have the following conclusions.*

1. $\boldsymbol{x}^{(1)} \sim \mathcal{N}_q\left(\mu^{(1)}, \Sigma_{11}\right)$ *and* $\boldsymbol{x}_{2\cdot1} \sim \mathcal{N}_{pq}\left(\mu_{21}, \Sigma_{22\cdot1}\right)$[3]
2. $\boldsymbol{x}^{(1)} \perp\!\!\!\perp \boldsymbol{x}_{2\cdot1}$
3. $\boldsymbol{x}_2|\boldsymbol{x}_1 \sim \mathcal{N}_{pq}\left(\mu^{(2)} + \Sigma_{21}\Sigma_{11}^{-1}\left(\boldsymbol{x}^{(1)} - \mu^{(1)}\right), \Sigma_{22\cdot1}\right)$

Proof. The main idea of proving the three properties is to use Jaccobian.

$$\begin{cases} x_{2\cdot1} &= x^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}x^{(1)} = \left(-\Sigma_{21}\Sigma_{11}^{-1}, I\right)\begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix} \\ x^{(1)} &= x^{(1)} \end{cases}$$

Now we create a Jaccobian of the transformation $(x^{(1)}, x^{(2)}) \to (x^{(1)}, x_{2\cdot1})$

$$z = \begin{pmatrix} x^{(1)} \\ x_{2\cdot1} \end{pmatrix} = \begin{pmatrix} I & 0 \\ -\Sigma_{21}\Sigma_{11}^{-1} & I \end{pmatrix}\begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix}$$

The differential of z is

$$\mathrm{d}z = \det\begin{pmatrix} I & 0 \\ -\Sigma_{21}\Sigma_{11}^{-1} & I \end{pmatrix}\mathrm{d}x$$

By defining the matrix $\begin{pmatrix} I & 0 \\ -\Sigma_{21}\Sigma_{11}^{-1} & I \end{pmatrix}$ as B, we have

$$x - \mu = B^{-1}z$$

$$\Leftrightarrow$$

□

[3] We name $\Sigma_{22\cdot1}$ as **Shur Complement** which is defined as $\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.

5.7 Homework

1. Compute the Laplace transformation of Gamma, Negative Nomial and Poisson Distribution.

2. Consider that

   - $\omega_1 = \omega\alpha$, $\omega_2(1-\alpha)$
   - $u_1 = u - \beta\sigma\sqrt{\frac{\omega_2}{\omega_1}}$, $u_2 = u_1 + \beta\sigma\sqrt{\frac{\omega_1}{\omega_2}}$
   - $\sigma_1^2 = \gamma(1-\beta^2)\sigma^2\frac{\omega}{\omega_1}$, $\sigma_2^2 = (1-\gamma)(1-\beta^2)\sigma\frac{\omega}{\omega_2}$

   where $\alpha, \beta, \gamma \in (0,1)$.

   Compute the Jacobian from $(\omega_1, \omega_2, u_1, u_2, \sigma_1^2, \sigma_2^2)$ to $(\omega, u, \sigma^2, \alpha, \beta, \gamma)$

   Hint: you can use the property

   $$\sum_{i=1}^{k} \omega_i N(\mu_i, \sigma_i^2) = \sum_{j=1}^{k+1} \omega_i N(\mu_j, \sigma_j^2)$$

   This equation means k numbers of gaussian can be seperated into $k+1$ sub gaussian.

3. By using multinomial theorem, show the marginal distribution (5.2) and conditional distribution (5.3). The multinomial theorem is:

   $$(p_1 + p_2 + \cdots + p_k)^n = \sum \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{j=1}^{k} p_j^{x_j}$$

   where $\sum_{i=1}^{k} x_i = n$

4. Suppose $p(\mathbf{x}|\alpha) \sim$ Multinomial distribution and $\theta \sim \mathrm{Dir}(\theta|\alpha)$. Compute $p(\theta|\mathbf{x})$

Chapter 6

Sufficient Statistics

6.1 Sufficient Statistics