12

PROBABILITY THEORY

12.1 Sigma algebra

12.1.1 sigma algebra concepts

Definition 12.1.1 (σ algebra). *Given a set Ω, a **σ-field**, or **σ-algebra** is a collection $\mathcal{F}$ of subsets of Ω, with the following properties:*

1. $\emptyset \in \mathcal{F}$
2. *if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$*
3. *(countable union)if $A \in \mathcal{F}$, then $\cup_{i=0}^{\infty} A_i \in \mathcal{F}$*

Example 12.1.1.

1. The trivial σ-field $\mathcal{F} = \{\emptyset, \Omega\}$
2. The collection $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$, where A is a fixed subset of Ω
3. The set of all the subsets of finite set Ω.
4. For a finite sample space Ω, the power set of Ω is the largest σ field, $\{\emptyset, \Omega\}$ is the smallest σ field.

Remark 12.1.1. The pair $(X, \mathcal{F})$ is called **measurable space**, the members $e \in \mathcal{F}$ are called **measurable sets** or $\Sigma-$measurable sets.

Lemma 12.1.1 (intersection theorem). *[1] If $\{\mathcal{F}_\alpha\}_{\alpha \in T}$ is a collection of σ fields on Ω, then $\cap_{\alpha \in T} \mathcal{F}_\alpha$ is σ field on Ω*

Proof. We consider the special case $T = \{1, 2\}$. Let $A = \mathcal{F}_1 \cap \mathcal{F}_2$. It is easy to see $\emptyset \in A$; since $A \in \mathcal{F}_1 \cap \mathcal{F}_2$, then $A \in \mathcal{F}_1, A \in \mathcal{F}_2$, then $A^c \in \mathcal{F}_1, A^c \in \mathcal{F}_2$, then $A^c \in \mathcal{F}_1 \cap \mathcal{F}_2$; Similarly, we can prove the union property. □

12.1.2 Generation of sigma algebra

Lemma 12.1.2 (Existence of smallest σ field, σ algebra generation). *If $\mathcal{A}$ is a collection of subsets of Ω, then there exist a unique smallest σ field on Ω, containing $\mathcal{A}$, which is contained by all the σ fields that contains $\mathcal{A}$. We denote this by $\mathcal{F}(\mathcal{A})$, and called the σ field generated by $\mathcal{A}$.*

Proof. Consider $\mathcal{B}$ as the set of all σ fields that contains $\mathcal{A}$. The intersections of all these sets will lead to $\mathcal{F}(\mathcal{A})$ due to theorem 12.1.1. □

Definition 12.1.2 (sigma algebra generated by an event). *Let A be a subset of a set Ω. The sigma algebra generated by A, denoted by $\sigma(A)$, is a set given by*

$$\sigma(A) = \{\emptyset, \Omega, A, A^c\}.$$

Remark 12.1.2 (sigma algebra generated by random variable and stochastic process). The generation of sigma algebra by random variables and stochastic processes are discussed in Definition 12.3.4Definition 19.1.5.

Corollary 12.1.0.1 (Properties of generated σ algebra). *[1] If $\mathcal{A}, \mathcal{A}_1$ and $\mathcal{A}_2$ are subsets of 2^Ω, then we have*

- *If $\mathcal{A}_1 \subset \mathcal{A}_2$, then $\mathcal{F}(\mathcal{A}_1) \subset \mathcal{F}(\mathcal{A}_2)$*
- *If $\mathcal{A}$ is a σ field, then $\mathcal{F}(\mathcal{A}) = \mathcal{A}$*
- *If $\mathcal{F}(\mathcal{F}(\mathcal{A})) = \mathcal{F}(\mathcal{A})$*

12.1.3 Partition of sample space

Definition 12.1.3 (partition of sample space). *A collection of subsets of Ω, $\{\mathcal{A}_i\}_{i\in I}$ (I can have size of uncountable infinite)is called a partition of Ω if*

$$\mathcal{A}_i \cap \mathcal{A}_j = \emptyset, \text{ if } i \neq j$$

and

$$\cup A_i = \Omega.$$

Lemma 12.1.3. *[1] If $\mathcal{P} = \{A_i\}_{i=1}^\infty$ is a countable partition of Ω, then the σ field generated from $\mathcal{P}$, $\mathcal{F}(\mathcal{P})$, consists of all sets of the form $\cup_{n\in M} A_n$ where M ranges over all subsets of $\mathbb{N}$.*

Lemma 12.1.4. *Let $\mathcal{P}_1, \mathcal{P}_2$ be the partitions of the same set Ω. If $\mathcal{P}_2$ is obtained by subdividing sets in $\mathcal{P}_1$ (i.e. $\mathcal{P}_2$ is finer), then we have*

$$\mathcal{F}(\mathcal{P}_1) \subseteq \mathcal{F}(\mathcal{P}_2) \Leftrightarrow \mathcal{P}_1 \subseteq \mathcal{P}_2$$

12.1.4 Filtration & information

Definition 12.1.4 (filtration). *Let $(\Omega, \mathcal{F})$ denote a measurable space.*

- *A **continuous filtration** is defined as: A family of σ algebras $\{\mathcal{F}_t | t >= 0\}$ where*

$$\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}, 0 \leq s \leq t$$

- *A **discrete filtration** on $(\Omega, \mathcal{F})$ is an increasing sequence of σ fields $\{\mathcal{F}_n\}$ such that:*

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 ... \subseteq \mathcal{F}$$

Note that as the time progresses, the finer the σ algebra will be. We call $\mathcal{F}_n$ the history up to time n.

Remark 12.1.3. Note that usually not all the subsets of X can be defined a measure with above properties. For example, all irrational numbers in the real line, the root to polynomial equation, are not measurable sets.[2]

Remark 12.1.4 (filtration and information).

- Let $\mathcal{F}_1, \mathcal{F}_2$ be two σ field on Ω, then $\mathcal{F}_1 \subseteq \mathcal{F}_2$ mean $\mathcal{F}_2$ contains more information than $\mathcal{F}_1$; For any A is measurable with respect to $\mathcal{F}_1$ then A is measurable with respect to $\mathcal{F}_2$. That is, if $A \subset \mathcal{F}_1$ then $A \subset \mathcal{F}_2$.

Example 12.1.2. For example, in a die toss example, $\mathcal{F}_1$ is generated by the events of odd number or even number, while $\mathcal{F}_2$ is generated by the event of all possible outcomes. Then, we have $\mathcal{F}_1 \subset \mathcal{F}_2$,i.e., knowing the probability measure on $\mathcal{F}_2$ will enable us to calculate the probability measure on $\mathcal{F}_1$.[1]

Now consider a series of experiment: Let Ω denote the set of all outcomes resulting from tossing a coin three times, the $\Omega = \{(H,H,H),(T,H,H)...,(T,T,T)\}$. Let $\mathcal{F}_i$ denote the events that have been determined by the end of the i toss. Then $\mathcal{F}_1 = \mathcal{F}(\{(H,\cdot,\cdot),(T,\cdot,\cdot)\})$, where $\cdot$ represent it will range over H, T,i.e., $\mathcal{F}_1$ is generated from a partition of 2. Since we have more information later, we have

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3$$

Note that if $\mathcal{F}_2$ represents events determined by the i toss instead of tosses upto i, the above will not hold.

12.1.5 Borel σ algebra

Definition 12.1.5. *[3]*

- *A **Borel set** is a set in a topological space that can be formed from open sets (or from closed sets) through the operations of countable union, countable intersection, and relative complement.*
- *For a topological space X, the collection of all Borel sets on X forms a σ-algebra, known as the **Borel σ-algebra**. The Borel σ algebra on X is the smallest σ-algebra generated by open sets.*

Remark 12.1.5. Note that the elements like low-dimensional manifold $S \subset \mathbb{R}^m, m < n$ in $\mathbb{R}^n$ will not be in the $\mathcal{B}(\mathbb{R}^n)$, i.e., they cannot be obtained from open set operation defined above.

Note 12.1.1 (open interval close interval conversion). Using countable union and intersection properties, we can convert between open interval and close intervals, for example

- $(a, b) = \cup_{n=1}^{\infty}[a + 1/n, b - 1/n]$
- $[a, b] = \cap_{n=1}^{\infty}(a - 1/n, b]$
- $(a, \infty) = \cup_{n=1}^{\infty}[a, a + n]$
- singleton: $\{a\} = [a, a]$

12.1.6 Measurable set and measurable space

Definition 12.1.6 (measure). *Given a set X with its σ field Σ, a function $\mu : \Sigma \to \mathbb{R}$ is called a **measure** if it satisfies: [4]:*

- *Non-negativity: For all $E \in \Sigma$, $\mu(E) >= 0$*
- *$\mu(\emptyset) = 0$*
- *Countable additivity: For all countable collections $\{E_i\}$ of pairwise disjoint sets in Σ:*

$$\mu(\cup_{i=1} E_i) = \sum_{i=1} \mu(E_i)$$

*The pair (X, Σ) is called **measurable space**, the members $e \in \Sigma$ are called **measurable sets** or $\Sigma-$measurable sets. A triple (X, Σ, μ) is called measure space.*

Remark 12.1.6. A **measure** on a set is a systematic way to assign a number of each suitable subset of set, as a generalization of the concepts of length, area, and volume.

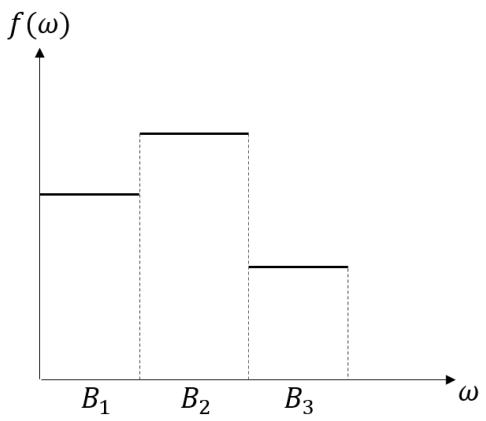
Example 12.1.3 (probability measure). A **probability measure** is a measure satisfying above three properties and has one additional requirement of total measure one $\mu(X) = 1$.

Definition 12.1.7 (measurable function, Borel measurable function). *Let $(\Omega, \mathcal{F})$ be a measurable space. A function $f : \Omega \to \mathbb{R}$ is said to be $\mathcal{F}$-measurable, or Borel measurable, if $f^{-1}(B) \in \mathcal{F}, B \in \mathcal{B}(\mathbb{R})$.*

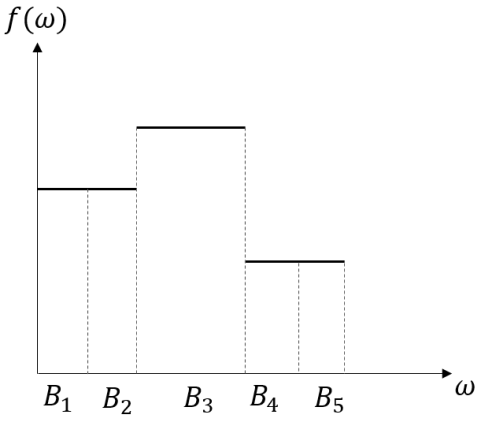
Example 12.1.4 (measurable function with coarse sigma field). Let $\mathcal{F}$ generated by a finite partitions $B_1, B_2, ..., B_m$ of Ω; let function $f : \Omega \to \mathbb{R}$ be **$\mathcal{F}$-measurable**. Then f take constant value on each element of $B_i, 1 \leq i \leq m$ [Figure 12.1.1].

Suppose f can take different values, say a_1, a_2, then the inverse image of the interval $[a_1, 0.5(a_1 + a_2)]$ is not a subset of $\mathcal{F}$ (note that $\mathcal{F}$ can only contain $\emptyset$ plus subsets due to unions of partition subset. See previous sections on partition of sample space), which contradicts the fact of f is measurable.

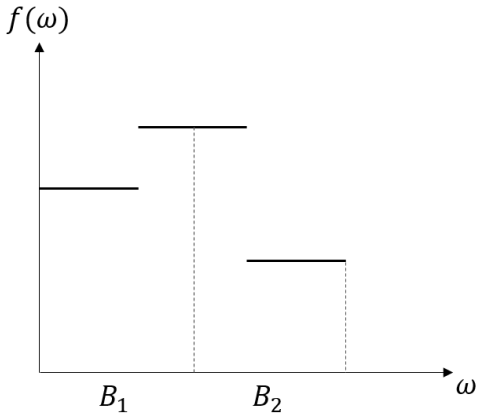
Therefore measurability usually limits the 'variation' of a function defined on a set.



(a) A mesurable function measurable with respect to the σ algebra generated by partition B_1, B_2, B_3.



(b) A mesurable function measurable with respect to the σ algebra generated by partition B_1, B_2, B_3, B_4, B_5.



(c) A function not measurable with respect to the σ algebra generated by partition B_1, B_2.

Figure 12.1.1: An illustration of measurable functions.

Note 12.1.2 (measurable functions vs. ordinary functions)**.**

- Ordinary functions from set A to set B simply establish a relationship between elements in A and elements in B. A measurable function from set A to set B also establish a relationship between elements in A and elements in B, however, under the constraint of measurability of σ fields.
- The level of coarseness constrain the number of values a measurable function can take. For a trivial $\mathcal{F}$, its measurable function can only take one value.
- For random variables, they are required to be measurable functions.

12.2 Probability space

12.2.1 Event, sample point and sample space

Definition 12.2.1 (event,sample point and sample space). *Consider a random experiment. The collection of all outcomes is the sample space Ω. Given a sample space Ω with its σ field $\mathcal{F}$, an event is simply an element in $\mathcal{F}$.*

Remark 12.2.1 (interpretation).

- The results of experiments or observations are called events. For example, the result of a measurement will be called an event. We shall distinguish between *compound*(or decomposable) and *simple*(or indecomposable) *events*. For example, saying that a throw with two dice resulted in "sum six" amounts to saying that it resulted in (1,5) or (2,4) ..., which can be decomposes to five simple events.
- The simple events will be called sample points. Every dis-decomposable result of the experiment is represented by one and only one, sample point. The aggregate of *all* sample points will be called sample space.

12.2.2 Probability space

Definition 12.2.2 (probability space). *[2] A probabilistic model is defined formally by a triple $(\Omega, \mathcal{F}, P)$, called a **probability space**, where*

1. *Ω is the sample space, the set of possible outcomes of the experiment.*
2. *$\mathcal{F}$ is a σ-field, a collection of subsets of Ω, containing Ω itself and the empty set $\emptyset$, and closed under the formation of complements, countable unions, and countable intersections.*
3. *P is a **probability measure** defined on σ-field $\mathcal{F}$, and has the property of:*
   - *$P(A) \geq 0, \forall A \in \mathcal{F}$*
   - *if $A_1, A_2, ... \in \mathcal{F}$ are **disjoints** subsets of Ω, we have **countable additivity** as: $P(A_1 \cup A_2 \cup A_3 ...) = P(A_1) + P(A_2) + P(A_3) + ...$*
   - *$P(\Omega) = 1$.*

Remark 12.2.2 (interpretation).

- Note that the σ-algebra is the collections of *measurable sets*. These are the subsets $A \subseteq \Omega$ where $P(A)$ is defined. In general, σ-field might not contain *all* subsets of Ω(For example, let Ω be a interval on the real line, then the set of all rational number in the interval is not in σ-field)

- **Note that we cannot extend to *uncountable unions*;** in this case, $\mathbb{F}$ would contain every subset A, since every subset can be written as $A = \cup_{x \in A}\{x\}$ and since the singleton sets $\{x\}$ are all in $\mathbb{F}$.

Definition 12.2.3 (discrete probability space). *A **discrete probability space** is a triplet $(\Omega, \mathbb{F}, \mathbb{P})$ such that*

1. *the sample space is finite or countable*
2. *the σ-field is the set of all subsets of Ω*
3. *the probability measure (a function) assigns a number in the set [0,1] to every pairwise disjoint subset of $A \subseteq \Omega$, given as*

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\})$$

and

$$\sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = 1$$

Example 12.2.1 (Infinite coin toss process(infinite Bernoulli experiments)). [5, p. 4]

- Consider the probability space for tossing a coin infinitely many time. We can define the sample space as $\Omega_\infty =$ the set of infinite sequences of Hs and Ts. A generic element of Ω_∞ will be denoted as $\omega = \omega_1\omega_2$...., where ω_n indicates the result of the nth coin toss $\omega_n = H \text{ or } T$.
- Example subsets in Ω are
  - A_H: the set of all sequences beginning with H. $A_H = \{\omega : \omega_1 = H\}$.
  - A_T: the set of all sequences beginning with H. $A_T = \{\omega : \omega_1 = H\}$.
  - A_{HT}: the set of all sequences beginning with HT. $A_{HT} = \{\omega : \omega_1 = H, \omega_1 = T\}$.
  - A_{TH}: the set of all sequences beginning with TH. $A_{TH} = \{\omega : \omega_1 = T, \omega_2 = H\}$.
- Possible σ algebra includes:
  - $\mathcal{F}_0 = \{0, \Omega_\infty\}$.
  - $\mathcal{F}_1 = \{0, \Omega_\infty, A_H, A_T\}$.
  -

$$\mathcal{F}_2 = \begin{array}{l} 0, \Omega_\infty, A_H, A_T, A_{HH}, A_{HT}, A_{TH}, A_{TT}, A_{HH}^C, A_{HT}^C, A_{TH}^C, A_{TT}^C, \\ A_{HH} \cup A_{TH}, A_{HH} \cup A_{TT}, A_{HT} \cup A_{TH}, A_{HT} \cup A_{TT} \end{array}$$

12.2.3 Properties of probability measure

Lemma 12.2.1 (basic properties of probability measure). *[6, p. 11]*

- $P(\emptyset) = 0$.
- *(finite additivity)if* $A_1, A_2, ..., A_n \in \mathcal{F}$ *are* ***disjoints*** *subsets of* Ω*, we have:* $P(A_1 \cup A_2 ... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n)$.
- *For each* $A \in \mathcal{F}$, $P(A^C) = 1 - P(A)$*, where* A^C *is the complement of* A *with respect to* Ω.
- *If* $A_1, A_2 \in \mathcal{F}$ *and* $A_1 \subset A_2$*, then* $P(A_1) \leq P(A_2)$.
- *For* $B \subset A$, $P(A - B) = P(A) - P(B)$.

Proof. (1) Directly from

$$P(\cup\emptyset) = P(\emptyset) = \sum P(\emptyset)$$

and $P(\emptyset) \geq 0$, we have $P(\emptyset) = 0$. (2) Set $A_{n+1}, A_{n+2}, ... = \emptyset$ and use (1). (3) from (2). (4) note that $A_2 = A_1 + (A_2 - A_1)$ and $P(A_2 - A_1) \geq 0$. (5) Note that $(A - B) \cup B = A$ such that

$$P(B) + P(A - B) = P(A).$$

□

Lemma 12.2.2 (union bound). *For any sequence* $A_1, A_2, ... \in \mathcal{F}$ $P(A_1 \cup A_2 \cup A_3 \cup ...) \leq P(A_1) + P(A_2) + ...$

Proof. Based on countable additivity of probability function, we have:

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3 \cup ...) &= P(A_1 \cup (A_2 \backslash A_1) \cup (A_3 \backslash A_2 \backslash A_1)...) \\ &= P(A_1) + P(A_2 \backslash A_1)... <= P(A_1) + P(A_2) + ... \end{aligned}$$

□

12.2.4 Conditional probability

12.2.4.1 *Basics*

In some random experiments, we are interested only in those outcomes that are elements of a subset C_1 of the sample space Ω. Then given the probability space $(\Omega, \mathcal{F}, P)$, and $C_1, C_2 \in \mathcal{F}$, the conditional probability of the event C_2, given C_1 is defined as

$$P(C_2|C_1) = \frac{P(C_1 \cap C_2)}{P(C_1)}.$$

Note that usually, for two events C_1, C_2 both occur, we can define an new event $C_3 = C_1 \cap C_2$, then we write $P(C_1, C_2) = P(C_3) = P(C_1 \cap C_2)$. $P(C_1 \cap C_2)$ is quite formal since it is based on set theory.

Definition 12.2.4 (conditional probability measure). *Given a probability space $(\Omega, \mathcal{F}, P)$ and an event $A \in \mathcal{F}, P(A) \neq 0$, we can define a conditional probability measure*

$$P_A(B) \triangleq P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Lemma 12.2.3 (basic properties of conditional probability measure). *[6] Consider the conditional probability measure conditioned on event A. We have*

- $P(B|A) \geq 0, \forall B \in \mathcal{F}$.
- $P(B|A) = 0, \forall B \in \mathcal{F}, A \cap B$.
- $P(A|A) = 1$.
- $P(\cup_{j=1}^{\infty} B_i|A) = \sum_{j=1}^{\infty} P(B_i|A)$, *provided that* $B_1, B_2, ... \in \mathcal{F}$ *are mutually exclusive event.*
- $\sum_{i=1}^{\infty} P(C_i|A) = 1$, *where* $C_1, C_2, ... \in \mathcal{F}$ *are the partition of* Ω.

Proof. (4)Use countable additivity property of the definition of probability space Definition 12.2.2, we have

$$P(\cup_{j=1}^{\infty} B_i|A) = \frac{P(\cup_{j=1}^{\infty} B_i \cap A)}{P(A)} = \sum_{j=1}^{\infty} \frac{P(B_i \cap A)}{P(A)} = \sum_{j=1}^{\infty} P(B_i|A).$$

(5) Note that $\cup_{i=1}^{\infty}(C_i \cap A) = A$. □

Theorem 12.2.1 (Law of total probability). *Given a set of subsets $C_1, C_2, ..., C_k$, which are mutual disjoint and partition the sample space Ω, then we have*

$$P(C) = P(C \cap C_1) + P(C \cap C_2) ... + P(C \cap C_k) = \sum_{i=1}^{k} P(C_i)P(C|C_i)$$

Proof. Note that we have $P(C \cap C_i) = P(C_i)P(C|C_i)$, then we get the law of total probability as:

$$P(C) = P(C_1)P(C|C_1) + P(C_2)P(C|C_2) + .. + P(C_k)P(C|C_k) = \sum_{i=1}^{k} P(C_i)P(C|C_i)$$

□

Theorem 12.2.2 (Bayes' theorem). *From the definition of the conditional probability, we have Bayes' theorem as:*

$$P(C_j|C) = \frac{P(C \cap C_j)}{P(C)} = \frac{P(C_i)P(C|C_i)}{\sum_{i=1}^{k} P(C_i)P(C|C_i)}$$

Proof. The law of total probability has been in the denominator. □

Theorem 12.2.3 (Conditional Bayes' theorem). *From the definition of the conditional probability, we have Bayes' theorem as:*

$$P(C_j|C) = \frac{P(C \cap C_j)}{P(C)} = \frac{P(C_i)P(C|C_i)}{\sum_{i=1}^{k} P(C_i)P(C|C_i)}$$

Proof. The law of total probability has been in the denominator. □

12.2.4.2 *Independence of events and sigma algebra*

Definition 12.2.5 (independence of event). *Given the probability space $(\Omega, \mathcal{F}, P)$, and $C_1, C_2 \in \mathcal{F}$, then we say C_1 and C_2 are independent if*

$$P(C_1 \cap C_2) = P(C_1)P(C_2).$$

Definition 12.2.6 (independence of σ algebras). *Given the probability space $(\Omega, \mathcal{F}, P)$, and $\mathcal{F}_1, \mathcal{F}_2 \subset \mathcal{F}$, then we say $\mathcal{F}_1$ and $\mathcal{F}_2$ are independent if*

$$P(A_1 \cap A_2) = P(A_1)P(A_2), \forall A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2.$$

Remark 12.2.3. Note that this is mathematical equivalent definition, which does not reveal the nature of independence in terms of set relationship. **The nature is that if two events are independent, then the occurrence of one event will not change our brief on the occurrence of the other event.**

Example 12.2.2. Consider the sample space of a random experiment is given as $\{(0,0),(0,1),(1,0),(1,1)\}$, with its σ field consists of all its subsets, and we define event $C_1 = \{(0,0),(0,1)\}$, and $C_2\{(0,1)\}$. So the occurrence of C_1 will change the our brief of C_2 from $1/4$ to $1/2$. Therefore, C_1 and C_2 are not independent to each other. Also, consider $C_3 = C_1^c$, then the occurrence of C_1 change our brief of C_1^c

to 0. If $C_4 = \Omega$, then the occurrence of C_4 will not change, and thus C_4 is always independent of other events. In summary, **independence between events is far more complicated than the simple set relations between events**

Remark 12.2.4. Here is an non-trivial example of independence. Consider the sample space as the product of two coin toss sample space, the event that the first toss get 1 is $\{(1,0),(1,1)\}$, which is independent of the other event that the second toss get 1(i.e. $\{(0,1),(1,1)\}$). The two events have finite intersections, but they are independent. Therefore, it seems that simply considering the set relationships between events can not yield complete information of independence. The nature of the random experiment, i.e., the probability measure, dictates the Independence. The intuition way to judge independence will be whether the occurrence of one events provides useful information, i.e., changes our brief, for the occurrence of the other event.

Lemma 12.2.4 (independence of complements). *If C_1, C_2 are independent, then C_1 and C_2^c,C_1^c and C_2, C_1^c and C_2^c are independent.*

Lemma 12.2.5. *[1]*

- *If $P(A) > 0$, then A and B are independent if and only if $P(B|A) = P(B)$*
- *If A and B are independent, then A and B^C are independent.*
- *If $P(A) = 0$ or 1, then for any $B \in \mathcal{F}, B \neq A$, A and B are independent.*

Proof. (1) Suppose A and B are independent, then

$$P(A \cap B) = P(A)P(B) = P(A)P(B|A) \implies P(B|A) = P(B).$$

(2)

$$P(A \cap B^C) = P(A \cap (\Omega - B)) = P(A \cap \Omega) - P(A \cap B) = P(A) - P(A)P(B) = P(A)P(B^C).$$

(3) If $A = \Omega$ such that $P(A) = 1$,then

$$P(A \cap B) = P(B) = P(A)P(B).$$

If $A = \emptyset$ such that $P(A) = 0$,then

$$P(A \cap B) = P(A) = 0 = P(A)P(B).$$

□

12.3 Measurable map and random variable

12.3.1 Random variable

Definition 12.3.1 (measurable map). *[7] Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ be two measurable space. A map $T : \Omega \to S$ is called $(\mathcal{F}, \mathcal{S})$-**measurable map** if*

$$T^{-1}(A) \in \mathcal{F}, \forall A \in \mathcal{S}$$

We can also write it as

$$T : (\Omega, \mathcal{F}) \to (S, \mathcal{S}).$$

Lemma 12.3.1 (basic properties of measurable map). *Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ be two measurable space. Let a map $T : \Omega \to S$ be a measurable map. Then we have:*

- *For any two disjoint sets $S_1, S_2 \in \mathcal{S}$, $T^{-1}(S_1)$ and $T^{-1}(S_2)$ are disjoint.*
- *$T^{-1}(S) = \Omega$.*
- *(measurable composition preserves measurability) Let G be a measurable map from $(S, \mathcal{S})$ to $(S, \mathcal{S})$. Then $G \circ T : \Omega \to S$ is a measurable map.*

Proof. (1) Suppose their inverse image intersection M is nonempty, then $T(m), m \in M$ will map a single element to two different elements in S, which violates the definitions of mapping. (2) Suppose $T^{-1}(S) = \Omega_1 \subset \Omega$ and $\Omega_1 \neq \Omega$, then $T(\Omega - \Omega_1) = \emptyset$(otherwise T will map a single element to two different elements). Therefore $T^{-1}(S \cup \emptyset) = T^{-1}(S) = \Omega$. (3) Note that $(G \circ T)^{-1}(B) = T^{-1} \circ G^{-1}(B) = T^{-1}(G^{-1}(B))$ Because G is measurable map, $G^{-1}(B) \in \mathcal{S}$. Because T is measurable map, $T^{-1}(G^{-1}(B)) \in \mathcal{F}$. Therefore, $G \circ T$ is a measurable from Ω to S. □

Definition 12.3.2 (random variable in real space). *Let $(\Omega, \mathcal{F})$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be two measurable spaces defined on sample space Ω and $\mathbb{R}$, respectively.*

- *. A **random variable** in real space is a measurable map $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.*
- *A **n-dimensional real-valued random vector** is a measurable map $X : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.*

Lemma 12.3.2 (basic measurability properties of random variables). *Let $(\Omega, \mathcal{F})$ and $(\mathbb{R}, \mathcal{B})$ be two measurable space. Let $X : \Omega \to \mathbb{R}$ be a random variable. Then we have:*

- *(measurable composition preserves measurability)Let f be a measurable map from $(S,\mathcal{S})$ to $(S,\mathcal{S})$. Then $f(X):\Omega\to\mathbb{R}$ is a measurable map.*
- *Let Y be another random variable from $(\Omega,\mathcal{F})$ to $(R,\mathcal{B})$. Then $\alpha X+\beta Y:\Omega\to\mathbb{R},\alpha,\beta\in\mathbb{R}$ is also a measurable map.*
- *Let Y be another random variable from $(\Omega,\mathcal{F})$ to $(R,\mathcal{B})$. Then $XY:\Omega\to\mathbb{R}$ is also a measurable map.*
- *Let $Y,Y\neq 0$ be another random variable from $(\Omega,\mathcal{F})$ to $(R,\mathcal{B})$. Then $1/Y:\Omega\to\mathbb{R}$ is also a measurable map.*

Proof. (1) Use the composition property of measurable map [Lemma 12.3.1]. (2)(3)(4) use Lemma 3.8.4. □

Remark 12.3.1 (implications). This theorem provides the foundation of when X and Y are random variables, usually, $f(X),X+Y,XY,X/Y,...$ are also random variables.

12.3.2 Image measure

Definition 12.3.3 (image measure). *[7] Let $X:(\Omega,\mathcal{F})\to(S,\mathcal{S})$ denote a random variable and P a measure on the measurable space $(\Omega,\mathcal{F})$. Then*

$$P_X(A):=P(X^{-1}(A)),\forall A\in\mathcal{S}$$

defines a probability measure on $(S,\mathcal{S})$, which we call the image measure of P_X with respect to X.

Theorem 12.3.1 (generation of probability space via random variable). *Let $(\Omega,\mathcal{F},P)$ be a probability space, and let $X:\Omega\to\mathbb{R}$ be a random variable. For each Borel set $B\in\mathcal{B},B\subset\mathbb{R}$, we have $X^{-1}(B)\in\mathcal{F}$, then we can define $P_X(B)=P(X^{-1}(B))$. Then $(\mathbb{R},\mathcal{B},P_X)$ is a probability space.*

Proof. First $(\mathbb{R},\mathcal{B})$ form a measurable space. So we only need to check the axiom property of P_X: (1) $P_X(A)\geq 0,\forall A\in\mathcal{B}$; (2) For any two disjoint sets A_1,A_2, then

$$P_X(A_1\cup A_2)=P(X^{-1}(A_1\cup A_2))=P(X^{-1}(A_1)\cup X^{-1}(A_2))=P(X^{-1}(A_1)+P(X^{-1}(A_2)).$$

We can directly generalize to countable additivity. (3) $P(X^{-1}(\mathbb{R}))=P(\Omega)=1$(from lemma on basic properties of measurable maps) □

Remark 12.3.2. This lemma has significant consequences that **it enables us to directly work on this generated probability space $(\mathbb{R}, \mathcal{B}, P_X)$ and investigate distribution, density functions etc without referring back to the original probability space.**

12.3.3 σ algebra of random variables

Definition 12.3.4 (σ algebra generated by random variables). *[5, p. 52] Let X be a random variable map from nonempty Ω to $\mathbb{R}$. The σ**algebra generated** by X, denoted by $\sigma(X)$, is the collection of all subsets of Ω of the form $\{\omega \in \Omega : X(\omega) \in B\}$, or equivalently $X^{-1}(B)$, where B ranges over all Borel subsets of $\mathbb{R}$.*

Remark 12.3.3 (interpretation).

- When we define the measurable map from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$, usually $\sigma(X) \subseteq \mathcal{F}$. For example, if $X = const$, then $\sigma(X) = \mathcal{F}_0 = \{\emptyset, \Omega\}$.
- We cannot have $\mathcal{F} \subset \sigma(X), \mathcal{F} \neq \sigma(X)$ because the definition of random variable require measurability.

Definition 12.3.5 (measurable random variables with respect to a σ algebra). *[5, p. 53] Let X be a random variable map from nonempty Ω to $\mathbb{R}$. Let $\mathcal{G}$ be the σ algebra defined on Ω. We say X is $\mathcal{G}$ measurable if $\sigma(X) \subseteq \mathcal{G}$.*

Remark 12.3.4 (interpretation).

- Note that for any $B \in \mathcal{B}$, $X^{-1}(B) \in \sigma(X) \subseteq \mathcal{G}$, therefore X is also $\mathcal{G}$-measurable.
- Given a set Ω, we can define different σ algebra, including $\mathcal{F}_0 = \{\emptyset, \Omega\}$. But only σ algebra finer than $\sigma(X)$ can measure the mapping X.

12.3.4 Independence of random variables

Definition 12.3.6 (independence of random variables). *Let $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ and $Y : (\Omega, \mathcal{F}) \to (S, \mathcal{S}))$ denote two random variables. We say X, Y are independent, if for all $A, B \in \mathcal{S}$ the events $X^{-1}(A)$ and $Y^{-1}(B)$ are independent in the sense that $P(X^{-1}(A) \cap Y^{-1}(B)) = P(X^{-1}(A))P(Y^{-1}(B))$.*

Definition 12.3.7 (independence of random variables, alternative). *Let* $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *and* $Y : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *denote two random variables. We say* X, Y *are independent, if for any events* $A, B, A \in \sigma(X), B \in \sigma(Y)$

$$P(A \cap B) = P(A)P(B)$$

Remark 12.3.5. Note that

- independence of random variables are much more than independence of a selected set of events, because it requires that *all* events are independent to each other.
- if X, Y are map from different sample space, then they are independent.

Lemma 12.3.3 (function composition preserves random variable independence). *Let* X, Y *be independent random variables defined from* Ω *to* $\mathbb{R}$ *, and let* f *and* g *be Borel-measurable functions on* $\mathbb{R}$*. Then* $f(X)$ *and* $g(Y)$ *are independent random variables.*

Proof. Note that for any $B \in \mathcal{B}$, $f^{-1}(B) \in \mathcal{B}$ since f is Borel measurable. Then $X^{-1}(f^{-1}(B)) \in \sigma(X)$ based on the definition of σ generation. Therefore, $\sigma(f(X)) \subset \sigma(X)$. Similarly, $\sigma(g(Y)) \subset \sigma(Y)$. Since every events in $\sigma(X)$ and $\sigma(Y)$ are independent, then every events in $\sigma(f(X))$ and $\sigma(g(Y))$ are independent; that is, $f(X)$ and $g(Y)$ are independent random variables. □

12.4 Distributions of random variables

12.4.1 Basic concepts

12.4.1.1 *Probability mass function*

Definition 12.4.1 (random variable, random vector). *[6, p. 75]*

- *Let X be a random variables maps from the probability space $(\Omega, \mathcal{F}, P)$ to $\mathbb{R}$. The **space** of the random variable X is the set*

$$\{(X(\omega) : \omega \in \Omega\}.$$

- *Let $X_1, X_2, ..., X_n$ be random variables maps from the probability space $(\Omega, \mathcal{F}, P)$ to $\mathbb{R}$. We say $(X_1, X_2, ..., X_n)$ is a random vector. The **space** of the random vector $(X_1, X_2, ..., X_n)$ is the set*

$$\{(X_1(\omega), X_2(\omega), ..., X_n(\omega)) : \omega \in \Omega\}.$$

Definition 12.4.2 (probability mass function). *[8]*

- *For a discrete random variable X with space $\mathcal{D}$, the **probability mass function** to characterize its distribution is given by*

$$f_X(x) = P(X = x), \forall x \in \mathcal{D}.$$

- *For a discrete random vector $(X_1, X_2, ..., X_n)$ with space $\mathcal{D}$, the **joint probability mass function** to characterize its distribution is given by*

$$f_{X_1, X_2, ..., X_n}(x_1, x_2, ..., x_n) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n), \forall (x_1, x_2, ..., x_n) \in \mathcal{D}.$$

12.4.1.2 *Distributions on $\mathbb{R}^n$*

Definition 12.4.3 (cumulative distribution functions).

- *Let X be a random variable with space $\mathcal{D} \subset \mathbb{R}$. The cumulative distribution function for X is given by*

$$F_X(x) = Pr(X \leq x).$$

- *Let $(X_1, X_2, ..., X_n)$ be a random vector with space $\mathcal{D} \subset \mathbb{R}^n$. The joint cumulative distribution function for X is given by*

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = Pr(X_1 \leq x_1, X_2 \leq x_2, ..., X_n \leq x_n).$$

Remark 12.4.1 (A rigorous interpretation). [7]

- Let P denotes a probability measure of the original probability space $(\Omega, \mathcal{F}, P)$. Let X be the random variable, then

$$F_X(x) := Pr(X \leq x) = P(X^{-1}((-\infty, x))).$$

  where X^{-1} maps a measurable subset in $\mathcal{B}(\mathbb{R})$ to a measurable set in $\mathcal{F}$.
- Note that every subset of such form $(-\infty, x), x \in \mathbb{R}$ is a member of $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and therefore $Pr(X \leq x) = P(X^{-1}((-\infty, x)))$ has a well-defined value.

Definition 12.4.4 (marginal cdf). *Let $(X_1, X_2, ..., X_n)$ be a random vector with joint cdf $F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)$. The marginal cdf of X_i is defined by*

$$F_{X_i}(x_i) = Pr(X_1 < \infty, ..., X_i \leq x_i, ..., X_n < \infty).$$

Lemma 12.4.1 (area probability formula). *[6, p. 76] Let X_1, X_2 be random variables with joint cdf $F_{X_1,X_2}(x_1, x_2)$. Then*

$$\begin{aligned} &Pr(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2) \\ =&F_{X_1,X_2}(b_1, b_2) - F_{X_1,X_2}(a_1, b_2) - F_{X_1,X_2}(b_1, a_2) + F_{X_1,X_2}(a_1, a_2) \end{aligned}$$

Proof. Straight forward. □

12.4.1.3 *Probability density function*

Definition 12.4.5 (probability density function, pdf).

- *Let X be a random variable with cdf $F_X(x)$. The probability density function for X is defined by*

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

- *Let $(X_1, X_2, ..., X_n)$ be a random vector with joint cdf $F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)$. The joint probability density function for $(X_1, X_2, ..., X_n)$ is given by*

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = \frac{\partial^n F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}.$$

Definition 12.4.6 (support of a random variable). *Let X be a random variable with pdf f_X and space $\mathcal{D}$. The support of X is defined as the set*

$$S_X = \{x \in \mathcal{D} : f_X(x) > 0\}.$$

Definition 12.4.7 (marginal pdf). *Let $(X_1, X_2, ..., X_n)$ be a random vector with joint pdf $f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)$. The marginal pdf of X_i is defined by*

$$f_{X_i}(x) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{X_1,X_2,...,X_n} dx_1 ... dx_{i-1} dx_{i+1} ... dx_n.$$

If the marginal cdf of X_i is F_{X_i}, then

$$f_{X_i}(x) = \frac{dF_{X_i}(x)}{dx}.$$

12.4.1.4 *Conditional distributions*

Definition 12.4.8 (conditional probability mass function (pmf)). *Let X_1 and X_2 be discrete random variables with joint pmf $p_{X_1,X_2}(x_1, x_2)$. Let $p_{X_1}(x_1)$ denote the marginal pmf Let x_1 be a point such that $p_{X_1}(x_1) > 0$.*

The conditional pmf of X_2 given $X_1 = x_1$ is defined as

$$p_{X_2|X_1}(x_2|x_1) = \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_1}(x_1)}.$$

Remark 12.4.2. The sum to 1 property can be verified by

$$\begin{aligned}
&\sum_{x_2} p_{X_2|X_1}\left(x_2|x_1\right) \\
&= \sum_{x_2} \frac{p_{X_1,X_2}\left(x_1,x_2\right)}{p_{X_1}\left(x_1\right)} \\
&= \frac{1}{p_{X_1}\left(x_1\right)} \sum_{x_2} p_{X_1,X_2}\left(x_1,x_2\right) = \frac{p_{X_1}\left(x_1\right)}{p_{X_1}\left(x_1\right)} = 1
\end{aligned}$$

Definition 12.4.9 (conditional probability density function (pdf)). *Let X_1 and X_2 be discrete random variables with joint pdf $f_{X_1,X_2}\left(x_1,x_2\right)$. Let $f_{X_1}\left(x_1\right)$ denote the marginal pmf Let x_1 be a point such that $f_{X_1}(x_1) > 0$.*

The conditional pdf of X_2 given $X_1 = x_1$ is defined as

$$f_{X_2|X_1}(x_2;x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)}.$$

Lemma 12.4.2 (basic properties). *[6, p. 97]*

- $f_{X_2|X_1}(x_2;x_1) > 0$
- $\int_{-\infty}^{\infty} f_{X_2|X_1}(x_2;x_1)dx_2 = 1.$
- $\int_{-\infty}^{\infty} f_{X_2|X_1}(x_2;x_1)f_{X_1}(x_1)dx_1 = f_{X_2}(x_2).$
- $E[u(X_2)|x_1] = \int_{-\infty}^{\infty} f_{X_2|X_1}(x_2;x_1)u(x_1)dx_2$

12.4.1.5 *Bayes law*

Theorem 12.4.1 (Bayes law for random variables). *Let X, Y, Z be random variables. It follows that*

- *(unconditional Bayesian law)*

$$f(X|Y) = \frac{f(Y|X)f(X)}{\int f(Y|X)f(X)dx}$$

- *(conditional Bayesian law)*

$$f(X|Y,Z) = \frac{f(X|Z)f(Y|X)}{\int f(Y|X)f(X|Z)dx}$$

Proof. (1) Note that the denominator $\int f(Y|X)f(X)dx = f(Y)$. Therefore

$$f(X|Y)\int f(Y|X)f(X)dx = f(X,Y) = f(Y|X)f(X).$$

(2) Note that

$$\int f(Y|X)f(X|Z)dx = \int f(Y,X|Z)dx = f(Y|Z).$$

Therefore,

$$f(X|Y,Z)\int f(Y|X)f(X|Z)dx = f(X|Y,Z)f(Y|Z) = f(X,Y|Z).$$

□

12.4.2 Independence

Definition 12.4.10 (independence of random variables). *[6, pp. 112, 115] Let the random variables X_1 and X_2 have joint pdf $f(x_1,x_2)$ and marginal pdfs $f_1(x_1), f_2(x_2)$.*

- *The random variables X_1 and X_2 are said to be independent if and only*

$$Pr(a < X_1 \leq b, c < X_2 \leq d) = Pr(a < X_1 \leq b, c < X_2 \leq d),$$

  for every $a < b, c < d$, where a, b, c, d are constants.
- *The random variables X_1 and X_2 are said to be independent if and only*

$$f(x_1,x_2) = f_1(x_1)f_2(x_2).$$

Lemma 12.4.3 (conditions for independence). *[6, p. 113] Let the random variables X_1 and X_2 have supports S_1 and S_2, and have joint pdf $f(x_1,x_2)$. Then X_1 and X_2 are independent if and only if $f(x_1,x_2)$ can be written as*

$$f(x_1,x_2) = g(x_1)h(x_2),$$

where $g(x_1) > 0, x_1 \in S_1$, zero elsewhere, and $h(x_2) > 0, x_2 \in S_2$, zero elsewhere.

Proof. (1) If $f(x_1,x_2) = g(x_1)h(x_2)$, then

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} g(x_1)h(x_2)dx_2 = c_2 g(x_1).$$

$$f_{X_2}(x_1) = \int_{-\infty}^{\infty} g(x_1)h(x_2)dx_1 = c_1 h(x_2).$$

Further we have

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1)h(x_2)dx_1dx_2 = c_1c_2.$$

Therefore, $f(x_1, x_2) = c_1c_2g(x_1)h(x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$; that is, X_1, X_2 are independent.

(2) The other direction directly from definition. □

Lemma 12.4.4 (independence criterion from mgf). *[6, p. 114] Let the random variables X_1 and X_2 have joint cdf $F(x_1, x_2)$ and marginal cdfs $F_1(x_1), F_2(x_2)$. Then X_1 and X_2 are independent if and only if*

$$F(x_1, x_2) = F_1(x_1)F_2(x_2).$$

Proof. (1) From Lemma 12.4.1, we have

$$\begin{aligned}
&Pr(a_1 \le X_1 \le b_1, a_2 \le X_2 \le b_2)\\
=&F_{X_1,X_2}(b_1, b_2) - F_{X_1,X_2}(a_1, b_2) - F_{X_1,X_2}(b_1, a_2) + F_{X_1,X_2}(a_1, a_2)\\
=&F_{X_1}(b_1)F_{X_2}(b_2) - F_{X_1}(a_1)F_{X_2}(b_2) - F_{X_1}(a_2)F_{X_2}(b_1) + F_{X_1}(a_1)F_{X_2}(a_2)\\
=&(F_{X_1}(b_1) - F_{X_1}(a_1))(F_{X_2}(b_2) - F_{X_2}(a_2))\\
=&Pr(a_1 \le X_1 \le b_1)Pr(a_2 \le X_2 \le b_2)
\end{aligned}$$

Since a_1, a_2, b_1, b_2 are arbitrary, X_1 and X_2 are independent. (2)The other direction directly from definition. □

Lemma 12.4.5 (independence from moment generating functions). *link Let X and Y be two random variables with space $\mathbb{R}$. Assume the moment generating functions for X, Y and $X + Y$ exist at the neighborhood of o. If for all t_X, t_Y in the neighborhood of o we have*

$$E[\exp(t_XX + t_YY)] = E[\exp(t_XX)]E[\exp(t_YY)],$$

then X and Y are independent.

Proof. Let (U, V) be such that U and V are independent; moreover, U and X have the same distribution and V and Y have the same distribution.

$$\begin{aligned}
E[\exp(t_XX + t_YY)] &= E[\exp(t_XX)]E[\exp(t_YY)]\\
&= E[\exp(t_XU)]E[\exp(t_YV)] = E[\exp(t_XU + t_YV)],
\end{aligned}$$

Therefore (X, Y) and (U, V) have the same joint distribution; that is, X and Y are independent. □

12.4.3 Conditional independence

Definition 12.4.11 (conditional independence). *Given discrete random variables X, Y, and Z, we say X and Y are conditionally independent on Z if we can be write:*

$$P(X, Y|Z = z) = P(X|Z = z)P(Y|Z = z).$$

If not conditionally independent, we will have

$$P(X, Y|Z = z) = P(X|Y, Z = z)P(Y|Z = z).$$

Remark 12.4.3. Intuitively, two random variable X, Y are conditional independence given Z is that: if the value of Z is known, X, Y are independent to each other,i.e., the occurrence of events about Y will not give extra information to the occurrence of events abouts X. We need to distinguish two different cases:

- If X, Y are independent, then they are conditionally independent to each other.
- If events about Z already gives information contained in events about Y, then X, Y are conditionaly independent given Z.

Remark 12.4.4. Conditionally independence will help us simplify calculation, for example:

$$P(X = x|Y = y, Z = z) = P(X = x|Z = z)$$

if X, Y are conditionally independent given Z.

12.4.4 Transformations

12.4.4.1 *Transformation for univariate distribution*

Lemma 12.4.6 (change of variable). *[8, p. 77] Let X have cdf $F_X(x)$ and Let $Y = g(X)$, where g is a **monotonely increasing** function. Then,*

$$F_Y(y) = F_X(g^{-1}(y)).$$

*If g is a **monotonely decreasing function**, then*

$$F_Y(y) = 1 - F_X(g^{-1}(y)).$$

Proof. If g is increasing function

$$P(Y < y) = P(g(X) < y) = P(X < g^{-1}(y)) = F_X(g^{-1}(y)).$$

If g is decreasing function

$$P(Y < y) = P(g(X) < y) = P(X > g^{-1}(y)) = 1 - F_X(g^{-1}(y)).$$

□

Lemma 12.4.7 (change of variable). *[8, p. 77] Let X have pdf $f_X(x)$ and Let $Y = g(X)$, where g is a **monotone** function. Let X and Y be defined as*

$$\mathcal{X} = \{x : f_X(x) > 0\}, \mathcal{Y} = \{y : y = g(x), x \in \mathcal{X}\}$$

then we have

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|, y \in \mathcal{Y} \\ 0, otherwise \end{cases}$$

Remark 12.4.5 (why monotonity). We require the $g(X)$ to be monotone because if $g'(x)$ has different sign on different regions, then $g'(x_0) = 0$ for some x_0 and $g(x)$ is not invertible near the neighborhood of x_0.

Corollary 12.4.1.1. *Let $Y = g(X)$, where g is a monotone function, let $m(x)$ be a function, then*

$$\int_{\mathcal{Y}} m(y) f_Y(y) dy = \int_{\mathcal{X}} m(g(x)) f_X(x) dx$$

Proof.

$$\int_{\mathcal{Y}} m(y) f_Y(y) dy = \int_{\mathcal{Y}} m(y) f_X(g^{-1}(y)) \left| dx/dy \right| dy$$

Let $y = g(x), dy = (dy/dx)dx$, then

$$\int_{\mathcal{Y}} m(y) f_X(g^{-1}(y)) \left| dx/dy \right| dy = \int_{\mathcal{X}} m(g(x)) f_X(x) dx$$

□

12.4.4.2 *Location-scale transformation*

Definition 12.4.12. *[8] (Location-scale family)Let $f(x)$ be any pdf. Then for any $\mu, -\infty < \mu < \infty$, and any $\sigma > 0$, the family of pdfs*

$$\frac{1}{\sigma} f(\frac{x - \mu}{\sigma})$$

is the location-scale family indexed by μ, σ.

Remark 12.4.6. When $\sigma > 1$, we strech the original pdf; when $\sigma < 1$, we contract it.

Lemma 12.4.8 (location-scale transformation). *[8, p. 116] Let $f(\cdot)$ be any pdf. Then for any $\mu, -\infty < \mu < \infty$, and any $\sigma > 0$:*

- *X is a random variable with pdf*

$$\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$$

  if and only if $X = \sigma Z + \mu$ and Z has pdf $f(z)$.
- $F_X(x) = F_Z(\frac{x-\mu}{\sigma})$
- $F_X^{-1}(\alpha) = \mu + \sigma F_Z^{-1}(\alpha), \forall \alpha \in [0,1]$, *where* $F_X^{-1}(\alpha) = \inf\{Pr(X < x) \geq \alpha\}$.
- $EX = \sigma EZ + \mu, Var(X) = \sigma^2 Var(Z)$

Proof. Directly from [Lemma 12.4.7, Corollary 12.4.1.1]. For (1)(2)

$$\begin{aligned} F_X(x) &= Pr(X < x) \\ &= Pr(\mu + \sigma Z < x) \\ &= Pr(Z < (x-\mu)/\sigma) \\ &= F_Z((x-\mu)/\sigma) \end{aligned}$$

Then

$$f_X(x) = dF_X(x)/dx = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma}).$$

(3)

$$\begin{aligned} \alpha &= Pr(X < F_X^{-1}(\alpha)) \\ &= Pr(\sigma Z + \mu < F_X^{-1}(\alpha)) \\ &= Pr(Z < (F_X^{-1}(\alpha) - \mu)/\sigma) \\ \alpha &= F_Z((F_X^{-1}(\alpha) - \mu)/\sigma) \end{aligned}$$

$$\begin{aligned} F_Z^{-1}(\alpha) &= (F_X^{-1}(\alpha) - \mu)/\sigma \\ \mu + \sigma F_Z^{-1}(\alpha) &= F_X^{-1}(\alpha). \end{aligned}$$

□

Example 12.4.1. Consider the random variable $X \sim N(0,1)$ with

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}).$$

Let $Y = \sigma X + \mu$, then

$$f_Y(y) = \frac{1}{\sigma} f_X(\frac{y-\mu}{\sigma}) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

12.4.4.3 *Transformation for multivariate distribution*

Lemma 12.4.9 (multivariate transformation). *[6, p. 128] Let $(X_1, X_2, ..., X_n)$ be a random vector with support $\mathcal{S}$. Let*

$$y_1 = y_1(x_1, ..., x_n), ..., y_n = y_n(x_1, ..., x_n)$$

define a set of transformations with inverse

$$x_1 = x_1(y_1, ..., y_n), ..., x_n = x_n(y_1, ..., y_n).$$

Let $\mathcal{T}$ be the image of $\mathcal{S}$ under the transformation.

Let $f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)$ be the joint pdf of $(X_1, X_2, ..., X_n)$. Then the joint pdf for the random vector $(Y_1, Y_2, ..., Y_n)$ is given by

$$f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = f_{Y_1,Y_2,...,Y_n}(y_1(x_1, x_2, ..., x_n), ..., y_n(x_1, x_2, ..., x_n))|J|$$

or

$$f_{X_1,X_2,...,X_n}(x_1(y_1, y_2, ..., y_n), ..., x_n(y_1, y_2, ..., y_n)) = f_{Y_1,Y_2,...,Y_n}(y_1, y_2, ..., y_n)|J|$$

where

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \dots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \dots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}.$$

Moreover,

$$\int_{\mathcal{T}} f_{Y_1,...,Y_n}(y_1, ..., y_n)|J| \, dy_1 ... dy_n = 1$$

Proof. (1)For S be a measurable subset in $\mathcal{S}$, let $T \in \mathcal{T}$ denote the its image under the transformation. We have

$$Pr((Y_1, Y_2, ..., Y_n) \in T) = Pr((X_1, X_2, ..., X_n) \in S)$$

Note that

$$Pr((X_1, X_2, ..., X_n) \in S) = \int_S f_{X_1,...,X_n}(x_1, ..., x_n) dx_1 ... dx_n$$

and

$$dx_1 dx_2 = |J| \, dy_1 dy_2.$$

Then

$$\int_S f_{X_1,...,X_n}(x_1, ..., x_n) dx_1 ... dx_n = \int_T f_{Y_1,...,Y_n}(y_1(x_1, ..., x_n), ..., y_n(x_1, ..., x_n)) |J| \, dy_1 ... dy_n.$$

Because S is arbitrary, we have

$$f_{X_1,...,X_n}(x_1, ..., x_n) = f_{Y_1,...,Y_n}(y_1(x_1, ..., x_n), ..., y_n(x_1, ..., x_n)) |J| \, .$$

(2)

$$\int_{\mathcal{T}} f_{Y_1,...,Y_n}(y_1, ..., y_n) |J| \, dy_1 ... dy_n = \int_{\mathcal{S}} f_{X_1,...,X_n}(x_1, ..., x_n) dx_1 ... dx_n = 1$$

□

Remark 12.4.7.

- We interpret $dx_1 dx_2$ as infinitesimal area in the original $\mathcal{S}$, and this area is mapped to an area in $\mathcal{T}$. Note that we divide $\mathcal{S}$ and $\mathcal{T}$ into the same number small areas and the sum them up to calculate the integral. The areas in both $\mathcal{T}$ and $\mathcal{S}$ have the following relation:

$$dx_1 dx_2 = |J| \, dy_1(x_1, ..., x_2) dy_2(x_1, ..., x_n) = |J| \, dy_1 dy_2.$$

- If we maps from larger support to a smaller support, for example, from $\mathbb{R}^2$ to $[0, \infty) \times [0, 2\pi]$, the density will increase.
-

Lemma 12.4.10 (polar transformation). *Let (X_1, X_2) be a random vector with support $S = \mathbb{R}^2$. Let $R = \sqrt{X_1^2 + X_2^2}, \Theta = \arctan(X_1 / X_2)$. Then*

-
$$f_{R,\Theta}(r, \theta) r = f_{X_1,X_2}(r \cos(\theta), r \sin(\theta)).$$

-
$$f_{X_1,X_2}(x_1, x_2) dx_1 dx_2 = f_{R,\Theta}(r(x_1, x_2), \theta(x_1, x_2)) r dr d\theta.$$

- $$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_1dx_2 = \int_0^{\infty}\int_0^{2\pi} f_{R,\Theta}(r,\theta)rdrd\theta = 1.$$
- *The support for* (R,Θ) *is*
$$\{(0,+\infty)\times[0,2\pi]\}$$

Proof. Note that
$$J = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{vmatrix} \implies |J| = r.$$

Therefore
$$f_{X_1,X_2}(x_1,x_2) = f_{R,\Theta}(r(x_1,x_2),\theta(x_1,x_2))r.$$

□

Example 12.4.2. Let $X \sim N(0,1)$ and $Y \sim N(0,1)$. Let $R = \sqrt{X^2+Y^2}, \Theta = \arctan(X/Y)$.

Then
$$f_{X,Y}(x,y) = \frac{1}{2\pi}\exp(-\frac{x^2+y^2}{2}).$$

$$f_{R,\Theta}(r,\theta) = f_{X,Y}(r\cos(\theta), r\sin(\theta)) = \frac{1}{2\pi}\exp(-\frac{r^2}{2}).$$

$$\int_0^{\infty}\int_0^{2\pi} f_{R,\Theta}(r,\theta)rdrd\theta = \int_0^{\infty}\int_0^{2\pi}\frac{1}{2\pi}\exp(-\frac{r^2}{2})rdrd\theta = 1.$$

Lemma 12.4.11 (convolution formula). *[6, p. 95] Let X_1 and X_2 be continuous random variables with joint pdf $f_{X_1,X_2}(x_1,x_2)$ with $\mathcal{D} = \mathbb{R}^2$. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_2$. Then*

- $f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}(y_1-y_2,y_2)$.
- *The pdf of Y_1 is given by*
$$f_{Y_1}(y) = \int_{-\infty}^{\infty} f_{X_1,X_2}(y-y_2,y_2)dydy_2.$$

Proof. It is easy to see $|J| = 1$. □

12.5 Expectation

12.5.1 Failure of elementary approach

Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$, if Ω is finite, we can simply define the expectation as

$$E[X] = \sum_{\omega \in \Omega} P(\omega) X(\omega)$$

However, if Ω is countably infinite, we can still list a sequence of $\omega_1, \omega_2, ...$ such that

$$E[X] = \sum_{i=1}^{\infty} P(\omega) X(\omega_i)$$

However, if Ω is uncountably infinite, then **uncountable** summation is not defined, and we need Lebesgue integral.

12.5.2 Formal definitions

Definition 12.5.1 (Lebesgue integral). *[5, p. 15] Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$, assume $0 \leq X(\omega) \leq \infty$ for every $\omega \in \Omega$, and let $\Pi : 0 = y_0 < y_1 < ...$ be a partition on the range of $X(\omega)$. For each subinterval $[y_k, y_{k+1}]$, we set*

$$A_k = \{\omega \in \Omega : y_k \leq X(\omega) \leq y_{k+1}\} = X^{-1}([y_k, y_{k+1}])$$

We define the lower Lebesgue sum to be

$$LS_{\Pi}^{-} = \sum_{k=1}^{\infty} y_k P(A_k)$$

We further define the limit

$$\lim_{\|\Pi\| \to 0} LS_{\Pi}^{-} = \int_{\Omega} X(\omega) dP(\omega)$$

Remark 12.5.1.

- Because X is measurable maps, its inverse image of any Borel set in $\mathbb{R}$ is measurable, i.e., $P(A)$ has value.
- For $X(\omega)$ that takes positive and negative part, we can simply decompose into two parts and use the linearity.

Definition 12.5.2 (expectation). *Let X be a random variable on a probability space $(\Omega, \mathbb{F}, P)$. The expectation of X is defined to be*

$$EX = \int_\Omega X(\omega) dP(\omega)$$

This definition makes sense if X is integrable, i.e., if

$$E|X| = \int_\Omega \left|X(\omega)\right| dP(\omega) < \infty$$

Remark 12.5.2. Note that the integral is defined using Lebesgue integral, and based on this definition we can recover the elementary definitions.

- If X takes only finitely many $x_0, x_1, ..., x_n$, but Ω is uncountable, then

$$EX = \sum_{x_k} x_k P(X = x_k)$$

  and $P(X = x_k)$ is the probability measure of all the subsets $X^{-1}(\{x_k\})$
- In particular, if Ω is finite, then

$$EX = \sum_{\omega \in \Omega} X(\omega) P(\omega)$$

Example 12.5.1. Let $\Omega = [0,1]$, and let P be the Lebesgue measure on $[0,1]$. Consider $X(\omega) = 1$, if ω is irrational; 0 otherwise. Then $E[X] = 1P(\omega \in [0,1] : \omega \text{ is irrational}) + 0P(\omega \in [0,1] : \omega \text{ is rational}) = 1$ since $P(\omega \in [0,1] : \omega \text{ is irrational}) = 1 = 1$, $P(\omega \in [0,1] : \omega \text{ is rational}) = 0$

12.5.3 Properties of expectation

Lemma 12.5.1 (linearity of expectation). *Let X, Y be two random variables over the same probability space. Then*

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y]$$

$$E[cY] = cE[X]$$

Theorem 12.5.1 (law of total expectation). *Let X be a random variable, Let $A_1, ..., A_n \in \mathcal{F}$ be the partition of the sample space, then*

$$E[X] = \sum_{i=1}^{n} E[X|A_i]P(A_i)$$

In concise form, we have

$$E[E[X|Y]] = E[X]$$

where Y is the random variable defined on measure space $(\Omega, \sigma(A_1, ..., A_n))$.

Definition 12.5.3 (expectation of function of random variable). *[9] Let $h : \mathbb{R} \to \mathbb{R}, f \in \mathcal{C}$, and let $X : \Omega \to \mathbb{R}$ be a random variable with probability density $f(x)$. Then the expectation of $h(X) : \Omega \to \mathbb{R}$ is given as:*

$$E[h(X)] = \int_{-\infty}^{+\infty} h(x)f(x)dx$$

12.6 Variance and covariance

12.6.1 Basic properties

Definition 12.6.1 (variance, covariance). *The variance of random variable X is defined as*

$$Var[X] = E[(X - EX)^2]$$

The covariance of random variable X and Y is defined as

$$cov(X, Y) = E[(X - EX)(Y - EY)]$$

The covariance matrix $Cov(Z)$ of a random vector $Z = [Z_1, ..., Z_m]^T$ is defined as

$$Cov(Z)_{ij} = cov(Z_i, Z_j)$$

Lemma 12.6.1 (basic properties for random variables). *Let X and Y be random variables, let $a, b \in \mathbb{R}$*

- $Var[X] = E[X^2] - E[X]^2$
- $cov(X, Y) = E[XY] - E[X]E[Y]$
- $cov(\sum_i^m a_i X_i, \sum_j^n b_j Y_j) = \sum_i^m \sum_j^n a_i b_j cov(X_i, Y_j)$
- $Var[X + a] = Var[X]$
- $Var[aX] = a^2 Var[X]$
- $Var[aX + bY] = a^2 Var[X] + b^2 Var[Y] + 2abcov(X, Y)$
- $Var[aX - bY] = a^2 Var[X] + b^2 Var[Y] - 2abcov(X, Y)$
- *More generally,*

$$Var[\sum_{i=1}^n a_i X_i] = \sum_{i=1}^n a_i^2 Var[X_i] + 2\sum_{i=1}^n \sum_{j>1}^n a_i a_j cov(X_i, X_j) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j cov(X_i, X_j)$$

Proof. Straight forward from definitions. For (3), use linearity of expectation. □

Lemma 12.6.2 (basic properties for random vectors). *Let X be a random vector, let A, B be non-random matrices, we have*

- $Cov(AX) = ACov(X)A^T$
- $Cov(X + B) = Cov(X)$

Proof. Straight forward from definitions. □

Lemma 12.6.3 (variance of a function of a random variable). *Let X be a random variable taking value in $\mathcal{X}$ with pdf $f(x)$, let g be a continuous function, then*

$$Var[g(X)] = E[(g(X) - E[g(X)])^2]] = \int_{\mathcal{X}} (g(x) - E[g(x)])^2 f(x) dx$$

Proof. Note that $E[g(X)]$ is a constant. We can calculate $Var[g(X)]$ using the expectation of a function of a random variable definition [Definition 12.5.3]. □

12.6.2 Conditional variance

Theorem 12.6.1 (conditional variance identity). *[8, p. 193] For any two random variables X and Y,*

$$Var[X] = E[Var[X|Y]] + Var[E[X|Y]],$$

provided that the expectation exists.

Example 12.6.1. Suppose the random variable $Y \sim Binomial(n, X)$, where $X \sim Uniform(0, 1)$ and n is a given constant. Then we can calculate

$$E[Y] = E[E[Y|X]] = E[nX]$$

and

$$Var[Y] = Var[E[Y|X]] + E[Var[Y|X]] = Var[nX] + E[nX(1 - X)].$$

12.7 Characteristic function and Moment generating functions

12.7.1 Moment generating function

Definition 12.7.1 (moment generating function). *[8, p. 62] The moment generating function of a random variable X is given as*

$$M_X(t) = E[e^{tX}],$$

provided that the expectation exists for t in some neighborhood of 0.

Remark 12.7.1 (existence of moment generating function). If the expectation does not exist for some t in the neighborhood of 0, then moment generating function does not exist.

Lemma 12.7.1 (generating moments). *[8, p. 62] Let X be a random variable with moment generating function $M_X(t)$. Under the assumption of exchange expectation and differential is legitimate, for $n > 1$, then*

- $$E[X^n] = M_X^{(n)}(0) = \frac{d^n M_X(t)}{dt^n}|_{t=0}.$$
- $$M_X(t) = 1 + \sum_{n=1}^{\infty} \frac{M^{(n)}(0)}{n!} t^n = 1 + \sum_{n=1}^{\infty} \frac{E[X^n]}{n!} t^n.$$

Proof. (1)

$$\begin{aligned} M_X(t) &= \int e^t x f(x) dx \\ M_X^{(n)}(t) &= \int x^n e^t x f(x) dx \\ M_X^{(0)}(t) &= \int x^n f(x) dx \end{aligned}$$

(2) Use Taylor expansion. □

Theorem 12.7.1 (fundamental relationship between distribution and moment generating functions). *[8, p. 65] Let $F_X(x)$ and $F_Y(y)$ be two cdfs all of whose moments exist. We have*

- *If X and Y have bounded support, then $F_X(u) = F_Y(u)$ for all u if and only if*

$$E[X^r] = E[Y^r]$$

for all integers $r = 0, 1, 2, ...$
- ***(uniqueness)**If the moment generating functions exist and $M_X(t) = M_Y(t)$ for all t in some neighborhood of o, then $F_X(u) = F_Y(u)$ for all u.*

Remark 12.7.2 (non-uniqueness of moments).

- Two distinct random variables might have the same moments.[8, p. 64].
- The problem of uniqueness of moments does not occur if the random variables have bounded support.

Remark 12.7.3. The assumption that expectation and differentiation operands can be exchanged holds whenever the moment generating function exists in a neighborhood of zero, which will be the case for common distributions.[10]

Lemma 12.7.2 (basic properties). *[8, p. 67] Let X and Y be two independent random variables, then*

- $M_{X+Y}(t) = M_X(t)M_Y(t)$
- *If $Z = aX + b$, then $M_Z(t) = e^{bt}M_X(at)$*

Proof. (1) Let $Z = X + Y$, $f_Z(z) = \int f_X(z-y)f_Y(y)dy$, then

$$M_Z = \int e^{zt} f_Z(z)dz = \int e^{zt} \int f_X(z-y)f_Y(y)dy = \int e^{(z-y)t} f_X(z-y)dz \int e^{yt} f_Y(y)dy$$

let $w = z - y$, then $dw = dz - dy, dzdy = dydw((dy)^2 = 0)$, we have

$$\int e^{(z-y)t} f_X(z-y)dz \int e^{y} f_Y(y)dy = \int e^{wt} f_X(w)dw \int e^{yt} f_Y(y)dy = M_X(t)M_Y(t)$$

(2) From Corollary 12.4.1.1, $M_Z(t) = \int e^{zt} f_Z(z)dz = \int e^{axt+bt} f_X(x)dx = e^{bt}M_X(at)$ □

12.7.2 Characteristic function

Definition 12.7.2 (characteristic function). *Given a random variable X with probability measure P, its characteristic function is given as*

$$\psi_X(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itx} dP(x) = \int_{-\infty}^{\infty} e^{itx} f(x) dx.$$

Remark 12.7.4 (interpretation and existence).

- We can interpret the characteristic function as the Fourier transform of the density function $f(x)$.
- Because $\left|e^{itx} f(x)\right| \leq \left|f(x)\right|$ is L^1 integrable, then characteristic function always exists.

Lemma 12.7.3 (characteristic function as bijections). *Every distribution has a unique characteristic function; and to each characteristic function there corresponds a unique distribution of probability.*

Remark 12.7.5 (Moment generating functions vs characteristic functions).

- Characteristic function always exists, whereas moment generating function not necessarily exists.
- Characteristic function is useful when we want to develop theory for more general pdf.

Lemma 12.7.4 (recovering probability distribution from characteristic function). *Let $\psi_X(t)$ be the characteristic function of random variable X. Then we can obtain its probability density function via*

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi_X(t) \exp(-itx) dt$$

Proof. Use the property of Fourier transform [Lemma 6.7.1]:

$$\begin{aligned}
\frac{1}{2\pi} \int_{-\infty}^{\infty} \psi_X(t) \exp(-itx) dt &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{itx} dP(x) \\
&= \int_{-\infty}^{\infty} e^{itx'} f(x') \exp(-itx) dt dx' \\
&= \int_{-\infty}^{\infty} f(x') \delta(x - x') dx' = f(x)
\end{aligned}$$

□

12.7.3 Joint moment generating functions for random vectors

Definition 12.7.3 (joint moment generating function). *The joint moment generating function for a random vector $X = (X_1, ..., X_n)^T$ is defined as*

$$m_X(t) = E[\exp(t^T X)]$$

where $t \in \mathbb{R}^n, m_X(t) \in \mathbb{R}$, if the expectation exists in the neighborhood of the origin.

Lemma 12.7.5 (constructing joint moment generating function). *Let X be a K-dimensional random vector with a joint mgf $M_X(t)$, then we have*

- *If $X_1, X_2, ..., X_K$ are mutually independent of each other, then $M_X(t) = M_{X_1}(t_1)...M_{X_K}(t_K)$*
- *Let A be a matrix and b a vector, then $Z = AX + b$ has joint mgf given as*

$$M_Z(t) = e^{t^T b} M_X(A^T t)$$

Proof. Directly from definitions. □

Lemma 12.7.6 (cross moment generation). *Let X be a K dimensional random vector possessing a joint mgf $M_X(t)$, then*

$$\mu_X(n_1, n_2, ..., n_K) = E[X_1^{n_1} X_2^{n_2} ... X_K^{n_K}]$$

is given by

$$\mu_X(n_1, n_2, ..., n_K) = \frac{\partial^{n_1+...+n_K} M_X(t_1, ..., t_K)}{\partial t_1^{n_1} ... \partial t_K^{n_K}}|_{t=0}$$

Remark 12.7.6 (some applications). With joint mgf, we can evaluate the mean and covariance easily. For example, $E[X_1]$ can be obtained by setting $n_1 = 1, n_2 = 0, n_K = 0$. $E[X_i X_j]$ can be obtained by setting $n_i = n_j = 1$.

12.7.4 Probability generating function

Definition 12.7.4 (sequence generating function). *[11, p. 148] Given a real-valued sequence $a = \{a_1, a_2, ...\}$. The generating function G of the sequence is*

$$G(s) = \sum_{i=0}^{\infty} a_i s^i$$

for $s \in \mathbb{R}$ such that the sum converges.

Definition 12.7.5 (convolution of real sequence). *The convolution of the real sequences $a = \{a_i, i \geq 0\}$ and $b = \{b_i, i \geq 0\}$ is the sequence $c = \{c_i, i \geq 0\}$ defined by*

$$c_n = a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0.$$

Lemma 12.7.7 (convolution theorem of real sequence). *[11, p. 150] If sequences $\{a_n\}$ and $\{b_n\}$ have generating function G_a and G_b respectively, then*

$$c = a * b, G_c(s) = G_a(s) G_b(s).$$

Proof.

$$G_c(s) = \sum_{n=0}^{\infty} c_n s^n = \sum_{n=0}^{\infty} (\sum_{i=0}^{n} a_i b_{n-i}) s^n = \sum_{i=0}^{\infty} a_i s^i \sum_{n=i}^{\infty} b_{n-i} s^{n-i} = G_a(s) G_b(s).c$$

□

Lemma 12.7.8 (term-by-term operation property). *Consider a sequence generating function $G(s)$. If s^* is the convergence radius, then*

- *$\sum_{i=0}^{\infty} a_i s^i$ is uniformly convergent within $|s| < s^*$.*
- *$\sum_{i=0}^{\infty} a_i s^i$ can be differentiated and integrated term-by-term within $|s| < s^*$.*

Proof. Note that the generating function is the power sequence. See Theorem 3.5.2. □

Definition 12.7.6 (probability generating function). *[11, p. 150] The probability generating function of the random variable X is defined to be the generating function $G(s) = E[s^X]$ of its probability mass function.*

Example 12.7.1.

- Bernoulli variable X.

$$G(s) = E[s^X] = (1-p) + ps.$$

- Binomial distribution X with parameter n and p.

$$G(s) = E[s^X] = ((1-p) + ps)^n.$$

- Poisson distribution $Poisson(\lambda)$ random variable X:

$$G(s) = E[s^X] = \sum_{k=0}^{\infty} s^k \frac{\lambda^k}{k!} e^{-\lambda} = e^{s\lambda} e^{-\lambda} = e^{\lambda(s-1)}.$$

Lemma 12.7.9. *[11, p. 151] If X has generating function $G(s)$ then*

- $E[X] = G'(1)$.
- $E[X(X-1)\cdots(X-k+1)] = G^{(k)}(1)$.

Proof. (1) Note that $G(s) = E[s^X]$, $G'(s) = E[Xs^{X-1}], G'(1) = E[X]$.(use term-by-term differentiation property Lemma 12.7.8.) (2) Same as (1). □

Lemma 12.7.10. *[11, p. 153] If X and Y are independent, then*

$$G_{X+Y}(s) = G_X(s) G_Y(s)$$

Proof. Use Lemma 12.7.7. □

12.7.5 Cumulants

Definition 12.7.7 (cumulant-generating function, cumulant).

- *The **cumulant-generating function** $K(t)$ of a random variable X is defined by*

$$K(t) = \ln E[\exp(tX)] = \ln M_X(t),$$

  where $M_X(t)$ is the moment generating function of X.
- *The **cumulants** κ_n are obtained via*

$$\kappa_n = K^{(n)}(0),$$

such that

$$K(t) = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!}.$$

Lemma 12.7.11 (connections between cumulants and moments). *Let $\mu_i, i = 1, 2, ...$ denote the central moments,i.e. $\mu_i = E[(X - E[X])^i]$ of a distribution of a random variable X.Let $m_i, i = 1, 2, ...$ denote the cumulants of the same distribution. Let $\kappa_i, i = 1, 2, ...$ denote the cumulants of the same distribution. Assume the existence of moment generating function. Then*

- $$\ln(1 + \sum_{n=1}^{\infty} \frac{m_n}{n!} t^n) = \sum_{n=1}^{\infty} \frac{\kappa_n}{n!} t^n$$
- *explicitly, we have*

$$\begin{aligned}
\kappa_1 &= m_1 \\
\kappa_2 &= m_2 - m_1^2 = \mu_2 \\
\kappa_3 &= \mu_3 \\
\kappa_4 &= \mu_4 - 3\mu_2^2 \\
\kappa_5 &= \mu_5 - 10\mu_3\mu_2.
\end{aligned}$$

Proof. (1) Based on the definition,

$$\sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!} = K(t) = \ln E[\exp(tX)] = \ln M_X(t) = \ln(1 + \sum_{n=1}^{\infty} \frac{m_n}{n!} t^n),$$

where we use the properties of moment generating functions [Lemma 12.7.1]. (2) Use Taylor expansion for $\ln(1 + x)$ [Lemma 3.6.4] given by

$$\ln(1 + x) = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + ...$$

and then match the coefficients for t^n. □

Example 12.7.2. Consider a Gaussian distribution given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

Then

- the cumulant generating function is given by

$$K(t) = \ln(e^{\mu} e^{\sigma^2 \frac{t^2}{2}}) = \mu t + \frac{\sigma^2 t^2}{2}.$$

- the cumulants are given by

$$\kappa_1 = \mu, \kappa_2 = \sigma^2, \kappa_n = 0, \forall n > 2.$$

12.8 Conditional expectation

12.8.1 General intuitions & comments

Consider a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$ and a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$ ($\mathcal{G}$ is a σ-algebra and $\mathcal{G} \subset \mathcal{F}$). We have the following situations:[5]

1. If X is independent of $\mathcal{G}$, then the information in $\mathcal{G}$ provides no help in determining the value X. In this case, $E[X|\mathcal{G}] = E[X]$.
2. If X is $\mathcal{G}$ measurable, then the information in $\mathcal{G}$ can fully determine X. In this case, $E[X|\mathcal{G}] = X$.
3. In the intermediate case, we can use information in $\mathcal{G}$ to estimate but not precisely evaluate X. The *conditional expectation* of X given $\mathcal{G}$ is such an estimate.
4. If $\mathcal{G}$ is the trivial σ algebra $\{\emptyset, \Omega\}$, then $\mathcal{G}$ barely contains any information: $E[X|\mathcal{G}] = E[X]$.

Another understanding in terms of random variables are: $E[X|Y]$ is the function of Y that bests approximates X. We consider a extreme case. Suppose that X is itself a function of Y, then the function of Y that best approximates X is X itself, i.e., $E[g(Y)|Y] = X = g(Y)$; If X is independent of Y, then the best estimate we can give is $E[X|Y] = E[X]$.

As a summary, we have

Definition 12.8.1 (conditional expectation as least-squared-best predictor). *[12] If $E[X^2] < \infty$, then the conditional expectation $Y = E[X|\mathcal{G}]$ is a version of the orthogonal projection of X onto the space $L^2(\Omega, \mathcal{G}, P)$. Hence, Y is the lease-squared-best $\mathcal{G}$-measurable predictor of X: among all $\mathcal{G}$-measurable functions, Y minimizes*

$$E[(Y - X)^2].$$

Remark 12.8.1. Note that the discussion on the existence and uniqueness of such Y can be found at [12][13, p. 28].

12.8.2 Formal definitions

Definition 12.8.2 (sub σ algebra). *Let X be a set and let $\mathcal{F}, \mathcal{G}$ be two σ algebras on X. then $\mathcal{G}$ is said to be sub-σ algebra of $\mathcal{F}$ if $\mathcal{G} \subseteq \mathcal{F}$.*

Definition 12.8.3 (conditional expectation as a random variable). *[5, p. 68] Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{G}$ be a **sub-σ algebra** of $\mathcal{F}$, and let X be a random variable that is either non-negative or integrable. The conditional expectation of X given $\mathcal{G}$, denoted $E[X|\mathcal{G}]$ is a **random variable** that satisfies:*

1. *(**measurability**) $E[X|\mathcal{G}]$ is $\mathcal{G}$ measurable*
2. *(partial averaging): For any element A in $\mathcal{G}$,*

$$\int_A E[X|\mathcal{G}](\omega)dP(\omega) = \int_A X(\omega)dP(\omega).$$

In particular,

- *if $\mathcal{G} = \mathcal{F}$, then $E[X|\mathcal{G}] = X$.*
- *If $G = \{\emptyset, \Omega\}$, then $E[X|\mathcal{G}] = E[X]$.*

[a]

[a] The meaning of X is $\mathcal{G}$ measurable can be understood as $\sigma(X) \subseteq \mathcal{G}$.

Remark 12.8.2.

- the filtration $\mathcal{G}$ in $E[X|\mathcal{G}]$ has to be $\mathcal{G} \subseteq \mathcal{F}$, otherwise P is defined for some elements in cG.
- the Partial averaging property reflects the **consistence** requirement between the new random variable $E[X|\mathcal{G}]$ and the old random variable X.
- If $\mathcal{G}$ is the σ algebra generated by some other random variable W, then we generally write $E[X|W]$ instead of $E[X|\sigma(W)]$.
- if $G = \{\emptyset, \Omega\}$, then the only $\mathcal{G}$- measurable function is a constant function. Among all the constant functions, the function that satisfies the partial averaging property is the expectation.

Note 12.8.1 (interpreting partial averaging property in partition set). Consider the case where $\mathcal{G}$ is countable. Let $\mathcal{P}$ be the smallest partition set of $\mathcal{G}$. Then the random variable $E[X|\mathcal{G}]$ can only take countable many values. In particular, the partial averaging property implies

$$E[X|\mathcal{G}](A_i) = \int_A X(\omega)dP(\omega) \forall A_i \in \mathcal{P}.$$

That is, $E[X|\mathcal{G}]$ can be viewed as a mapping from Ω to $\mathbb{R}$ that has been **coarsened via local averaging**.

Note 12.8.2 (Generalization on expectation). When we talk about expectation, there are two items we should consider: which measure the expectation is taken with respect to and which filtration the expectation is taken with respect to.

- We can view expectation as a special case of conditional expectation: for example
$$E[X] = E[X|\mathcal{G}], \mathcal{G} = \{\emptyset, \Omega\}.$$
- Conditional expectations with respect to different measure can equal if the two measures agree on the filtration. For example,
$$E_P[X|\mathcal{G}] = E_Q[X|\mathcal{G}],$$
if $P(A) = Q(A), \forall A \in \mathcal{G}$.

12.8.3 Different versions of conditional expectation

Remark 12.8.3. For different versions of conditional expectation, see [13, p. 17] for details.

12.8.3.1 *Conditioning on an event*

Definition 12.8.4. *For any integrable random variable η and any event $B \in \mathcal{F}$ such that $P(B) \neq 0$, the conditional expectation given B is defined as*
$$E[\eta|B] = \frac{\int_B \eta dP}{\int_B dP} = \frac{1}{P(B)} \int_B \eta dP$$

12.8.3.2 *Conditioning on a discrete random variable as a new random variable*

Definition 12.8.5. *Let X be an integrable random variable, let Y be a discrete random variable. Then the conditioning expectation of X given Y is defined to be a random variable $E[X|Y]$ such that*
$$E[X|Y](\omega) = E[X|\{Y(\omega) = y_i\}]$$

Lemma 12.8.1. *If X is an integrable random variable, and Y is a discrete random variable, then*

- *$E[X|Y]$ is $\sigma(Y)$-measurable*

- *For any $A \in \sigma(Y)$:*
$$\int_A E[X|Y]dP = \int_A XdP$$

Proof. When Y is a discrete random variable, $E[X|Y]$ can only take discrete values. For any Borel set on $\mathbb{R}$, we find the inverse image $B \in \sigma(Y)$. Therefore it is measurable. (2) directly form partial averging property of conditional expectation. □

12.8.3.3 *Condition on random variable vs. event vs σ algebra*

- Conditional expectations for discrete random variables, such as $E[X|Y = 2], E[X|Y = 5]$ are numbers. These are examples of condition on events. $E[X|Y]$ can be interpreted as $E[X|Y = y]$, a function depends on y.
- When we write $E[X|Y]$, we should interpret as conditioning on the σ algebra generated by Y.

12.8.4 Properties

For a comprehensive treament, see [13, p. 70]. Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$,let X, Y be integrable random variables. We have:

12.8.4.1 *Linearity*

Lemma 12.8.2 (linearity of conditional expectation). *[5, p. 69] Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$,let X, Y be integrable random variables. We have:*
$$E[c_1X + c_2Y|\mathcal{G}] = c_1E[X|\mathcal{G}] + c_2E[Y|\mathcal{G}].$$

Proof. (1) First, $c_1E[X|\mathcal{G}]$ is $\mathcal{G}$ measurable, $c_2E[Y|\mathcal{G}]$ is $\mathcal{G}$ measurable, therefore, $E[c_1X + c_2Y|\mathcal{G}] = c_1E[X|\mathcal{G}] + c_2E[Y|\mathcal{G}]$ is $\mathcal{G}$ measurable [Lemma 12.3.2]. (2) For every $A \in \mathcal{G}$,

$$\begin{aligned}
&\int_A (c_1E[X|\mathcal{G}](\omega) + c_2E[Y|\mathcal{G}](\omega))dP(\omega) \\
=&\int_A (c_1E[X|\mathcal{G}](\omega) + c_2E[Y|\mathcal{G}](\omega))dP(\omega) \\
=&c_1\int_A (E[X|\mathcal{G}](\omega)dP(\omega) + c_2\int_A E[Y|\mathcal{G}](\omega))dP(\omega) \\
=&c_1\int_A X(\omega)dP(\omega) + c_2\int_A Y(\omega)dP(\omega) \\
=&\int_A c_1X(\omega) + c_2Y(\omega)dP(\omega)
\end{aligned}$$

that is $E[c_1X + c_2Y|\mathcal{G}]$ satisfies the partial averaging property. □

12.8.4.2 *Taking out what is known*

Lemma 12.8.3. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Let X, Y be integrable random variables. If XY is integrable, X is $\mathcal{G}$-measurable, then*

- $$E[XY|\mathcal{G}] = XE[Y|\mathcal{G}], E[g(X)Y|\mathcal{G}] = g(X)E[Y|\mathcal{G}].$$
- $$E[X|\mathcal{G}] = X, E[X|X] = X, E[g(X)|X] = g(X)$$

Proof. Note that from Lemma 12.3.2, $g(X), XY, g(X)Y$ are all $\mathcal{F}$ measurable random variables. □

12.8.4.3 *Law of iterated expectations*

Lemma 12.8.4 (iterated conditioning). *If $\mathcal{H}, \mathcal{G}$ are both σ algebra on Ω, and $\mathcal{G} \subset \mathcal{H}$ (in some sense $\mathcal{G}$ has less information), then for random variable X, we have*

$$E[E[X|\mathcal{H}]|\mathcal{G}] = E[X|\mathcal{G}]$$

$$E[E[X|\mathcal{G}]|\mathcal{H}] = E[X|\mathcal{G}].$$

In particular,

$$E[E[X|\mathcal{G}]] = E[X],$$

or equivalently, in terms of conditioning on random variables, we have

$$E[E[X|Y]] = E[X].$$

Proof. (1)(a) First $E[X|\mathcal{G}]$ is $\mathcal{G}$-measurable. (b) For any $A \in \mathcal{G} \subseteq H$, we have

$$\begin{aligned}
&\int_A E[E[X|\mathcal{H}]|\mathcal{G}](\omega)dP(\omega) \\
=&\int_A E[X|\mathcal{H}](\omega)dP(\omega) \\
=&\int_A X(\omega)dP(\omega) \\
=&\int_A E[X|\mathcal{G}](\omega)dP(\omega)
\end{aligned}$$

(2) Note that the random variable $E[X|\mathcal{G}]$ is $\mathcal{G}$-measurable therefore $\mathcal{H}$-measurable.

□

Example 12.8.1. Let B_t be a Brownian motion, let $\mathcal{F}_t = \sigma(B_s, s \leq t)$ be the filtration, then

$$E[B_s|\mathcal{F}_t, t > s] = B_s, E[B_s|\mathcal{F}_t, t < s] = E[B_t + (B_s - B_t)|\mathcal{F}_t] = B_t$$

12.8.4.4 *Conditioning on independent random variable/σ algebra*

Lemma 12.8.5. *[5, p. 70]Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Let X, Y be integrable random variables. Let f be Borel measurable function and $f(X)$ be integrable.*

- *If $\sigma(X)$ and $\mathcal{G}$ are independent, then*

$$E[X|\mathcal{G}] = E[X], E[g(X)|\mathcal{G}] = E[g(X)].$$

- *If X and Y are independent, then*

$$E[X|Y] = E[X|\sigma(Y)] = E[X].$$

Proof. (1)(a) $E[X]$ is a constant, therefore is $\mathcal{G}$ measurable. (b)(informal) Consider the special case where $X = \mathbf{1}_B$, where $B \in \mathcal{F}$ but B is independent of $\mathcal{G}$. Then

$$\int_A X(\omega)dP(\omega) = P(A \cap B) = P(A)P(B) = E[X]P(A) = E[X]\int_A dP(\omega) = \int_A E[X]dP(\omega).$$

Since X can be represented by the sum of indicator function, such relation can hold when X is an arbitrary random variable.(See reference for more details). □

12.8.4.5 *Least Square minimizing property*

Lemma 12.8.6 (least square minimizing property of conditional expectation). *Let $Y \in \mathcal{L}_2(\Omega, \mathcal{G}, P)$ and $\mathcal{F}$ be a sub-σ of $\mathcal{G}$, then*

$$E[(Y - E[Y|\mathcal{F}])^2] = \min\{E[(Y - Z)^2], \forall Z \in \mathcal{L}_2(\Omega, \mathcal{F}, P)\}$$

Proof. For any $Z \in \mathcal{L}_2(\Omega, \mathcal{F}, P)$, we have

$$\begin{aligned}
&E[(Y - E[Y|\mathcal{F}])^2] \\
&= E[(Y - Z + Z - E[Y|\mathcal{F}])^2] \\
&= E[(Y - Z)^2] + E[(Z - E[Y|\mathcal{F}])^2] + 2E[(Y - Z)(Z - E[Y|\mathcal{F}])] \\
&= E[(Y - Z)^2] + E[(Z - E[Y|\mathcal{F}])^2] + 2E[E[(Y - Z)(Z - E[Y|\mathcal{F}])|\mathcal{F}]] \\
&= E[(Y - Z)^2] + E[(Z - E[Y|\mathcal{F}])^2] + 2E[E[(Y - Z)|\mathcal{F}](Z - E[Y|\mathcal{F}])] \\
&= E[(Y - Z)^2] + E[(Z - E[Y|\mathcal{F}])^2] - 2E[(Z - E[Y|\mathcal{F}])(Z - E[Y|\mathcal{F}])] \\
&= E[(Y - Z)^2] - E[(Z - E[Y|\mathcal{F}])^2] \leq E[(Y - Z)^2]
\end{aligned}$$

Note that we use $E[(Y - Z)(Z - E[Y|\mathcal{F}])|\mathcal{F}] = (Z - E[Y|\mathcal{F}])E[(Y - Z)|\mathcal{F}]$ since $(Z - E[Y|\mathcal{F}])$ is $\mathcal{F}$ measurable. □

12.9 The Hilbert space of random variables

12.9.1 Definitions

Definition 12.9.1. *The vector space $L^2(\Omega, \mathcal{F}, P)$ of real-valued random variables on $(\Omega, \mathcal{F}, P)$ can be defined as the Hilbert space of random variables with finite second moment. The inner product is then defined as*

$$\langle x, y \rangle = E[xy].$$

The norm of a random variable is

$$\|X\| = \sqrt{E[X^2]}.$$

Lemma 12.9.1 (correlation and orthogonality for zero mean random variables). *Let X and Y be two zero mean random variables in the Hilbert space $L^2(\Omega, \mathcal{F}, P)$. Then X and Y are uncorrelated if and only if they are orthogonal, i.e., $\langle X, Y \rangle = 0$.*

Proof. (1) If $\langle X, Y \rangle = 0$, then

$$E[XY] = E[X]E[Y] + Cov(X, Y) = 0 \implies Cov(X, Y) = 0$$

. (2) If $Cov(X, Y) = 0$, then

$$\langle X, Y \rangle = E[XY] = E[X]E[Y] + Cov(X, Y) = 0.$$

□

12.9.2 Subspaces, projections, and approximations

Theorem 12.9.1 (projection onto closed subspace , recap). *Let U be a closed subspace of L^2 and $X \in L^2$. Then the projection of X onto U is the vector/random variable $V \in U$ such that*

- $$\langle X - V, u \rangle = E[(X - V)u] = 0, \forall u \in U$$
- *V is unique;*
- *V is minimizer,i.e., $\|X - V\|^2 \leq \|X - u\|^2, \forall u \in U$.*

[a].

[a] Note that in a Hilbert space(also a normed linear space), any finite-dimensional subspace is closed [Theorem 6.2.1]

Proof. See the projection theorem [Theorem 6.3.5] guarantees the existence of solution. □

Lemma 12.9.2 (projection onto the subspace of constant random variables).

- *Let real-valued random variable $X \in L^2$, we define the root mean square error function by*
$$d_2(X,t) = \|X - t\|_2 = \sqrt{E[(X-t)^2]}, t \in \mathbb{R}$$
then $d_2(X,t)$ is minimized when $t = E[X]$ and that the minimum value is $\sqrt{Var[X]}$.
- *Let real-valued random variable $X \in L^2$, we define the 1d subspace $W = \{a : a \in \mathbb{R}\}$(the subspace spanned by constant random variable 1). Then the projection of X onto W is $E[X]$.*

Proof. (1)directly minimize with respect t. (2) We can see that the orthogonality condition implies that
$$0 = \langle X - a, b \rangle = E[(X-a)b] = 0, \forall b \in \mathbb{R},$$
which gives $a = E[X]$. □

Theorem 12.9.2 (best linear predictor for random variables).

- *Given $X, Y \in L^2$, the best linear predictor for Y given X is to find a projection onto the subspace $W = \{a + bX : a \in \mathbb{R}, b \in \mathbb{R}\}$(the subspace spanned by random variable 1 and X), given as*
$$L(Y|X) = E[Y] + \frac{Cov(X,Y)}{Var(X)}(X - E[X])$$
and the variance/mean square error for the prediction is
$$Var(Y - L(Y|X)) = Var(Y) - \frac{Cov(X,Y)^2}{Var(X)}.$$
- *Given $X_1, X_2, ..., X_n, Y \in L^2$, the best linear predictor for Y given $X_1, X_2, ..., X_n$ is*
$$L(Y|X) = E[Y] + \sum_{i=1}^{n}(X_i - E[X_i])[\sum_{j=1}^{n}(\Sigma_{XX}^{-1})_{ij}Cov(X_j, Y)],;$$

or in vector form

$$L(Y|X) = E[Y] + (X - E[X])^T \beta,$$

where $X = (X_1, X_2, ..., X_n)^T, \beta = \Sigma_{XX}^{-1}\Sigma_{XY}$. *In particular, if* $Cov(X_i X_j) = Var[X_i]\delta_{ij}$, *then*

$$L(Y|X) = E[Y] + \sum_{i=1}^{n} \frac{Cov(X_i, Y)}{Var(X_i)}(X_i - E[X_i]).$$

- *The estimation error is given by*

$$E[(Y - L(Y|X))^2] = Var[Y] - \Sigma_{XY}^T(\Sigma_{XX}^{-1})\Sigma_{XY}.$$

- *The single coefficient associated with* X_i *is given by*

$$\beta_i = \frac{Cov(Y - L(Y|X_{-i}), X_i - L(X_i|X_{-i})}{Var[X_i - L(X_i|X_{-i})]},$$

where X_{-i} *denotes the subspace associated with span* $\{1, X_1, ..., X_{i-1}, X_{i+1}, ..., X_n\}$.
- *un-correlation of the residual and* X*:*

$$Cov(Y - L(Y|X), X) = 0.$$

Proof. (1) To verify that $L(Y|X)$ is the projection, we only need to verify the orthogonality conditions [Theorem 6.3.5]:

$$\langle Y - L(Y|X), X \rangle = 0, \langle Y - L(Y|X), 1 \rangle = 0.$$

We have

$$\begin{aligned}
\langle Y - L(Y|X), X \rangle &= E[(Y - L(Y|X))X] \\
&= E[(Y - E[Y] - \frac{Cov(X, Y)}{Var(X)}(X - E[X]))X] \\
&= E[(Y - E[Y])X] - E[\frac{Cov(X, Y)}{Var(X)}(X - E[X]))X] \\
&= Cov(X, Y) - Cov(X, Y) \\
&= 0
\end{aligned}$$

where we used the fact that $E[X(X - E[X])] = Var[X]$. For another,

$$\begin{aligned}\langle Y - L(Y|X), 1 \rangle &= E[(Y - L(Y|X))] \\ &= E[(Y - E[Y] - \frac{Cov(X,Y)}{Var(X)}(X - E[X]))] \\ &= 0 - 0 \\ &= 0.\end{aligned}$$

The variance is given by

$$\begin{aligned}&Var[Y - L(Y|X)] \\ &= E[(Y - L(Y|X))^2] \\ &= E[(Y - E[Y] - \frac{Cov(X,Y)}{Var(X)}(X - E[X]))(Y - E[Y] - \frac{Cov(X,Y)}{Var(X)}(X - E[X]))] \\ &= E[(Y - E[Y])^2] + \frac{Cov(X,Y)^2}{Var(X)} - 2E[(Y - EY)(\frac{Cov(X,Y)}{Var(X)}(X - E[X]))] \\ &= E[(Y - E[Y])^2] + \frac{Cov(X,Y)^2}{Var(X)} - 2\frac{Cov(X,Y)^2}{Var(X)} \\ &= E[(Y - E[Y])^2] - \frac{Cov(X,Y)^2}{Var(X)}\end{aligned}$$

(2)

We can obtain the vector form via the optimization

$$\min f = E[(Y - \beta_0 - \beta^T X)^2]$$

over β_0, β_1, we have

$$f(\beta_0, \beta_1) = E[(Y^2 + \beta_0^2 + (\beta^T X)^2 + 2\beta_0\beta^T X - 2\beta_0 Y - 2Y\beta^T X)]$$

The first order condition on β_0 gives that

$$\beta_0 = E[Y] - \beta^T E[X];$$

Plug in β_0 and the first order condition on β_1 gives that

$$\begin{aligned}&f(\beta_0, \beta_1) = E[((Y - EY)^2 - 2\beta^T(X - EX)(Y - EY) + \beta^T E[(X - EX)(X - EX)^T]\beta \\ \implies &\partial f/\partial\beta = -2E[(X - EX)(Y - EY)] + 2E[(X - EX)(X - EX)^T]\beta = 0 \\ \implies &\beta = (E[(X - EX)(X - EX)^T])^{-1}E[(X - EX)(Y - EY)] = (\Sigma_{XX}^{-1})\Sigma_{XY}\end{aligned}$$

Note that the problem has semi-positive definite Hessian we are sure that the minimizer exists.

From the Hilbert space projection point of view, we can also verify the orthogonality conditions [Theorem 6.3.5]:

$$\langle Y - L(Y|X), X_k \rangle = 0, k = 1, 2, ..., n.$$

We have

$$\begin{aligned}
\langle Y - L(Y|X), X_k \rangle &= E[(Y - E[Y] + \sum_{i=1}^{n}(X_i - E[X_i])[\sum_{j=1}^{n}(\Sigma_{XX}^{-1})_{ij}Cov(X_j, Y)]))X_k] \\
&= E[(Y - E[Y])X_k] - E[(\sum_{i=1}^{n}(X_i - E[X_i])[\sum_{j=1}^{n}(\Sigma_{XX}^{-1})_{ij}Cov(X_j, Y)])X_k] \\
&= Cov(X_k, Y) - Cov(X_j, Y)\delta_{jk} \\
&= 0
\end{aligned}$$

where we used the fact that

$$\sum_{i=1}^{n}(X_i - E[X_i]) \sum_{j=1}^{n}(\Sigma_{XX}^{-1})_{ij}X_k = \delta_{jk}.$$

Note that for an invertible matrix A, $\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}A_{jk}^{-1} = \delta_{ik}$.

(3) Note that $L(Y|X)$ is unbiased because of the orthogonality condition

$$\langle Y - L(Y|X), 1 \rangle = 0 \implies E[(Y - L(Y|X))1] = E[Y] - E[L(Y|X)] = 0.$$

In the following we use the notation

$$\hat{Y} = L(Y|X), E[Y] = \mu_Y = E[L(Y|X)] = \mu_{\hat{Y}}.$$

We have

$$\begin{aligned}
E[(Y - \hat{Y})^2] &= E[((Y - \mu_Y) - (\hat{Y} - \mu_{\hat{Y}}))^2] \\
&= E[((Y - \mu_Y))^2] - 2E[(Y - \mu_Y)(\hat{Y} - \mu_{\hat{Y}})] + E[((\hat{Y} - \mu_{\hat{Y}}))^2] \\
&= Var[Y] - 2E[(Y - \mu_Y)(\hat{Y} - \mu_{\hat{Y}})] + Var[\hat{Y}] \\
&= Var[Y] - 2E[(Y - \mu_Y)(\beta^T X - \mu_{\hat{Y}})] + Var[\hat{Y}] \\
&= Var[Y] - 2Cov(Y, \beta^T X) + Var[\beta^T X] \\
&= Var[Y] - 2\beta^T Cov(Y, X) + \beta^T Cov(X, X)\beta \\
&= Var[Y] - 2\Sigma_{XY}^T(\Sigma_{XX}^{-1})Cov(Y, X) + \Sigma_{XY}^T(\Sigma_{XX}^{-1})Cov(X, X)(\Sigma_{XX}^{-1})\Sigma_{XY} \\
&= Var[Y] - \Sigma_{XY}^T(\Sigma_{XX}^{-1})\Sigma_{XY}
\end{aligned}$$

(4) Direct generalization from Hilbert space approximation theory [Theorem 6.4.4].

(5)

$$\begin{aligned} E[Y - L(Y|X), X - E[X]] &= E[Y - E[Y] - (X - E[X])^T\beta, X - E[X]] \\ &= Cov(X, Y) - Var[X]\beta \\ &= Cov(X, Y) - Cov(X, Y) \\ &= 0. \end{aligned}$$

□

Remark 12.9.1.

- The more correlated X and Y are, the more information X can provide to predict Y
- The more volatile X is, the less information X can provide.
- The magnitude of $\frac{Cov(X,Y)^2}{Var(X)}$ reflects the importance of X in prediction.

12.9.3 Connection to conditional expectation

Theorem 12.9.3 (conditional expectation with respect to a σ algebra as a projection).

- *Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, then the set*

$$U = \{X \in L^2 | X \text{ is measurable with respect to } \mathcal{G}\}$$

  is a subspace of L^2.
- *If $X \in L^2$, then $E[X|\mathcal{G}]$ is the projection of X onto the subspace U defined as*

$$U = \{X \in L^2 | X \text{ is measurable with respect to } \mathcal{G}\}.$$

Proof. (1) the zero element o is both $\mathcal{F}$ and $\mathcal{G}$ measurable. (2) If X, Y are $\mathcal{G}$ measurable, then $cX, X + Y$ are $\mathcal{G}$ measurable [Lemma 3.8.4]. (2) directly from the definition of conditional expectation [Definition 12.8.3]. □

Definition 12.9.2 (conditional expectation and projection). *The conditional expectation of $X \in L^2$ given $X_1, X_2, ..., X_n \in L^2$ is defined to be the projection of X onto*

*the closed subspace $M(X_1, X_2, ..., X_n)$ spanned by **all random variables of the form** $g(X_1, X_2, ..., X_n)$, where g is some measurable function $g : \mathbb{R}^n \to \mathbb{R}$, i.e.*

$$E[X|X_1, X_2, ..., X_n] = P_S[M(X_1, X_2, ..., X_n)].$$

Definition 12.9.3 (conditional expectation and projection onto a subspace, special case). *The conditional expectation of $X \in L^2$ given a closed subspace $S \subseteq L^2$, which contains the constant random variable 1, is defined to be the projection of X onto S,i.e.,*

$$E[X|S] = P_S[X].$$

Remark 12.9.2. Note that the subspace has to contains the constant random variable to make the definition and conditional expectation and projection match.

Remark 12.9.3.

- $$span(1, X_1, ..., X_n) \subseteq M(X_1, X_2, ..., X_n),$$
  therefore
  $$\left\|X - E[X|X_1, X_2, ..., X_n]\right\|^2 \leq \left\|X - E[X|span(1, X_1, X_2, ..., X_n)]\right\|^2.$$
- The definition of
  $$E[X|X_1, X_2, ..., X_n] = P_S[M(X_1, X_2, ..., X_n)]$$
  coincides with the usual definition of conditional expectation with respect to a σ algebra [Definition 12.8.3].
- The conditional expectation with respect to a subspace is not the general definition of conditional expectation.

Lemma 12.9.3 (conditional expectation and best predictor for multivariate normal random variables). *Let $(Y, X_1, X_2, ..., X_n), X = (X_1, X_2, ..., X_n)$ be a random vector with multivariate normal distribution with parameter*

$$\mu = [\mu_Y^T, \mu_X^T]^T, \Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}$$

Then

- *$Y|X_1, X_2, ..., X_n$ has the same distribution of*
  $$\hat{Y} = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X) + \epsilon,$$

conditioning on X and $\epsilon \sim N(0, \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})$.

- $$E[Y|X_1, X_2, ..., X_n] = E[Y|span(1, X_1, X_2, ..., X_n)] = P_{span(1,X_1,X_2,...,X_n)}[Y].$$

  That is, the best predicator(in terms of minimum variance) given $X_1, X_2, ..., X_n$ is the best linear predictor.

Proof. From Theorem 15.1.2, the martinal distribution is Gaussian given by

$$N(\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X), \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}).$$

Therefore, Y has the conditional expectation of

$$\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X),$$

and the conditional variance of

$$\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}.$$

□

12.10 Probability inequalities

12.10.1 Some common inequalities

Theorem 12.10.1 (General Chebychev's inequality). *Let X be a random variable and $g(x) \geq 0$. Then for any $r > 0$, we have:*

$$P(g(X) > r) \leq \frac{E[g(X)]}{r}.$$

Proof.

$$\begin{aligned} E[g(x)] &= \int_{\infty}^{\infty} g(x) f_X(x) dx \\ &\geq \int_{x:g(x)\geq r} g(x) f_X(x) dx \\ &\geq r \int_{x:g(x)\geq r} f_X(x) dx \\ &= rP(g(X) \geq r) \end{aligned}$$

□

Corollary 12.10.1.1 (Chebychev's inequality).

$$P(\frac{(X - EX)^2}{\sigma^2} \geq t^2) \leq \frac{1}{t^2} E[\frac{(X-\mu)^2}{\sigma^2}].$$

Or equivalently,

$$P(|X - EX| \geq t) \leq \frac{1}{t^2} Var[X], t \geq 0.$$

Proof. Let $g(X) = (X - \mu)^2/\sigma^2$ and use above theorem. □

Example 12.10.1. Let X be a random variable with mean $\mu = 4$ and standard deviation $\sigma = 1$. Then the probability that $X < 1$ or $X > 7$ is bounded by

$$P(|X - 4| > 3) \leq \frac{1^2}{3^2} = \frac{1}{9}.$$

Corollary 12.10.1.2. *Let $g : [0, \infty) \to [0, \infty)$ be a strictly increasing non-negative function, and set $h(x) = g(|x|)$ to obtain*

$$P(|X| \geq a) \leq \frac{E[g(|X|)]}{g(a)}$$

where $a > 0$.

Corollary 12.10.1.3 (Markov's inequality). *If $X \geq 0$, then*

$$P(X \geq r) \leq E[X]/r$$

Proof. Let $g(X) = X$ in the general Chebychev's inequality. □

Lemma 12.10.1 (Jensen's inequality). *For any random variable X, if $g(x)$ is a convex function then*

$$E[g(X)] \geq g(E[X]).$$

Proof. Note that for convex function

$$g(\sum_{i=1}^{n} w_i x_i) \leq \sum_{i=1}^{n} w_i g(x_i), \forall w_i \geq 0, \sum_{i=1}^{n} w_i = 1, i = 1, ..., n.$$

□

Example 12.10.2 (Jensen's inequality application). Use Jensen's inequality, it can be showed that

$$E[X]^2 \leq E[X^2]$$

with $g(x) = x^2$.

Theorem 12.10.2 (Holder's inequality). *[11, p. 319] If $p, q > 1$ and $1/p + 1/q = 1$, then*

$$E[|XY|] \leq (E[|X|^p])^{1/p} (E[|Y|^q])^{1/q}.$$

The equality holds when there exists real numbers $\alpha, \beta > 0$ such that $\alpha |X|^p = \beta |Y|^q$ almost everywhere.

Proof. Let $A = (\int |x|^p \, dP)^{1/p} = E[\![X^p|]\!])^{1/p}$ and $B = (\int |y|^q \, dP)^{1/q} = (E[\![Y^q|]\!])^{1/q}$. Then let $a = |X| / A$, $b = |Y| / B$, and then apply Young's inequality:

$$ab = |XY| / AB \leq \frac{|X|^p}{pA^p} + \frac{|Y|^q}{qA^q} = \frac{a^p}{p} + \frac{b^q}{q}$$

Integrate(Lebesgue) both sides use probability measure and notice that A, B are constant, $A^p = E[\![X^p|]\!]$, then

$$\frac{E[\![XY|]\!]}{(E[\![X^p|]\!])^{1/p}(E[\![Y^q|]\!])^{1/q}} \leq 1/p + 1/q = 1$$

□

Remark 12.10.1. Let $q = p = 2$, and we get the Cauchy-Schwarz inequality.

Theorem 12.10.3 (Minkowski's inequality). *[11, p. 319] If $p \geq 1$, then*

$$(E[\![X + Y|^p]\!])^{1/p} \leq (E[\![X|^p]\!])^{1/p} + (E[\![Y|^p]\!])^{1/p}$$

Proof. Because L^p space are normed vector space, we can prove this using triangle inequality. □

Theorem 12.10.4 (Cauchy-Schwarz inequality). *[2][8, p. 187][14]*

- *Let X and Y be random variables with $E[X^2] < \infty, E[Y^2] < \infty$. Then*

  $$\left|E[XY]\right| \leq \sqrt{E[X^2]E[Y^2]}.$$

  The equality holds when there exists real numbers $\alpha, \beta > 0$ such that $\alpha|X|^2 = \beta|Y|^2$ almost everywhere.
  Further more,

  $$(Cov(X, Y))^2 \leq Var[X] \cdot Var[Y].$$

- *Let X and Y be two p dimensional random vectors with bounded variance. Then*

  $$Var[Y] \geq Cov(Y, X)Var[X]^{-1}Cov(X, Y).$$

Proof. (1)(a) define inner product between two random variable as $\langle X, Y \rangle = \int xy\rho(x, y)dxdy$, since each random variable can be viewed as a functional. (b) Similarly we can use Holder's inequality. [Theorem 12.10.2]

A simple derivation: Since the covariance matrix of random vector (X, Y) much be positive semi-definite, we have

$$\big|Cov([XY])\big| = \begin{vmatrix} Var[X] & Cov(X,Y) \\ Cov(X,Y) & Var[Y] \end{vmatrix} \geq 0.$$

Expand the determinant, we have

$$Cov(X,Y)^2 \leq Var[X] \cdot Var[Y].$$

(2) See reference. □

Corollary 12.10.4.1 (bounds on correlations).

- *Let X and Y be two random variables with mean μ_X and μ_Y. Define correlation by*

$$\rho \triangleq \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]}\sqrt{E[(Y-\mu_Y)^2]}}.$$

  Then

$$\big|\rho\big| \leq 1$$

  .

- *Let $X_1, X_2, ..., X_n$ be the iid random sample of X. Let $Y_1, Y_2, ..., Y_n$ be the iid random sample of Y. Define sample correlation by*

$$\hat{\rho} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})(Y - \overline{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2}}.$$

  For each realizations of $X_1, ..., X_n, Y_1, ..., Y_n$, we have

$$\big|\hat{\rho}\big| \leq 1.$$

Proof. (1) From Cauchy-Schwartz inequality, we have

$$\big|\rho\big| = \frac{\big|E[(X-\mu_X)(Y-\mu_Y)]\big|}{\sqrt{E[(X-\mu_X)^2]}\sqrt{E[(Y-\mu_Y)^2]}} \leq 1.$$

(2) Suppose we have a realization of $x_1, x_2, ..., x_n, y_1, y_2, ..., y_n$, we can define a random variable X with probability $1/n$ in taking discrete values $x_1, x_2, ..., x_n$;similarly define a random variable Y. Then

$$\big|\hat{\rho}\big| = \big|\rho_{XY}\big| \leq 1.$$

□

Lemma 12.10.2 (Popoviciu's inequality for variance). *link* *Consider a random variable X with support on a finite interval $[m, M]$. Then*

its variance is bounded via

$$Var[X] = \frac{(M-m)^2}{4}.$$

- *the bound is tight and can be achieve by a discrete distribution of*

$$p(X) = \begin{cases} \frac{1}{2}, X = m \\ \frac{1}{2}, X = M \end{cases}$$

Proof. Define a function $g(t) = E[(X-t)^2]$. The derivative of g with respect to t is given by $g'(t) = -2E[X] + 2t = 0$. And the g achieves its minimum at $t = E[X]$(note that $g''(E[X]) > 0$) with minimum value $g(E[X]) = Var[X]$. Consider the special point $t = \frac{M+m}{2}$, we have

$$Var[X] = g(E[X]) \leq g(\frac{M+m}{2}) = E[(X - \frac{M+m}{2})^2].$$

Now our goal is to find an upper bound on $E[(X - \frac{M+m}{2})^2] = \frac{1}{4}E[((X-m) + (X-M))^2]$.

Since $X - m \geq 0, X - M \leq 0$, we have

$$\begin{aligned} (X-m)^2 + 2(X-m)(X-M) + (X-M)^2 &\leq (X-m)^2 - 2(X-m)(X-M) + (X-M)^2 \\ ((X-m) + (X-M))^2 &\leq ((X-m) - (X-M))^2 = (M-m)^2 \\ \implies \frac{1}{4}E[((X-m) + (X-M))^2] &\leq \frac{1}{4}E[((X-m) - (X-M))^2] = \frac{(M-m)^2}{4}. \end{aligned}$$

We therefore have

$$Var[X] = \frac{(M-m)^2}{4}.$$

□

Example 12.10.3. Consider a discrete random variable X with support on $[-1, 1]$, then the upper bound for its variance is given by

$$\frac{1}{4}(2)^2 = 1.$$

The bound can be achieve by a discrete distribution of

$$p(X) = \begin{cases} \frac{1}{2}, X = -1 \\ \frac{1}{2}, X = 1 \end{cases}$$

12.10.2 Chernoff bounds

Theorem 12.10.5 (Chernoff bounds). *[10]The Chernoff bound for a random variable X: for $t > 0$,*

$$P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq E[e^{tX}]/e^{ta}$$

minimize t, we have

$$P(X \geq a) \leq \min_{t>0} E[e^{tX}]/e^{ta}$$

similarly, for $t < 0$, we have

$$P(X \leq a) \leq \min_{t<0} E[e^{tX}]/e^{ta}$$

Proof. We use the Markov inequality [Corollary 12.10.1.3] □

12.11 Convergence of random variables

12.11.1 Different levels of equivalence among random variables

Given two random variables A and B defined on the same probability space $(\Omega, \mathcal{F}, P)$, we can have the following different levels of equivalence:

- We say A is identical to B if
$$A(\omega) = B(\omega), \forall \omega \in \Omega.$$
- We say A is almost surely identical to B if
$$P(\mathcal{N}) = 0, \mathcal{N} = \{\omega, A(\omega) \neq B(\omega)\}.$$
- We say A and B have the same distribution if
$$P(A < x) = P(B < x).$$
- We say A and B have the same moments upto K if
$$E[A^k] = E[B^k], k = 1, 2, ..., K.$$

12.11.2 Convergence almost surely

Definition 12.11.1 (convergence almost surely). *[11, p. 308]Let $\{X_n\}$ be a sequence of random variables. Then X_n converges to X almost surely if, for arbitrary $\delta > 0$ and for all $\omega \in \Omega$, we have:*
$$P(\lim_{n\to\infty} \left|X_n(\omega) - X(\omega)\right| < \delta) = 1;$$
or
$$X_n(\omega) \to X(\omega), \text{ as } n \to \infty, \forall \omega \in \Omega.$$

Remark 12.11.1 (interpretation).

- X_n converges to X almost surely if the functions $X_n(\omega)$ converges to $X(\omega)$ for all $\omega \in \Omega$ except perhaps for $s \in N, N \subset \Omega, P(N) = 0$. The probability measure o for the non-convergent point is the key point here.
- Note that if we view X_n as a type of function mapping, then the almost surely convergence says that X_n and X are the same (in the limit) when maps from sample space to $\mathbb{R}^n$.

Remark 12.11.2 (convergence almost surely vs. converge pointwise). If the partition the sample space Ω into two sets D and N such that $P(D) = 1$ and $P(N) = 0$. Then $X_1, X_2, ...$ converges to X almost surely is equivalently to $X_1, X_2, ...$ converges to X **pointwise** on the set of D.

Example 12.11.1. For example, let $\Omega = [0, 1]$, and $X_n(\omega) = \omega + \omega^n$ and $X(\omega) = \omega$. For every $s \in [0, 1)$, X_n converges to X; the non-convergent point 1 has measure of 0.

12.11.3 Convergence in probability

12.11.3.1 *Basics*

Definition 12.11.2 (convergence in probability). *[6]Let $\{X_n\}$ be a sequence of random variables and let X be a random variable defined on a sample space. We say that X_n converges in probability to X if, $\forall \epsilon > 0$,*

$$\lim_{n\to\infty} P(|X_n - X| >= \epsilon) = 0,$$

or equivalently

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1,$$

then we write

$$X_n \xrightarrow{P} X$$

Remark 12.11.3. Note that if the random variable X is degenerate, i.e., X has close to 1 but not 1 probability of taking a constant value a. Not 1 probability means that there will infinitely often $X_n \neq a$ as $n \to \infty$. The convergence in probability is NOT like the real sequence convergence in which when n is large enough, X_n will be arbitrarily closer to a, but in probability convergence, X_n might have small chances to **take value far from** a.

Lemma 12.11.1. *[15][8]Convergence almost surely will imply convergence in probability.*

Proof. Convergence almost surely says that given $\epsilon > 0$, there exist an N such that for all $n > N$, we have $\left|X_n(\omega) - X(\omega)\right| < \epsilon, \forall \omega \in A \in \mathcal{F}, P(A) \neq 0$. Therefore, $P(|X_n - X| < \epsilon) = 1, \forall n > N$, therefore, $\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$ □

Remark 12.11.4. Convergence in probability cannot imply convergence almost surely. For example, consider $X_n(\omega) = \omega + I_{[0,1/n]}(\omega), \omega \in [0, 1]$, $P_n = 1 - 1/n$ therefore it converges

in probability but not almost surely since the non-convergent region has measure greater than 0.

However, **if sequence$\{X_n\}$ converges to X in probability, then there is a subsequence converges to X almost surely.**

12.11.3.2 *Algebraic properties*

Theorem 12.11.1 (Algebraic properties of convergence in probability). *[6, p. 297][16, p. 1165] If $X_n \xrightarrow{P} x$ and $Y_n \xrightarrow{P} y$, then*

- $X_n + Y_n \xrightarrow{P} x + y$
- $aX_n \xrightarrow{P} ax$ *for any constant* a.
- $X_n \xrightarrow{P} x \Rightarrow g(X_n) \Rightarrow g(x)$, *for any real valued function* g *continuous at* x
- $X_n Y_n \xrightarrow{P} xy$
- $X_n / Y_n \xrightarrow{P} x/y$,*if* $y \neq 0$.
- *If W_n is a matrix whose elements are random variables and if* $\operatorname{plim} W_n = \Omega$, *then*

$$\operatorname{plim} W_n^{-1} = \Omega^{-1}.$$

- *If X_n, Y_n are random matrices with* $\operatorname{plim} X_n = A, \operatorname{plim} Y_n = B$, *then*

$$\operatorname{plim} X_n Y_n = AB.$$

Proof. (1)

$$\begin{aligned} P(\left|X_n + Y_n - x - y\right| > \epsilon) &\leq P(|X_n - x| + \left|Y_n - y\right| > \epsilon) \\ &\leq P(|X_n - x| > \epsilon/2) + P(\left|Y_n - y\right| > \epsilon/2) \to 0 \end{aligned}$$

where we have used the fact that **probability measure is monotone relative to set containment**. For the first line, $|X_n - x| + \left|Y_n - y\right| \geq \left|X_n + Y_n - x - y\right| > \epsilon$, therefore when we randomly sample X_n, Y_n, we have a higher chance to have $|X_n - x| + \left|Y_n - y\right| > \epsilon$, therefore $P(\left|X_n + Y_n - x - y\right| > \epsilon) \leq P(|X_n - x| + \left|Y_n - y\right| > \epsilon)$. For the second line, $|X_n - x| > \epsilon/2, \left|Y_n - y\right| > \epsilon/2 \Rightarrow \left|X_n + Y_n - x - y\right| > \epsilon$

(2) $P(|aX_n - ax| > \epsilon) = P(|a||X_n - x| > \epsilon) = P(|X_n - x| > \epsilon/|a|) \to 0$

(3) For any $\epsilon > 0$, there exist a δ such that $|x_n - x| < \delta \Rightarrow \left|g(x_n) - g(x)\right| < \epsilon$, therefore

$$P(\left|g(X_n) - g(x)\right| < \epsilon)) \leq P(\left|X_n) - x\right| < \delta) \to 0$$

where we have used the fact that **probability measure is monotone relative to set containment**.

(4) $X_nY_n = \frac{1}{2}X_n^2 + \frac{1}{2}Y_n^2 - \frac{1}{2}(X_n - Y_n)^2$, use (1)(2)(3) to prove.

(5) use (3) to prove $1/Y_n \xrightarrow{P} 1/y$. (6)(7) We can approximately view matrix inversion and matrix multiplication as a series of algebraic operations on the matrix elements. □

12.11.4 Mean square convergence

Definition 12.11.3. *Let $\{X_n\}$ be a sequence of random variables. Then X_n converges to a random variable X in mean square if:*

$$\lim_{n\to\infty} E[(X_n - X)^2] = 0.$$

Theorem 12.11.2 (mean square convergence to a constant). *Let $\{X_n\}$ be a sequence of random variables and c be a constant. We say X_n converges to c if*

- $\lim_{n\to\infty} E[X_n] = c$.
- $\lim_{n\to\infty} Var[X_n] = 0$.

Proof. Use notation $\mu_n = E[X_n]$. Note that

$$\begin{aligned}
\lim_{n\to\infty} E[(X_n - c)^2] &= \lim_{n\to\infty} E[(X_n - \mu_n + \mu_n - c)^2] \\
&= \lim_{n\to\infty} E[(X_n - \mu_n)^2] + 2\lim_{n\to\infty} E[(X_n - \mu_n)(\mu_n - c)] + \lim_{n\to\infty} E[(\mu_n - c)^2] \\
&= \lim_{n\to\infty} E[(X_n - \mu_n)^2] + 0 + 0 \\
&= 0
\end{aligned}$$

□

Theorem 12.11.3 (convergence in mean square implies convergence in probability.). *Let $\{X_n\}$ be a sequence of random variables. If X_n converges to X in mean square, then X_n converges to X in probability.*

Proof. Given $\epsilon > 0$, we have

$$P(|X_n - X| > \epsilon) = P(|X_n - X|^2 > \epsilon^2) < E[(X_n - X)^2]/\epsilon^2 \to 0.$$

□

12.11.5 Convergence in rth mean

Definition 12.11.4. *[11, p. 308] Let $\{X_n\}$ be a sequence of random variables. Then X_n converges to X in rth mean $r \geq 1$, if $E[X_n^r] < \infty$:*

$$\lim_{n\to\infty} E[(X_n - X)^r] = 0$$

12.11.6 Convergence in distribution

Definition 12.11.5 (Convergence in distribution). *[6, p. 300]Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. Let F_{X_n} and F_X be the cumulative distribution function of X_n and X. Let $C(F_X)$ denote the set of all points where F_X is continuous. We say X_n converges in distribution to X if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \forall x \in C(F_X)$$

We denote as

$$X_n \xrightarrow{D} X$$

12.11.6.1 *Convergence in probability vs in distribution*

Theorem 12.11.4. *[6, p. 304][11, p. 311] If X_n converges to X in probability, then X_n converges to X in distribution.*

Proof. Let x be a point of continuity of $F_X(x)$. For every $\epsilon > 0$,

$$\begin{aligned} F_{X_n}(x) &= P(X_n \leq x) \\ &= P(X_n \leq x \cap |X_n - X| < \epsilon) + P(X_n \leq x \cap |X_n - X| \geq \epsilon) \\ &\leq P(X_n < x + \epsilon) + P(|X_n - X| \geq \epsilon) \end{aligned}$$

where the inequality is established by using a containing set. Then we have

$$\limsup_{n\to\infty} F_{X_n}(x) \leq P(X_n < x + \epsilon) = F_X(x + \epsilon)$$

since the second term can be arbitrarily small. Similarly, we have

$$\liminf_{n\to\infty} F_{X_n}(x) \geq P(X < x - \epsilon) = F_X(x - \epsilon)$$

We therefore have

$$F_X(x-\epsilon) \leq \liminf_{n\to\infty} F_{X_n}(x) \leq \limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x+\epsilon)$$

As $\epsilon \to 0$, we have $\liminf_{n\to\infty} F_{X_n}(x) = \limsup_{n\to\infty}$ as required by the continuity of F_X, then $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$. □

Remark 12.11.5.

- In the above proof, we cannot directly use $\lim_{n\to\infty} F_{X_n}$ because it might not exist; however $\limsup_{n\to\infty} F_{X_n}$ always exists for bounded sequence.
- Convergence in distribution is weaker than convergence almost surely, because **it says nothing on the mapping from random experiment outcomes to** $\mathbb{R}$. For example, let X be a normal random variable, let $Y = -X$, then Y and X are the same in distribution, but X and Y are totally different mappings.

Theorem 12.11.5 (Convergence to a constant). *[6, p. 305]If X_n converges to a constant b in distribution, then X_n converges to b in probability.*

Proof. for any $\epsilon > 0$, we have $P(|X_n - b| > \epsilon) = F_{X_n}(b+\epsilon) - F_{X_n}(b-\epsilon) \to 1 - 0 = 1$. □

12.12 Finite sampling models

12.12.1 Counting principles

Theorem 12.12.1 (Fundamental counting principle). *Suppose that two events occur in order. If the first can occur in m ways and the second in n ways(after the first has occurred), then the two events can occur in order in $m \times n$ ways.*

Definition 12.12.1 (permutation).

- *A **permutation** of any r elements taken from a set of n elements is an arrangement of the r elements. We denote the number of such permutations by $P(n,r)$.*
- *A **permutation** is an arrangement of objects. For example, the permutations of three letters abc are the six arrangements:*

$$abc, acb, bac, bca, cab, cba.$$

Theorem 12.12.2 (number of permutations).

- *The number of permutations for n objects is*

$$P(n,n) = n!$$

.

- *The number of permutations of n objects taken from r at a time is*

$$P(n,r) = \frac{n!}{(n-r)!}$$

Proof. Choosing r elements from a set of size n, we have:

- the first element can be selected n ways.
- the second element can be selected $n-1$ ways(since now there are $n-1$ left).
- the third element can be selected $n-2$ ways.
- Continue the process, and the r^{th} element can be selected $n-r+1$ ways.

Using the fundamental counting principle [Theorem 12.12.1], we have

$$P(n,r) = n(n-1)(n-2)\cdots(n-r+1).$$

□

Lemma 12.12.1 (number of distinguishable permutations). *If a set of n objects consists of k different kinds of objects with n_i objects of the i kind such that $\sum_{i=1}^{k} = n$. **Objects from the same kind is not distinguishable**. Then then the number of distinguishable permutations of these objects is*

$$\frac{n!}{n_1!n_2!...n_k!}.$$

Definition 12.12.2 (combination). *A combination is a subset of elements of a set.*

Example 12.12.1. The combinations of size $r = 1, 2, 3$ taken from the set $\{a, b, c\}$ is given in the following table.

$r = 1$	$r = 2$	$r = 3$
$\{a\}$	$\{a, b\}$	$\{a, b, c\}$
$\{b\}$	$\{a, c\}$	
$\{c\}$	$\{b, c\}$	

Theorem 12.12.3 (number of combinations). *The number of combinations(or subsets) of size r which can be selected from a set of size n, denoted by $C(n, r)$ or $\binom{n}{r}$, is*

$$C(n, r) = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Proof. Because combinations are essentially permutations where order does not matter. Then

$$C(n, r) = \frac{P(n, r)}{r!} = \frac{n!}{r!(n-r)!}.$$

□

Lemma 12.12.2 (decomposition).

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

Proof. Choosing k objects from n objects can be done by choosing k objects from $n-1$ objects or choosing $k-1$ objects from $n-1$ objects plus the rest. □

Lemma 12.12.3.

- *Given a set of n objects, the number of ways to divide them into k groups, each with n_i objects such that $\sum_{i=1}^{k} n_i = n$, is given by*

$$\frac{n!}{n_1!n_2!...n_k!}$$

- *Select n_1 objects from n objects to form a group, the number of ways is given by*

$$\frac{n!}{n_1!(n-n_1)!}.$$

Example 12.12.2. Assume 365 days a year. Among N people, the probability of exact 2 people has the same day of birthday is given as

$$365 \times \frac{1}{365} \times \frac{1}{365} \times (364 \cdot 363 \cdot \ldots (364 - (n-2) + 1)/365^{n-2}.$$

Example 12.12.3. Assume 365 days a year. Among N people, the probability of at least 2 people has the same day of birthday is given as

$$1 - \frac{365 \cdot 364 \cdots 365 - n + 1}{365^n}.$$

Example 12.12.4. 52 cards are randomly distributed to 4 players with each player getting 13 cards. What is the probability that each of them will have an ace.
Solution: The total possibilities are

$$N_0 = \frac{52!}{13!13!13!13!}.$$

The possibilities that each of them has an ace is

$$N_1 = \frac{48!}{12!12!12!12!}4!.$$

Then, we have

$$p = \frac{N_1}{N_0}.$$

Example 12.12.5. Imagine you have the following setup:

`_A1_A2_A3_A4_`

Each ace separated out evenly and we are interested in the pile that's before A1. For a standard deck of cards you have 52 cards - 4 aces = 48 cards left, and

$$\frac{48}{5} = 9.6,$$

cards for each pile. So basically you would have to turn all 9.6 cards + the A1 card in order to see the first ace. So the answer is

$$1 + \frac{48}{5}.$$

12.12.2 Matching problem

Example 12.12.6. A secretary randomly stuffs 5 letters into 5 envelopes. We want to find the probability of exactly k matches, with $k \in \{0, 1, ..., 5\}$.

Lemma 12.12.4 (sampling with replacement). *Define* $I_j = 1(X_j = j)$.

- $(I_1, I_2, ..., I_n)$ *is a sequence of n Bernoulli trials, with success probability* $\frac{1}{n}$.
- *The number of matches* N_n *is binomial distribution with parameter n and* $1/n$.

Lemma 12.12.5 (probability of the union of n events). *For any n events* $E_1, E_2, ..., E_n$ *that are defined on the same sample space, we have the following formula:*

$$P(E_1 \cup E_2 \cup ... \cup E_n) = \sum_{m=1}^{n} (-1)^{m+1} S_m,$$

where

$$\begin{aligned} S_1 &= \sum_{i=1}^{n} P(E_i) \\ S_2 &= \sum_{1\le j<k\le n} P(E_i \cap E_j) \\ &\cdots\cdots \\ S_m &= \sum_{1\le i_1<i_2<\ldots<i_m\le n} P(E_{i_1} \cap E_{i_2} \cap \ldots \cap E_{i_m}). \end{aligned}$$

In particular,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2),$$

and

$$\begin{aligned} P(E_1 \cup E_2 \cup E_3) = {} & P(E_1) + P(E_2) + P(E_3) \\ & - P(E_1 \cap E_2) - P(E_1 \cap E_3) - P(E_1 \cap E_3) + P(E_1 \cap E_2 \cap E_3). \end{aligned}$$

Lemma 12.12.6 (The matching problem). *link Suppose that the n letters are numbered* $1, 2, \cdots, n$. *Let* E_i *be the event that the* i^{th} *letter is stuffed into the correct envelop.*

$P(E_1 \cup E_2 \cup \cdots \cup E_n)$ *is the probability that at least one letter is matched with the correct envelop.*

- $1 - P(E_1 \cup E_2 \cup \cdots \cup E_n)$ *is the probability that* ***all*** *letters matched incorrectly.*
- *The probability of the intersection of m events is:*

$$P(E_{i(1)} \cap E_{i(2)} \cap \cdots \cap E_{i(m)}) = \frac{(n-m)!}{n!}.$$

- $P(E_1 \cup E_2 \cup \cdots \cup E_n)$ *can be calculated using the probability of event union lemma [Lemma 12.12.5].*

Proof. (3) The calculation of the probability of intersection of m events can use the following model. There are totally $n!$ ways putting letters into envelops; there are totally $(n-m)!$ ways putting letters into envelops such that at least m specified letters are in the correct envelops. Therefore,

$$P(E_{i(1)} \cap E_{i(2)} \cap \cdots \cap E_{i(m)}) = \frac{(n-m)!}{n!}.$$

□

12.12.3 Birthday problem

Definition 12.12.3. *The sampling experiment as a distribution of n balls into m cells; X_i is the cell number of ball i. In this interpretation, our interest is in the number of empty cells and the number of occupied cells.*

Example 12.12.7. In a set of n randomly chosen people, some pair of them will have the same birthday.

Lemma 12.12.7. *Let Y_i to denote the number of balls falling into the i box, then*

$$p(Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m) = \frac{n!}{y_1! y_2! \ldots y_m!} \frac{1}{m^n}, \sum_{i=1}^{m} y_i = n$$

That is, the random vector $(Y_1, ..., Y_m)$ has the multinomial distribution with parameter n and $(1/m, ..., 1/m)$.

Example 12.12.8. Assume 365 days a year.

- Among N people, the probability of exact 2 people has the same day of birthday is given as

$$365 \times \frac{1}{365} \times \frac{1}{365} \times \frac{364 \cdot 363 \cdot \ldots (364 - (n-2) + 1)}{/} 365^{n-2}.$$

- The probability that at least 2 have the same birthday is

$$1 - \frac{1}{365^n} \frac{365!}{(365-n)!}.$$

Example 12.12.9. If you randomly put 18 balls into 10 boxes, what is the expected number of empty boxes? For each box, the probability of being empty is $(\frac{9}{10})^18$, then the expected number of empty boxes is $10(\frac{9}{10})^18$.

Lemma 12.12.8 (generalized birthday problem). *link Given a year with d days, the generalized birthday problem asks for the minimal number $n(d)$ such that, in a set of n*

randomly chosen people, the probability of a birthday coincidence is at least 50%. It follows that $n(d)$ is the minimal integer n such that

$$1 - (1 - \frac{1}{d})(1 - \frac{2}{d} \cdots (1 - \frac{n-1}{d})) \geq 1/2.$$

12.12.4 Coupon collection problem

Definition 12.12.4 (coupon collection problem). *Suppose that there is an urn of n different coupons. How many coupons do you expect you need to draw **with replacement** before having drawn each coupon at least once?*

Lemma 12.12.9. *Consider the coupon collection problem with m different coupons. Let Z_i denote the number of additional samples needed to go from $i-1$ distinct coupons to i distinct coupons. Let W_k denote the number of samples needed to get k distinct coupons. Then*

- *Then $Z_1, ..., Z_m$ is a sequence of independent random variables, and Z_i has the geometric distribution with parameter $p_i = \frac{m-i+1}{m}$.*
- $W_k = \sum_{i=1}^{k} Z_i$.
- $E[W_k] = \sum_{i=1}^{k} \frac{m}{m-i+1}$.

Proof. (1) When $i = 1$, Z_1 has a geometric distribution with parameter $p_1 = 1$. Similarly, Z_2 has a geometric distribution with parameter $p_2 = (m-1)/m$; Z_3 has a geometric distribution with parameter $p_3 = (m-2)/m$. Then, we can generalize to Z_i has a geometric distribution with parameter $p_i = (m-(i-1))/m$. (3) From the property of geometric distribution [Lemma 13.1.25],

$$E[W_k] = E[Z_1] + E[Z_2] + ... + E[Z_k] = \frac{1}{p_1} + \frac{1}{p_2} + ... + \frac{1}{p_k}.$$

□

Lemma 12.12.10. *Consider the coupon collection problem with m different coupons. Among m different coupons, there are $n, n \leq m$ are special coupons. Let Z_i denote the number of additional samples needed to go from $i-1$ distinct special coupons to i distinct special coupons. Let W_k denote the number of samples needed to get k distinct special coupons. Then*

- *Then $Z_1, ..., Z_m$ is a sequence of independent random variables, and Z_i has the geometric distribution with parameter $p_i = \frac{n-i+1}{m}$.*
- $W_k = \sum_{i=1}^{k} Z_i$.

- $E[W_k] = \sum_{i=1}^{k} \frac{m}{n-i+1}$.

Proof. (1) When $i = 1$, Z_1 has a geometric distribution with parameter $p_1 = n/m$. Similarly, Z_2 has a geometric distribution with parameter $p_2 = (n-1)/m$; Z_3 has a geometric distribution with parameter $p_3 = (n-2)/m$. Then, we can generalize to Z_i has a geometric distribution with parameter $p_i = (n-(i-1))/m$. (3) From the property of geometric distribution [Lemma 13.1.25],

$$E[W_k] = E[Z_1] + E[Z_2] + ... + E[Z_k] = \frac{1}{p_1} + \frac{1}{p_2} + ... + \frac{1}{p_k}.$$

□

12.12.5 Balls into bins model

Definition 12.12.5 (balls into bins problems). *Suppose there are m balls and n bins, balls are thrown into bins where each ball is thrown into a bin uniformly at random.*

- *Pick a bin. What is the probability for this box to be empty? What is the expected number of bins that are empty?*
- *Pick a bin. What is the probability for this box to contain exactly 1 ball? What is the expected number of bins that contain exactly 1 ball.*
- *Pick a bin. What is the probability for this box to contain exactly i balls? What is the expected number of bins that contain exactly i balls?*

Example 12.12.10. Suppose there are N types of coupons in a box. If a child draws with replacement m times from the box, what is the expected number of distinct coupon types?

View each coupon as a box. And this problem is equivalent to throw m balls into N boxes and ask the expected number of non-empty boxes.

- For each box, the probability of being empty is $(\frac{N-1}{N})^m$; therefore, the probability of being non-empty is $1 - (\frac{N-1}{N})^m$.
- The expected number of empty boxes is $N(\frac{N-1}{N})^m$, and nonempty boxes is $N - N(\frac{N-1}{N})^m$.

Definition 12.12.6 (balls-into-bins distribution problems).

- *(distribution of distinguishable balls into indistinguishable bins without restriction) Suppose we want to put m distinguishable balls into n labeled bins. What is the number of ways that the balls are in different bins?*
- *(distribution of indistinguishable balls into indistinguishable bins without restriction) Suppose we want to put m indistinguishable balls into n labeled bins. What is the number of ways that the balls are in different bins?*
- *(distribution of distinguishable balls into distinguishable bins without restriction) Suppose we want to put m indistinguishable balls into n labeled bins. What is the number of ways that the balls are in different bins?*
- *(distribution of indistinguishable balls into distinguishable bins without restriction) Suppose we want to put m indistinguishable balls into n labeled bins. What is the number of ways that the balls are in different bins?*
- *(distribution without restriction I) Suppose we want to put m labeled balls into n labeled bins. What is the number of ways that the balls are in different bins such that each bin has at least has one ball?*
- *(distribution without restriction II) Suppose we want to put m labeled balls into n labeled bins. What is the number of ways that the balls are in different bins such that each bin has at least k balls?*

Lemma 12.12.11. *link*

The number of ways of putting m distinguishable balls in n distinguishable bins is m^n.

- *The number of ways of putting m distinguishable balls in n indistinguishable bins is $m^n/n!$.*
- *The number of ways of putting m indistinguishable balls in n distinguishable bins is*

$$\binom{m+n-1}{n-1}.$$

- *The number of ways of putting m indistinguishable balls in n indistinguishable bins is*

$$\binom{m+n-1}{n-1}\frac{1}{n!}.$$

- *The number of ways of putting m indistinguishable balls in n distinguishable bins and ensure each bin has at least one ball is*

$$\binom{(m-n)+n-1}{n-1}.$$

Proof. (1) The number of ways of putting m balls in n bins is m^n since each ball has n bins to go. (2) Use (1) and divide it by double counting. (3, 4) Transform to selecting $n-1$ separations from $m+n-1$ possibilities. (5) We need to first put one ball into each bin. □

Remark 12.12.1 (equivalence of distinct root problem)**.**

- The number of ways to distribute m indistinguishable balls into n distinguishable bins is equivalent to the number of solutions to the equation:

$$x_1 + x_2 + x_3 + \ldots + x_n = m, x_i \geq 0.$$

- The number of ways to distribute m indistinguishable balls into n distinguishable bins and ensure each bin to have at least one ball is equivalent to the number of non-negative solutions to the equation:

$$x_1 + x_2 + x_3 + \ldots + x_n = m, x_i \geq 1.$$

Example 12.12.11. If there are 200 students in the library, how many ways are there for them to be split among the floors of the library if there are 6 floors? The answer is 6^200.

12.13 Law of Large Number and Central Limit theorem

12.13.1 Law of Large Numbers

Theorem 12.13.1 (Weak Law of Large Numbers). *[8, p. 232]Let $\{X_n\}$ be a sequence of **iid random variables** having common mean $EX_i = \mu$ and the variance $\sigma^2 < \infty$. Let $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$. Then for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P(\left|\bar{X}_n - \mu\right| < \epsilon) = 1;$$

that is, X_n converges in probability to μ.

Proof. The weak law can be easily proved using probability Markov inequalities. □

Remark 12.13.1 (Cauchy random variable does not hold). An example where the law of large numbers does not apply is the standard Cauchy distributionLemma 13.1.46, which does not have the expectation. And the average of n such variables has the same distribution as one such variable. The probability of the averaging deviation from μ does not tend toward zero as n goes to infinity.

Theorem 12.13.2 (Strong Law of Large Numbers). *[8, p. 235]Let $\{X_n\}$ be a sequence of iid random variable having common mean $EX_i = \mu$,$E[|X_i|] < \infty$ and the variance $\sigma^2 < \infty$. Let $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$. for arbitrary $\delta > 0$:*

$$P(\lim_{n\to\infty}\left|X_n - \mu\right| < \delta) = 1$$

that is, $\bar{X}_n$ converges almost surely to μ.

Remark 12.13.2 (discussion).

- Compared to weak law, strong law requires one more moment condition $E[|X_i|] < \infty$
- The weak law states that for a specified large n, the average $\overline{X}_n$ will be concentrated on μ. However, it may still have nonzero possibility that $\left|\overline{X}_n - \mu\right| > \epsilon$; that is, such situation will happen an infinite number of times, although at infrequent intervals.
- The strong law shows with probability 1, we have that for any $\epsilon > 0$, there exists an $N > 0$ such that the inequality $\left|\overline{X}_n - \mu\right| < \epsilon$ holds for all large enough $n > N$, except possible at zero-measure set.

Remark 12.13.3 (Cauchy random variable does not hold). An example where the law of large numbers does not apply is the standard Cauchy distributionLemma 13.1.46, which does not have the expectation. And the average of n such variables has the same distribution as one such variable. The probability of the averaging deviation from μ does not tend toward zero as n goes to infinity.

12.13.2 Central limit theorem

Theorem 12.13.3 (central limit theorem). *[8, p. 236][6, p. 313] Let $X_1, X_2, ..., X_n$ be a sequence **iid random variables** that have mean μ and variance $\sigma^2 < \infty$. Then the random variable*

$$Y_n = \frac{(\sum_{i=1}^n X_i/n - \mu)}{\sigma/\sqrt{n}} = \frac{(\sum_{i=1}^n X_i - n\mu)}{\sqrt{n}\sigma} = \sqrt{n}(\bar{X}_n - \mu)/\sigma$$

***converges in distribution to** $N(0,1)$.*

Proof. Use moment generating function(if exists) or characteristic function to prove.

Let $\phi(t) = E[\exp(it(X-\mu))] = \exp(\frac{i\sigma^2 t^2}{2})$ be the characteristic function of X. Then the characteristic function for Y_n can be derived via

$$\begin{aligned}
\Phi(t,n) &= E[\exp(it\frac{(\sum_{j=1}^n X_j/n - \mu)}{\sigma/\sqrt{n}})] \\
&= \phi(\frac{t}{\sigma\sqrt{n}})^n \\
&= (1 - \frac{t^2}{2n} + O((t/\sqrt{n})^3))^n \\
&\to \exp(-\frac{t^2}{2}), n \to \infty
\end{aligned}$$

where we use the Taylor expansion of $\phi(t)$ given by

$$\phi(t) = \phi(0) + \phi'(0)t + \phi''(0)\frac{t^2}{2} + O(t^3) = 1 - \sigma^2\frac{t^2}{2} + O(t^3),$$

and the limit theorem to e [Lemma 1.5.2].

That is, as $n \to \infty$, Y_n will have its characteristic function converge to the characteristic function of the standard normal. □

Remark 12.13.4 (convergence rate). We can view the sample mean $\overline{X}_n$ has distribution similar to $N(\mu, \sigma/\sqrt{n})$ at large n. Therefore, the convergence rate is $O(1/\sqrt{N})$.

Remark 12.13.5 (Situations where central limit theorem breaks down).

- The sample mean of the iid standard Cauchy distribution random variable will not converge in distribution to standard normal; instead, the sample mean will converge to standard Cauchy distribution [Lemma 13.1.46]. Note that standard Cauchy does not have finite mean and variance.

Example 12.13.1 (application of CLT for normal approximation).

- Let $X_1, ..., X_n$ be independent iid random variable of $Exp(\lambda)$, then

$$Y = \sum_{i=1}^{n} X_i$$

can be approximated(when $n \to \infty$) by

$$\frac{Y - n\mu}{\sqrt{n\sigma}} \sim N(0,1),$$

where $\mu = n/\lambda$,and $\sigma = 1/\lambda^2$.

- Let $X_1, ..., X_n$ be independent iid random variable of $Poisson(\theta)$, then

$$Y = \sum_{i=1}^{n} X_i$$

can be approximated by

$$\frac{Y - n\theta}{\sqrt{n\theta}} \sim N(0,1),$$

or equivalently

$$Y \sim N(n\theta, \theta/n).$$

Theorem 12.13.4 (general central limit theorem). *Let $X_1, X_2, ...$ be independent random variables. Suppose they have*

$$E[X_k] = \mu_k, Var[X_k] = \sigma_k^2.$$

Further let

$$B_n^2 = \sum_{k=1}^{n} \sigma_k^2.$$

If there exists a $\delta > 0$ such that as $n \to \infty$, we have

$$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^{n} E[|X_k - \mu_k|^{2+\delta}] \to 0,$$

then

$$Z = \frac{\sum_{k=1}^{n} X_k - \mu}{\sigma}$$

converges to $N(0,1)$ in distribution, where μ and σ^2 are the mean and variance of $\sum_{k=1}^{n} X_k$.

Lemma 12.13.1 (Slutsky's theorem). *[8, p. 239] If $X_n \to X$ is distribution and $Y_n \to a$, a constant, in probability, then*

- *$Y_n X_n \to aX$ in distribution*
- *$X_n + Y_n \to X + a$ in distribution.*

Remark 12.13.6 (linearity in convergence in distribution). This importance of Slutsky's theorem is that it provides the sufficient condition for linearity in convergence in distribution.

12.13.3 Delta method & generalized CLT

Lemma 12.13.2 (first-order approximation to mean and variance of a function). *[8, p. 242] Let $T_1, ..., T_k$ be random variables with mean $\mu_1, ..., \mu_k$, and define $T = (T_1, ..., T_k)$ and $\mu = (\mu_1, ..., \mu_k)$. Define a differentiable function $g : \mathbb{R}^k \to \mathbb{R}$. Then we have the following first-order approximate mean and variance:*

$$E[g(T)] \approx g(\mu)$$
$$Var[g(T)] \approx \sum_{i=1}^{k} [g_i'(\mu)]^2 Var[T_i] + 2 \sum_{i=1}^{k} \sum_{j>i}^{k} g_i'(\mu) g_j'(\mu) Cov_{ij}[T].$$

Proof. (1)

$$\begin{aligned} g(T=t) &= g(\mu) + \sum_{i=1}^{k} g_i'(\mu)(t-\mu) + o((t-\mu)) \\ &\approx g(\mu) + \sum_{i=1}^{k} g_i'(\mu)(t-\mu) \\ E[g(T)] &\approx g(\mu) + \sum_{i=1}^{k} g_i'(\mu)(E[T]-\mu) = g(\mu) \end{aligned}$$

(2)

$$\begin{aligned} g(T=t) &\approx g(\mu) + \sum_{i=1}^{k} g_i'(\mu)(t-\mu) \\ Var[g(T)] &\approx Var[\sum_{i=1}^{k} g_i'(\mu)(T-\mu)] = \sum_{i,j} g_i'(\mu) g_j'(\mu) Cov_{ij} \end{aligned}$$

□

Corollary 12.13.4.1. *Let $T_1, ..., T_k$ be a iid random sample of T. Assume $E[T] = \mu$. Let $g : \mathbb{R} \to \mathbb{R}$ be a differentiable function. Then, we have the first-order approximation:*

$$\begin{aligned} E[g(T)] &\approx g(\mu) \\ Var[g(T)] &\approx [g'(\mu)]^2 Var[T]. \end{aligned}$$

Moreover, let $\overline{T}$ be the sample mean. Then,

$$\begin{aligned} E[g(\overline{T})] &\approx g(\mu) \\ Var[g(\overline{T})] &\approx [g'(\mu)]^2 \frac{Var[T]}{k}. \end{aligned}$$

Example 12.13.2. Let X and Y are random variables with means μ_X and μ_Y, respectively. Let $g(x,y) = x/y$. $\frac{\partial g}{\partial x} = \frac{1}{\mu_X}, \frac{\partial g}{\partial y} = -\frac{\mu_X}{\mu_Y^2}$.

We have

$$E[\frac{X}{Y}] \approx \frac{\mu_X}{\mu_Y}$$

and

$$E[\frac{X}{Y}] \approx \frac{1}{\mu_Y^2} Var[X] + \frac{\mu_X^2}{\mu_Y^4} Var[Y] - 2\frac{\mu_X}{\mu_Y^3} Cov(X,Y).$$

Theorem 12.13.5 (Delta method for central limit theorem). *[8, p. 243] Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to N(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is not o. Then*

$$\sqrt{n}(g(Y_n) - g(\theta)) \to N(0, \sigma^2[g'(\theta)]^2)$$

in distribution.

12.14 Order statistics

Definition 12.14.1. *The order statistics of a random sample $X_1, ..., X_n$ are the sample values placed in ascending order. And they are denoted by $X_{(1)}, X_{(n)}, ..., X_{(n)}$, where $X_{(1)} \leq X_{(2)} ... \leq X_{(n)}$*

Theorem 12.14.1 (Discrete order statistics). *[8] Let $X_1, ..., X_n$ be a random sample from a discrete distribution with pmf $f_X(x_i) = p_i$, where $x_1 < x_2 < ...$ are possible values of X in ascending order. Define*

$$\begin{aligned} P_0 &= 0 \\ P_1 &= p_1 \\ &... \\ P_i &= \sum_{k=0}^{i} p_i \end{aligned}$$

Let $X_{(1)}, X_{(n)}, ..., X_{(n)}$ denote the order statistics from the sample. Then

$$P(X_{(j)} \leq x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k} \tag{6}$$

and

$$P(X_{(j)} = x_i) = \sum_{k=j}^{n} \binom{n}{n-k} [P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k} \tag{7}$$

Proof. We can treat P_i as discrete version of cdf, and it means the probability of one X satisfies the inequality. The order statistics connected to binomial distribution as:

- If the minimum of Xs are less then x, then there are 1,2,...,n out of n are less than x.
- If the second minimum of Xs are less then x, then there are 2,3,...,n out of n are less than x.

□

Theorem 12.14.2 (Continuous order statistics). *[8] Let $X_1, ..., X_n$ be a random sample from a continuous distribution with pmf f_X and cdf $F_X(x)$. Let $X_{(1)}, X_{(n)}, ..., X_{(n)}$ denote the order statistics from the sample. Then*

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x)[F_X(x)]^{j-1}[1-F_X(x)]^{n-j}$$

Proof. We can use

$$F_{X_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k (1-F_X(x))^{n-k}$$

and take derivative.
Another proof: When jth order statistic at x, that means we from n variables, we first select 1 variable to be at x, then from rest of $n-1$ variables, we select $j-1$ to be smaller than x then the rest greater than x. From combinatorics, we know

$$f_j(x) = nf(x)\binom{n-1}{j-1}(F(x))^{j-1}(1-F(x))^{n-j}$$

□

Lemma 12.14.1 (Two order statistics). *Let $X_1, ..., X_n$ be a random sample from a continuous distribution with pmf f and cdf $F(x)$. Let $X_{(1)}, X_{(n)}, ..., X_{(n)}$ denote the order statistics from the sample. Then the joint density for $X_{(r)}$ and $X_{(s)}$ is:*

$$f_{r,s}(u,v) = \frac{n!}{(r-1)!(n-s)!(s-r-1)!} f(u)f(v)(F(u))^{r-1}(1-F(v))^{n-s}(F(v)-F(u))^{s-r-1}$$

Proof. Use the argument similar to above: just divide the variables into five groups. □

Corollary 12.14.2.1. *Let $X_1, ..., X_n$ be a random sample from a continuous distribution with pmf f and cdf $F(x)$. Let $X_{(1)}, X_{(n)}, ..., X_{(n)}$ denote the order statistics from the sample. Then we have*

- $$f_{1,n}(u,v) = n(n-1)(F(v)-F(u))^{n-2}f(u)f(v)$$
- *(density of range) Let $W = X_{max} - X_{min}$, then*

$$f_W(w) = \int_u f_{1,n}(u, u+w)du$$

Lemma 12.14.2 (joint density of all the order statistics). *Let $X_1, ..., X_n$ be a random sample from a continuous distribution with pmf f and cdf $F(x)$. Let $X_{(1)}, X_{(n)}, ..., X_{(n)}$ denote the order statistics from the sample. Then the conditional joint density function of $X_{(1)}, X_{(n)}, ..., X_{(n)}$ is given by*

$$f_{1,2,...,n}(y_1, ..., y_n) = n! f(y_1) f(y_2) ... f(y_n) I_{y_1 < y_2 ... < y_n}$$

Proof. The sample space of $(X_1, ..., X_n)$ can be partition into $n!$ **equal-sized** subspaces such that $X_1 < X_2 < ... < X_n, ...$ In each this subspace, there exist a map from $(X_1, ..., X_n)$ to $(X_{(1)}, X_{(n)}, ..., X_{(n)})$ with the Jacobian being 1(since it is a permutation matrix). The density for $(X_{(1)}, X_{(n)}, ..., X_{(n)})$ is $f(y_1) f(y_2) ... f(y_n) I_{y_1 < y_2 ... < y_n}$. Use the law of total probability [Theorem 12.2.1]. □

Lemma 12.14.3 (distribution of max and min). *Let X be a random variable with cdf $F_X(x)$. Let $Y_n = \min(X_1, ..., X_n)$ and $Z_n = \max(X_1, ..., X_n)$, where $X_1, ..., X_n$ are n iid random sample of X. Then*

$$f_{Y_n}(x) = n f_X(x)(1 - F_X(x))^{n-1}$$

and

$$f_{Z_n}(x) = n f_X(x)(F_X(x))^{n-1}.$$

Proof.

$$P(Y_n \geq x) = (P(X \geq x))^n \implies 1 - F_{Y_n}(x) = (1 - F_X(x))^n \implies f_{Y_n}(x) = n f_X(x)(1 - F_X(x))^{n-1}$$

and

$$P(Z_n \leq x) = (P(X \leq x))^n \implies F_{Z_n}(x) = (F_X(x))^n \implies f_{Z_n}(x) = n f_X(x)(F_X(x))^{n-1}.$$

□

Corollary 12.14.2.2 (order statistics of uniform random variables). *Let X be a uniform random variable at [o,1]. Let $Y_n = \min(X_1, ..., X_n)$ and $Z_n = \max(X_1, ..., X_n)$, where $X_1, ..., X_n$ are n iid random sample of X. Then*

- $$f_{Y_n}(x) = n f_X(x)(1 - F_X(x))^{n-1} = n(1 - x)^{n-1}$$
- $$f_{Z_n}(x) = n f_X(x)(F_X(x))^{n-1} = n x^{n-1}$$

- $$f_j = \frac{n!}{(j-1)!(n-j)!}[x]^{j-1}[1-x]^{n-j} = Beta(j, n-j+1)$$

Proof. Note that we use the fact that for $U(0,1)$ distribution, $F_X(x) = x, f_X(x) = 1$. □

12.15 Information theory

12.15.1 Concept of entropy

Definition 12.15.1 (entropy of a random variable).

- *Let X be a discrete random variable taking values $x_k, k = 1, 2...$ with probability mass function*
$$Pr(X = x_k) = p_k, k = 1, 2, ...$$
Then the entropy of X is defined by
$$H(X) = -\sum_{k \geq 1} p_k \ln p_k.$$
- *If X is a continuous random variable with pdf $f(x)$, then entropy of X is defined by*
$$H(X) = -\int_{-\infty}^{\infty} f(x) \ln f(x) dx.$$

Remark 12.15.1 (entropy, information and probability distribution).

- Entropy is a measure of the uncertainty of a random variable: the larger the value, the uncertainty the random variable is.
- When the random variable is deterministic, the entropy is at the minimum.

Example 12.15.1.

- The entropy of the Gaussian density on $\mathbb{R}$ with mean μ and variance σ^2 is
$$\begin{aligned} H &= -\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma} \exp(-1/2((x-\mu)^2/\sigma^2))(-\ln(\sqrt{2\pi}\sigma) - 1/2((x-\mu)^2/\sigma^2))dx \\ &= \frac{1}{2} + \ln(\sqrt{2\pi}\sigma). \end{aligned}$$
Note that the mean μ does not enter the entropy; therefore the entropy for Gaussian distribution is translational invariant.
- The entropy of the exponential distribution with mean λ and pdf
$$f(x) = \frac{1}{\lambda} \exp(-x/\lambda)$$

is

$$H = -\int_0^\infty \frac{1}{\lambda}\exp(-x/\lambda)(-\ln\lambda - x/\lambda)dx = \ln\lambda + 1.$$

Lemma 12.15.1 (basic properties of entropy).

- $H(X) \geq 0$.
- $H(X) = 0$ *if and only if there exists a* x_0 *such that* $P(X = x_0) = 1$.
- *If* X *can take on finite number* n *values, then* $H(X) \leq \log(n)$. $H(X) = \log(n)$ *if and only if* X *is uniformly distributed.*
- *Let* $X_1, X_2, ..., X_n$ *be discrete valued random variables on a common probability space. Then*
$$H(X_1, X_2, ..., X_n) = H(X_1) + \sum_{i=2}^{n} H(X_n|X_1, ..., X_{n-1}).$$
- $H(X) + H(Y) \geq H(X, Y)$, *with equality if and only if* X *and* Y *are independent.*

Proof. (1) note that every term $\log(p)$ is non-positive, therefore $H(X) \geq 0$. (2) direct verification. (3) direct verification. (4) It can be showed that $H(X, Y) = H(X|Y) + H(Y)$. (5) $H(X, Y) = H(X|Y) + H(Y) \leq H(X) + H(Y)$(using chain rule and conditioning entropy). □

12.15.2 Entropy maximizing distributions

Theorem 12.15.1 (continuous distribution with maximum entropy). *Suppose* S *is a closed subset of* $\mathbb{R}$. *Let* X *be a random variable with support* S *and pdf* $f(x)$.

Then, the probability density function $f(x)$ *maximizing the entropy*

$$H(X) = -\int_{-\infty}^{\infty} f(x)\ln f(x)dx,$$

and satisfying the following n *constraints*

$$E[g_j(X)] = a_j, \forall j = 1, 2, ..., n.$$

and sum-to-unit constraint

$$\int_S f(x)dx = 1,$$

has the form

$$f(x) = c\exp(\sum_{j=1}^{n} \lambda_j g_j(x)), \forall x \in S,$$

where the constant c and the n multipliers λ_i are determined by the above $n+1$ constraints.

Proof. Note that our constraints can be written as

$$\int_{-\infty}^{\infty} g_j(x) f(x) dx = a_j, j = 1, 2, ..., n$$

$$\int_{-\infty}^{\infty} f(x) dx = 1, j = 1, 2, ..., n.$$

The Lagrange of our minimizing problem is given by

$$J[p(x)] = \int_{-\infty}^{\infty} f(x) \ln f(x) dx - \lambda_0 (\int_{-\infty}^{\infty} f(x) dx - 1) - \sum_{j=1}^{n} \lambda_j (\int_{-\infty}^{\infty} g_j(x) f(x) dx - a_j).$$

where $\lambda_i, i = 0, 1, 2, ..., n$ are Lagrange multipliers.

The first order optimality condition gives

$$\frac{\delta J}{\delta f(x)} = \ln f(x) + 1 - \lambda_0 - \lambda_j g_j(x),$$

or equivalently

$$f(x) = \exp(-1 + \lambda_0) \exp(\sum_{j=1}^{n} \lambda_j g_j(x)) = c\exp(\sum_{j=1}^{n} \lambda_j g_j(x)).$$

Note that the second order conditions gives $\frac{\delta^2 J}{\delta f(x)^2} = 1/f(x) > 0$, which ensures we have unique global minimum solution. □

Corollary 12.15.1.1.

- *The uniform distribution on the interval $[a, b]$ is the maximum entropy distribution among all continuous distribution supported on $[a, b]$.*
- *The exponential distribution, for which the density function with parameter λ is*

$$f(x|\lambda) = \begin{cases} \lambda \exp(-\lambda x), x \geq 0 \\ 0, x < 0 \end{cases}$$

is the maximum entropy distribution among all continuous distributions supported in $[0, \infty]$ that have a specified mean of $1/\lambda$.

- *The normal distribution with parameter μ and σ, for which the density function is*

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}),$$

has the maximum entropy among all distributions supported on $\mathbb{R}$ with a specified mean μ and variance σ^2.

Proof. (1) from Theorem 12.15.1, we know that the $f(x)$ should have the following form

$$f(x) = c.$$

and c is determined by

$$\int_a^b f(x)dx = c(b-a) = 1 \implies c = \frac{1}{b-a}.$$

Therefore,

$$f(x) = \frac{1}{b-a}, x \in [a,b].$$

(2)Similarly, we know that the $f(x)$ should have the following form

$$f(x) = c\exp(\mu x),$$

where μ is the Lagrange multiplier and c is determined by

$$\int_0^\infty f(x)dx = \frac{c}{\mu} = 1 \implies c = \mu.$$

and then

$$\int_0^\infty xf(x)dx = -\frac{1}{\mu} = 1/\lambda \implies \mu = -\lambda.$$

(3)Similarly, we know that the $f(x)$ should have the following form

$$f(x) = c\exp(\lambda_1 x + \lambda_2(x-\mu)^2).$$

Then we can determine c, λ_1, λ_2 using constraints. □

Theorem 12.15.2 (discrete distribution with maximum entropy). *Suppose $S = \{x_1, x_2, ...\}$ is a (finite or infinite) discrete subset of of $\mathbb{R}$. Let X be a random variable with support S and probability mass function given by $Pr(X = x_k)$.*

Then, the probability mass function $Pr(X)$ maximizing the entropy

$$H(X) = -\sum_{k\geq 1} Pr(X = x_k)\ln Pr(X = x_k),$$

and satisfying the following n constraints

$$E[g_j(X)] = a_j, \forall j = 1, 2, ..., n.$$

and sum-to-unit constraint

$$\sum_{k \geq 1} Pr(X = x_k) = 1,$$

has the form

$$Pr(X = x_k) = c \exp(\sum_{j=1}^{n} \lambda_j g_j(x_k)), \forall x_k \in S,$$

where the constant c and the n multipliers λ_i are determined by the above $n + 1$ constraints.

Proof. Note that our constraints can be written as

$$\int_{-\infty}^{\infty} g_j(x) f(x) dx = a_j, j = 1, 2, ..., n$$

$$\int_{-\infty}^{\infty} f(x) dx = 1, j = 1, 2, ..., n.$$

Let $p_k = Pr(X = x_k)$. The Lagrange of our minimizing problem is given by

$$L = \sum_{i=1}^{n} p_k \ln p_k - \lambda_0 (\sum_{i \geq 1} p_k - 1) - \sum_{j=1}^{n} \lambda_j (\sum_{i \geq 1} g_j(x_i) p_i - a_j).$$

where $\lambda_i, i = 0, 1, 2, ..., n$ are Lagrange multipliers.

The first order optimality condition for p_i gives

$$\frac{\partial L}{\partial p_i} = \ln p_i + 1 - \lambda_0 - \lambda_j g_j(x_i), i \geq 1$$

or equivalently

$$Pr(X = x_i) = p_i = \exp(-1 + \lambda_0) \exp(\sum_{j=1}^{n} \lambda_j g_j(x_i)) = c \exp(\sum_{j=1}^{n} \lambda_j g_j(x_i)).$$

Note that the second order conditions gives $\frac{\partial^2 L}{\partial p_i} = 1/p_i > 0$, which ensures we have unique global minimum solution. □

Corollary 12.15.2.1. *For a probabilistic mass function p on a finite set $\{x_1, x_2, ..., x_n\}$, the entropy H is bounded by*

$$H \leq \ln n$$

with equality holds if and only if p is uniform, i.e., $p(x_i) = 1/n, \forall i$.

Proof. From Theorem 12.15.2, we know that the $p(x)$ should have the following form

$$p(x_i) = c.$$

and c is determined by

$$\sum_{i=1}^{n} c = 1 \implies c = \frac{1}{n}.$$

Therefore,

$$p(x_i) = 1/n, \forall i.$$

□

12.15.3 KL divergence

Definition 12.15.2 (Kullback-Leibler divergence, KL divergence). *Given two discrete probability distribution P and Q defined on the same set $\mathcal{X}$, the KL divergence from Q to P is defined as*

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$$

Lemma 12.15.2 (non-negativeness of KL divergence). *Given two discrete probability distribution P and Q defined on the same set $\mathcal{X}$,*

$$D_{KL}(P||Q) \geq 0.$$

And the equality holds if $P = Q$.

Proof.

$$D_{KL}(P||Q) = -\sum_{x \in \mathcal{X}} P(x) \log \frac{Q(x)}{P(x)} \geq -\log(\sum_{x \in \mathcal{X}} \frac{Q(x)}{P(x)} P(x)) = 0$$

where the fact that $-\log(x)$ is a convex function and Jensen's inequality has been used [Lemma 12.10.1]. □

12.15.4 Conditional entropy and mutual information

Definition 12.15.3 (conditional entropy).

- ***Specific conditional entropy*** $H(X|Y=v)$ *of* X *given* $Y=v$*:*

$$H(X|Y=v) = -\sum_{i=1}^{n} P(X=i|Y=v)\log P(X=i|Y=v).$$

- ***Conditional entropy*** $H(X|Y)$ *of* X *given* Y*:*

$$H(X|Y) = \sum_{v\in Val(Y)} P(Y=v)H(X|Y=v).$$

Definition 12.15.4 (mutual information). *[17] Consider two discrete random variables* X *and* Y *taking values in* $\mathcal{X}$ *and* $\mathcal{Y}$*. The* ***mutual information, or information gain)*** *of* X *and* Y *is given as:* $I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

Lemma 12.15.3.

$$I(X,Y) \geq 0$$

where $I(X,Y)=0$ *if* X *and* Y *are independent.*

Proof. (1)

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x)) + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x|y)p(y)\log(p(x|y)) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x)) + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(\frac{p(x,y)}{p(y)}) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)}) = DL(p(x,y)||p(x)p(y)) \geq 0. \end{aligned}$$

(2) When X and Y are independent, we have

$$H(X|Y) = \sum_{v\in\mathcal{Y}} (Y=v)H(X|Y=v) = \sum_{v\in\mathcal{Y}} P(Y=v)H(X) = H(x).$$

□

Corollary 12.15.2.2 (conditioning reduce entropy). *Given discrete random variables X and Y, we have*

$$H(X|Y) \leq H(X),$$

which is also known as conditioning reduces entropy (i.e., conditioning provides information); and this equality holds if and only if X and Y are independent.

Chain rule: $H(X,Y) = H(X) + H(Y|X)$ can be proved using $P(X,Y) = P(X|Y)P(Y)$.

12.15.5 Cross-entropy

Definition 12.15.5 (cross-entropy of two probability distributions). *Consider a probability distribution on N value with probability $y_i, i = 1, 2, ..., N$. Consider another distribution on the same support and probabilities $y'_i, i = 1, 2, ..., N$. Then the cross-entropy of the two distributions is defined by*

$$H(y, y') = \sum_{i=1}^{N} y_i \log \frac{1}{y'_i} = -\sum_{i=1}^{N} y_i \log y'_i.$$

Lemma 12.15.4 (properties of cross entropy). *Consider two discrete distributions, characterized by probability mass vectors y and y', on the same N values.*

- *The KL divergence on the two distributions is the difference between cross entropy and entropy; that is*

$$KL(y||y') = \sum_{i=1}^{N} y_i \log \frac{y_i}{y'_i} = \underbrace{\sum_{i=1}^{N} y_i \log \frac{1}{y'_i}}_{cross\ entropy} - \underbrace{\sum_{i=1}^{N} y_i \log \frac{1}{y_i}}_{entropy}.$$

- *Cross entropy is no smaller than entropy*

$$H(y, y') \geq H(y).$$

Proof. (1) Straight forward. (2) Use the fact that

$$KL(y||y') = H(y, y') - H(y) \geq 0.$$

□

Remark 12.15.2 (cross entropy, maximum likelihood, and classificiation accuracy). Consider a K-class classification problem with N training examples. The target of each example is represented by a K-dimensional one-hot vector. The classification output generated by the classifier can be represented by a discrete distribution vector.

For example, let $y^{(1)} = (1, 0, 0,)$ be the target vector of example 1 and $\hat{y}^{(1)} = (0.4, 0.1, 0.5)$ be a prediction output based on input of example 1.

Note that the likelihood for example i is given by

$$L(y^{(i)}; \hat{y}^{(i)}) = \prod_{k=1}^{K} [\hat{y}_j^{(i)}]^{y^{(i)}},$$

whose the logarithm form is

$$\log L(y^{(i)}; \hat{y}^{(i)}) = \sum_{k=1}^{K} y_k^{(i)} \log \hat{y}_k^{(i)} = -H(y^{(i)}, \hat{y}^{(i)}).$$

For overall N examples, the overall negative log likelihood is

$$-\log L = -\sum_{n=1}^{N} \log L(y^{(n)}; \hat{y}^{(n)}) = \sum_{n=1}^{N} H(y^{(i)}, \hat{y}^{(i)}).$$

Therefore, **minimizing the negative log likelihood is equivalent to minimizing the cross-entropy.**

12.16 Notes on bibliography

For excellent treatment on the whole topic, see[18][19]. For clear treatment on conditional expectation, see [20],[13].

For clear treatment on σ field and measure, see[1][21].

For problems in probability, see [22][23].

For treatment on measure and integral, see [24].

An excellent online resource is http://www.math.uah.edu/stat/, including random variable vector space theory(http://www.math.uah.edu/stat/expect/Spaces.html), finite sampling model(http://www.math.uah.edu/stat/urn/index.html), Brownian motion (http://www.math.uah.edu/stat/brown/Standard.html).

BIBLIOGRAPHY

1. Dineen, S. *Probability theory in finance: a mathematical guide to the Black-Scholes formula* (American Mathematical Soc., 2013).
2. Rosenthal, J. S. *A first look at rigorous probability theory* (World Scientific, 2006).
3. Wikipedia. *Borel set — Wikipedia, The Free Encyclopedia* [Online; accessed 18-May-2016]. 2016.
4. Wikipedia. *Measure (mathematics) — Wikipedia, The Free Encyclopedia* [Online; accessed 18-May-2016]. 2016.
5. Shreve, S. E. *Stochastic calculus for finance II: Continuous-time models* (Springer Science & Business Media, 2004).
6. Hogg, R. V., McKean, J. & Craig, A. T. Introduction to Mathematical Statistics, 7 ed (2012).
7. Fries, C. *Mathematical finance: theory, modeling, implementation* (John Wiley & Sons, 2007).
8. Casella, G. & Berger, R. L. *Statistical inference* (Duxbury Pacific Grove, CA, 2002).
9. Stefanica, D. *A primer for the Mathematics of Financial Engineering* (Fe Press, 2008).
10. Mitzenmacher, M. & Upfal, E. *Probability and computing: Randomized algorithms and probabilistic analysis* (Cambridge University Press, 2005).
11. Grimmett, G. & Stirzaker, D. *Probability and Random Processes* ISBN: 9780198572220 (OUP Oxford, 2001).
12. Williams, D. *Probability with martingales* (Cambridge university press, 1991).
13. Mikosch, T. *Elementary stochastic calculus with finance in view* (World scientific, 1998).
14. Tripathi, G. A matrix extension of the Cauchy-Schwarz inequality. *Economics Letters* **63,** 1–3 (1999).
15. De Micheaux, P. L. & Liquet, B. Understanding convergence concepts: A visual-minded and graphical simulation-based approach. *The American Statistician* (2012).
16. Greene, W. *Econometric Analysis* ISBN: 9780134461366 (Pearson, 2017).
17. Cover, T. M. & Thomas, J. A. *Elements of information theory* (John Wiley & Sons, 2012).
18. Shiryaev, A. N. Probability: Volume 1 (Graduate Texts in Mathematics) (1996).

19. Feller, W. *An introduction to probability theory and its applications* (John Wiley & Sons, 2008).

20. Brzezniak, Z. & Zastawniak, T. *Basic stochastic processes: a course through exercises* (Springer Science & Business Media, 1999).

21. Koralov, L. & Sinai, Y. G. *Theory of probability and random processes* (Springer Science & Business Media, 2007).

22. Capinski, M. & Zastawniak, T. J. *Probability through problems* (Springer Science & Business Media, 2013).

23. Grimmett, G. & Stirzaker, D. *One thousand exercises in probability* (Oxford University Press, 2001).

24. Capinski, M. & Kopp, P. E. *Measure, integral and probability* (Springer Science & Business Media, 2013).

13

STATISTICAL DISTRIBUTIONS

13.1 Common distributions and properties

13.1.1 Bernoulli distribution

Definition 13.1.1 (Bernoulli distribution). *A random variable Y with sample space $\{0,1\}$ is said to have Bernoulli distribution $Ber(\theta)$ with parameter θ if it has a pmf given as*

$$p(y) = \theta^y(1-\theta)^{1-y}, y \in \{0,1\}.$$

Example 13.1.1. Consider the experiment of toss a biased coin. The probability of getting head is p and getting tail is $1-p$. The outcome of coin toss can be modeled by a Bernoullie random variable.

Lemma 13.1.1 (basic properties). *Let X be a random variable with distribution $Ber(p)$. Then*

- $M_X(t) = (1-p+pe^t)$.
- $E[X] = p, E[X^2] = p, Var[X^2] = p - p^2 = p(1-p)$.

Proof. Straight forward. □

13.1.2 Normal distribution

Definition 13.1.2 (normal distribution). *A random variable X with normal distribution $N(\mu, \sigma^2)$, characterized by parameters μ and σ, has its pdf given by*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}(x-\mu)^2/\sigma^2), -\infty < x < \infty.$$

X is called normal random variable, or Gaussian random variable. If $\mu = 0, \sigma = 1$, X is also called standard normal random variable.

Lemma 13.1.2 (moment generating function). *Let X be a random variable with normal distribution $N(0,1)$, then the moment generating function is*

$$m_X(t) = \exp(\frac{1}{2}t^2).$$

If Y is a random variable with normal distribution $N(\mu, \sigma^2)$, then the moment generating function is

$$m_Y(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2).$$

Proof. (1) $m_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ complete the square and get the result. (2)Let $Y = \sigma X + \mu$ and use Lemma 12.7.2. Then $m_Y(t) = e^{\mu t} m_X(\sigma t)$ □

Lemma 13.1.3 (basic properties of normal random variable). *Consider $X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2)$.*

- *If X, Y **are independent**, then we have*
  - $$aX + b \sim N(a\mu + b, a^2\sigma_x^2)$$
  - $$X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$
    $$aX + bY \sim N(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$$
- *If X **and** Y **are not independent but jointly normal**, then $X + Z$ will be normal, and*
  $$aX + bZ \sim N(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_z^2 + 2abCov(X, Z)).$$
- *Assume X, Y are independent. Further let $W = \rho X + \sqrt{1-\rho^2}Y, \rho \in [-1, 1]$. Then W is normal, correlated with X and Y, and the sum $X + W$ is also normal; that is,*
  $$W \sim N, aX + bW \sim N.$$
- ***In general, the sum of two dependent normal random variable is not necessarily normal.*** *See Corollary 13.1.1.1.*

Proof. (1) Directly from the properties of moment generating functions at Lemma 12.7.2. (2)The proof of two general jointly normal random variable will be showed in Corollary 13.1.1.1. (3) Note that

$$(X, W)^T = (X, \rho X + \sqrt{1-\rho^2}Y)^T = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} (X, Y)^T,$$

therefore (X, W) are jointly normal [Theorem 15.1.1]. Then we use (2). □

Lemma 13.1.4 (moments of standard normal distribution). *Let $X \sim N(0,1)$, then*

$$E[X] = 0, E[X^2] = 1, E[X^3] = 0, E[X^4] = 3$$

Moreover, all odd moments are 0.

Proof. The mgf is $m(t) = e^{t^2/2}$, then

$$m'(t) = te^{t^2/2}$$

$$m''(t) = e^{t^2/2} + t^2 e^{t^2/2}$$

$$\dots$$

For all odd moments

$$\int x^{2k+1} f(x) dx$$

has integrand as odd function. □

Corollary 13.1.0.1 (moments of normal distribution). *Let $X \sim N(0, \sigma^2)$, then*

$$E[X] = 0, E[X^2] = \sigma^2, E[X^3] = 0, E[X^4] = 3\sigma^4$$

Moreover, all odd moments are 0.

Proof. The mgf is $m(t) = e^{\sigma^2 t^2/2}$, then

$$m'(t) = \sigma^2 te^{\sigma^2 t^2/2}$$

$$m''(t) = \sigma^2 e^{\sigma^2 t^2/2} + \sigma^4 t^2 e^{t^2/2}$$

$$\dots$$

For all odd moments

$$\int x^{2k+1} f(x) dx$$

has integrand as odd function. □

13.1.3 Half-normal distribution

Definition 13.1.3 (half-normal distribution). *Let X follow an ordinary normal distribution $N(0,\sigma^2)$. Then $Y=|X|$ follows a **half-normal distribution** with parameter σ. It has probability density function*

$$f_Y(y;\sigma)=\frac{\sqrt{2}}{\sigma\sqrt{\pi}}\exp(-\frac{y^2}{2\sigma^2}).$$

Lemma 13.1.5 (basic properties of half normal distribution). *Let Y follow a half-normal distribution with parameter σ. Then*

- $E[Y]=\frac{\sigma\sqrt{2}}{\sqrt{\pi}}$.
- $Var[Y]=\sigma^2(1-\frac{2}{\pi})$

13.1.4 Laplace distribution

Definition 13.1.4 (Laplace distribution). *A random variable X has a **Laplace distribution**, denoted by $Lap(\mu,b)$, if its probability density function is*

$$f(x|\mu,b)=\frac{1}{2b}\exp(-\frac{|x-\mu|}{b})=\begin{cases}\frac{1}{2b}\exp(-\frac{\mu-x}{b}),if\ x<\mu\\ \frac{1}{2b}\exp(-\frac{x-\mu}{b}),if\ x\geq\mu\end{cases}.$$

Lemma 13.1.6 (properties of Laplace distribution). *Let X be a random variable with Laplace distribution with parameter μ,b. It follows that*

- *The mean and the median are μ.*
- *The variance is $2b^2$.*
- *The cdf is given by*

$$\begin{aligned}F(x)=\int_{-\infty}^{x}f(u)du&=\begin{cases}\frac{1}{2}\exp(-\frac{\mu-x}{b}),if\ x<\mu\\ 1-\frac{1}{2}\exp(-\frac{x-\mu}{b}),if\ x\geq\mu\end{cases}\\ &=\frac{1}{2}+\frac{1}{2}sgn(x-\mu)(1-\exp(-\frac{|x-\mu|}{b})).\end{aligned}$$

- *The inverse cdf is given by*

$$F^{-1}(p)=\mu-b\cdot sgn(p-0.5)\ln(1-2|p-0.5|).$$

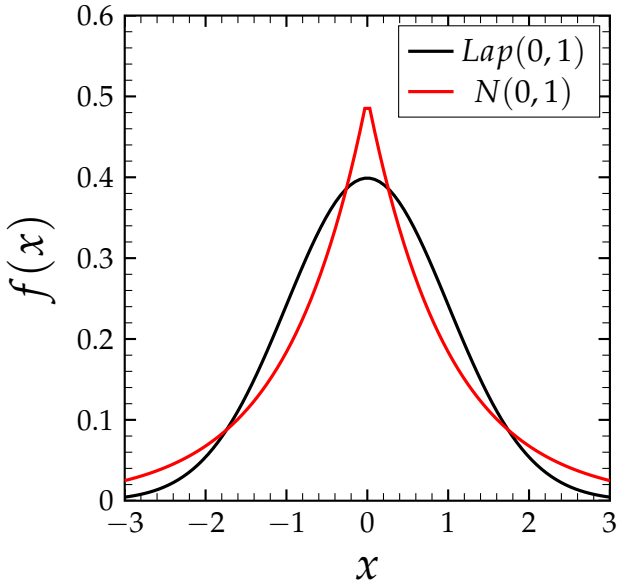


Figure 13.1.1: Comparison of Laplace distribution and normal distribution.

13.1.5 Multivariate Gaussian/normal distribution

13.1.5.1 *Basic definitions*

Definition 13.1.5 (multivariate Gaussian/normal distribution). *A random vector is said be multivariate Gaussian/normal random variable if it's pdf is multivariate Gaussian/normal distribution, whose support is $\mathbb{R}^n$ and its pdf is*

$$\rho(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|det\Sigma|^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$.

Lemma 13.1.7 (mgf of multivariate Gaussian/normal random variables). *[1, p. 181] An n-dimensional random vector $X \sim MN(\mu,\Sigma)$ has its mgf given by*

$$M_X(t) \triangleq E[\exp(t^T X)] = \exp(t^T\mu + \frac{1}{2}t^T\Sigma t)$$

for all $t \in \mathbb{R}^n$.

Proof.

$$\begin{aligned} M_X(t) &= E[\exp(t^T X)] \\ &= \exp(E[t^T X] + \frac{1}{2} Var[t^T X]) \\ &= \exp(t^T \mu + \frac{1}{2} t^T \Sigma t) \end{aligned}$$

where we use the fact that $t^T X \sim N(t^T \mu, t^T \Sigma t)$ from Theorem 15.1.1. □

Remark 13.1.1 (implication). Given a random vector X, if we want to check whether X is a multivariate Gaussian, we can check its mgf. If its mgf is the exponential of a linear form plus a quadratic form, then it is multivariate Gaussian.

Lemma 13.1.8 (criterion via linear combination). *A vector $X = (X_1, X_2, ..., X_n)^T$ is a multivariate Gaussian distribution if every linear combination*

$$S = a^T X, a \in \mathbb{R}^n$$

has a normal distribution.

Proof. Because $a^T X$ is normal, then it has characteristic function

$$E[\exp(ita^T X)] = \exp(ita^T \mu_X - \frac{1}{2} t^2 (a)^T \Sigma_X a).$$

Since a is arbitrary, we can say for any $t' \in \mathbb{R}^n$, we have

$$E[\exp(it' X)] = \exp(i[t']^T \mu_X - \frac{1}{2} [t']^T \Sigma_X t).$$

That is, X is multivariate Gaussian. □

Example 13.1.2 (bivariate Gaussian distribution). Let $f(x, y)$ be the density of a bivariate Gaussian distribution $MN(\mu, \Sigma)$, where

$$\mu = \begin{Bmatrix} \mu_X \\ \mu_Y \end{Bmatrix}, \Sigma = \begin{Bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{Bmatrix}.$$

Then,

$$f(x,y) = \frac{\exp(-\frac{1}{2(1-\rho^2)}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\frac{\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}]).$$

13.1.5.2 *Affine transformation and its consequences*

Theorem 13.1.1 (affine transformation for multivariate normal distribution). *Let X be an n-dimensional random vector with $MN(\mu, \Sigma)$ distribution. Let $Y = AX + b, A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$. Then Y is an m-dimensional random vector having a $MN(A\mu + b, A\Sigma A^T)$ distribution.*

Proof. Use moment generating function to prove. Let $Y = AX + b$, then from Lemma 12.7.5

$$M_Y(t) = e^{t^T b} M_X(A^T t) = e^{t^T(A\mu+b)+\frac{1}{2}t^T A\Sigma A^T t}$$

which indicates $Y \sim MN(A\mu + b, A\Sigma A^T)$. □

Corollary 13.1.1.1 (sum of two multivariate normal random vectors). *Let $X_1 \sim MN(\mu_1, \Sigma_1)$ and $X_2 \sim MN(\mu_1, \Sigma_1)$ be two n dimensional multivariate normal random variable. It follows that*

- *If X_1 and X_2 are independent, then $Y = X_1 + X_2$ is a multivariate normal random vector with $MN(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$.*
- *If X_1 and X_2 are dependent and (X_1, X_2) are **jointly normal** [a] with covariance matrix given by*

$$\Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12} & \Sigma_2 \end{bmatrix},$$

then $Y = X_1 + X_2$ is a multivariate normal random vector with $MN(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2 + 2\Sigma_{12})$.

[a] If X_1 and X_2 are not jointly normal, Y is not normal.

Proof. (1) Consider a $2n$-dimensional multivariate normal random variable Z with distribution $\mu = [\mu_1; \mu_2]$, $\Sigma = \Sigma_1 \oplus \Sigma_2$ (Σ is a diagonal block matrix with two blocks Σ_1 and Σ_2). Then we can construct a linear transformation matrix

$$A = \begin{bmatrix} I_n & I_n \end{bmatrix}$$

to construct $Y = AZ$. Apply affine transformation theorem [Theorem 15.1.1] to $Y = AZ$, we have $Y \sim MN(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$. (2) same as (1). □

Note 13.1.1. caution! The joint distribution of two Gaussian margins are not necessarily joint Gaussian:

- Two multivariate normal random variables are not necessarily joint normal.[a]. For example, consider two marginal distribution of Gaussian. For Gaussian copula, the joint distribution is multivariate Gaussian; however, for other copulas including Frank copula and Clayton copula, the joint distribution is not multivariate Gaussian.
- If two multivariate normal random variables are independent, then they are joint normal.

[a] link

Corollary 13.1.1.2 (orthonormal transformation maintains independence). *Let X be a n dimensional random vector with $MN(0, I)$. If C is an orthonormal matrix, then $Y = CX$ has distribution $MN(0, I)$. That is, orthonormal transformation will preserve independence.*

Proof. $Cov(Y) = C^T IC = I$. □

13.1.5.3 *Marginal and conditional distribution*

Lemma 13.1.9 (marginal distribution). *The multivariate Gaussian distribution $\rho(x; \mu, \Sigma)$ on $\mathbb{R}^n$ has marginal distribution on $\mathbb{R}^k, k \leq n$ given as $\rho(x_1; \mu_1, \Sigma_{11}), x_1 \in \mathbb{R}^k$ where we decompose*

$$\mu = [\mu_1, \mu_2]^T, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Proof. Use above Theorem 15.1.1. Let

$$A = \begin{bmatrix} I & 0 \end{bmatrix}$$

Then $X_1 = AX$. □

Lemma 13.1.10 (full joint distribution can be constructed from pair joint distribution). *Let $X = (X_1, X_2, ..., X_n)^T$ be a random multivariate Gaussian vector with mean $\mu = (\mu_1, \mu_2, ..., \mu_n)^T$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$. The the pair $(X_i, X_j), i \neq j$ has joint distribution*

$$\hat{\mu} = (\mu_i, \mu_j), \hat{\Sigma} \in \mathbb{R}^{2\times 2}, \hat{\Sigma}_{11} = \Sigma_{ii}, \hat{\Sigma}_{12} = \Sigma_{ij}.$$

That is, all the pair joint distribution can construct the full joint distribution.

Proof. Directly from Lemma 15.1.3. □

Remark 13.1.2 (caution! not all the distribution has this property). If the full joint distribution is not Gaussian, then such property (reconstruct full distribution from pair distribution) will not generally hold.

Theorem 13.1.2 (conditional distribution). *The multivariate Gaussian distribution $\rho(x; \mu, \Sigma)$ on $\mathbb{R}^n$ has a conditional Gaussian distribution on $\mathbb{R}^k, k \leq n$ given by*

$$\frac{\rho(x_1, x_2)}{\rho(x_2)} = \rho(x_1; \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

where we decompose

$$\mu = [\mu_1^T, \mu_2^T]^T, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

with $\mu_1 \in \mathbb{R}^k, \mu_2 \in \mathbb{R}^{n-k}$.

Proof. See link □

Remark 13.1.3 (gaining information). From the conditional distribution, we can see that given the information of x_2, the mean of x_1 will be corrected and the variance of x_1 will be reduced.

Example 13.1.3 (bivariate Gaussian distribution). Let $f(x, y)$ be the density of a bivariate Gaussian distribution $MN(\mu, \Sigma)$, where

$$\mu = \begin{Bmatrix} \mu_X \\ \mu_Y \end{Bmatrix}, \Sigma = \begin{Bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{Bmatrix}.$$

Then,

$$X|Y \sim N(\mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y - \mu_Y), (1-\rho^2)\sigma_X^2).$$

13.1.5.4 *Box Muller transformation*

Lemma 13.1.11 (Box Muller transformation). *Let $X, Y \sim N(0,1)$ and X, Y be independent. Let*

$$R = \sqrt{X^2 + Y^2}, \Theta = \arctan(Y/X)$$

. Then

- *R and Θ are independent.*
- *$\Theta \sim U(0, 2\pi)$ and $F_R(r) = 1 - \exp(-r^2/2)$.*
- *Suppose we have U_1, U_2 being independent uniform on $[0,1]$. Then $2\pi U_1$ and $\sqrt{-2\ln(1-U_2)}$ are independent and have the same distribution of R and Θ.*
- *Further, $\sqrt{-2\ln(1-U_2)}\cos(2\pi U_1)$ and $\sqrt{-2\ln(1-U_2)}\sin(2\pi U_1)$ are independent and have the same distribution of X and Y.*

Proof. (1)Using polar transformation Lemma 12.4.10, we have

$$\begin{aligned} Pr(R < r, \Theta < \theta) &= \int_0^r \int_0^\theta \frac{1}{2\pi}\exp(-\frac{r^2}{2}) r dr d\theta \\ &= \int_0^r \int_0^\theta \exp(-\frac{r^2}{2}) r dr \frac{1}{2\pi} d\theta \\ &= F_R(R < r) F_\Theta(\Theta < \theta) \end{aligned}$$

Using independence condition Lemma 12.4.3, we know that R and Θ are independent. (2) Integrate directly in (1). (3) Let $U = 1 - \exp(-R^2/2)$. Based on probability integral transform [Lemma 15.4.2], we know U is an uniform random variable. Or equivalently, $R = \sqrt{-2\ln(1-U)}$ has the same distribution of R. (4) Note that $X = R\cos(\Theta), Y = R\sin(\Theta)$. □

13.1.6 Lognormal distribution

13.1.6.1 *Univariate lognormal distribution*

Definition 13.1.6 (lognormal distribution). *A random variable Y has a lognormal distribution with parameters μ and σ^2, written as*

$$Y \sim LN(\mu, \sigma^2)$$

if $\log(Y)$ is normally distributed as $N(0, \sigma^2)$. Several equivalent definitions are:

- *$Y \sim LN(\mu, \sigma^2)$ if and only if $\log(Y) \sim N(\mu, \sigma^2)$.*
- *$Y \sim LN(\mu, \sigma^2)$ if and only if $Y = e^X$ with $X \sim N(\mu, \sigma^2)$.*
- *The distribution function is given as*

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp(-\frac{(\ln y - \mu)^2}{2\sigma^2}).$$

Lemma 13.1.12 (basic properties of lognormal distribution). *Let $Y \sim LN(\mu, \sigma^2)$, or equivalently $Y = \exp(X), X \sim N(\mu, \sigma^2)$ then*

- *The distribution function for Y is given as*

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp(-\frac{(\ln y - \mu)^2}{2\sigma^2}).$$

- $E[Y] = \exp(E[X] + \frac{1}{2}Var[X^2]) = \exp(\mu + \sigma^2/2)$.
- $E[Y^2] = \exp(2\mu + 2\sigma^2), E[Y^m] = \exp(m\mu + \frac{1}{2}m^2\sigma^2)$
- *$Var[Y] = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$. In particular $\mu = 0$, we have*

$$E[Y] = \exp(\frac{1}{2}\sigma^2), E[Y^m] = \exp(\frac{1}{2}m^2\sigma^2), Var[Y] = \exp(2\sigma^2) - \exp(\sigma^2).$$

- *If $X_1 \in N(\mu_1, \sigma_1^2), X_2 \in N(\mu_2, \sigma_2^2)$, then*

$$E[\exp(X_1 + X_2)] = \exp(E[X_1] + E[X_2] + \frac{1}{2}Var[X] + \frac{1}{2}Var[X_2] + Cov(X_1, X_2)).$$

- $$\mu = \log(\frac{E[Y]^2}{\sqrt{E[Y^2]}}), \sigma^2 = \ln(\frac{E[Y^2]}{E[Y]^2}).$$

- *The median of Y is $\exp(\mu)$.*
- *skewness*

$$(\exp(\sigma^2) + 2)\sqrt{\exp(\sigma^2) - 1} > 0.$$

Proof. (1) Note that

$$x = \ln y, f_Y(y) = f_X(\ln y)\left|\frac{d\ln y}{dy}\right|.$$

(2)(3) Note that for $X \sim N(\mu, \sigma^2)$, $M_X(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$. Then

$$E[Y] = E[\exp(X)] = M_X(1) = \exp(\mu + \frac{1}{2}\sigma^2),$$

and

$$E[Y^2] = E[\exp(2X)] = M_X(2) = \exp(2\mu + 2\sigma^2),$$

and

$$Var[Y] = E[Y^2] - (E[Y])^2 = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2).$$

(4) Note that $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2)$. Then we use (1). (5) Note that the exponential is a monotone function, the median of Y will be $\exp(median\ X) = \exp(\mu)$, where we used the fact that median of X is μ. □

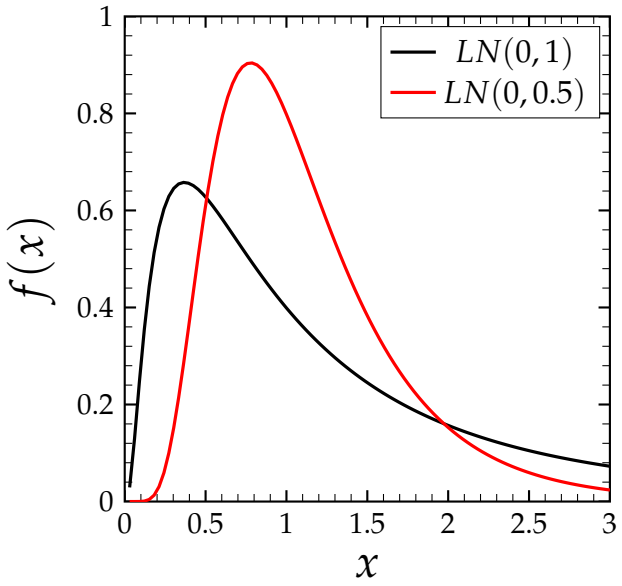


Figure 13.1.2: Density of $LN(0,1)$ and $LN(0,0.5)$. Note the positive skewness.

13.1.6.2 *Extension to univariate lognormal distribution*

Definition 13.1.7. *[2]*

- ***regular log-normal distribution*** *with parameter* (μ, σ^2) *is given by*

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp(-\frac{(\ln x - \mu)^2}{2\sigma^2}), x > 0.$$

- ***negative log-normal distribution*** *with parameter* (μ, σ)*, denoted by* $NLN(\mu, \sigma^2)$ *is given by*

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp(-\frac{(\ln -x - \mu)^2}{2\sigma^2}), x < 0.$$

- ***shifted log-normal distribution*** *with parameter* (μ, σ, τ)*,denoted by* $SLN(\mu, \sigma^2, \tau)$ *is given by*

$$f(x) = \frac{1}{(x-\tau)\sigma\sqrt{2\pi}} \exp(-\frac{(\ln x - \tau - \mu)^2}{2\sigma^2}), x > \tau.$$

- ***negative shifted log-normal distribution*** *with parameter* (μ, σ^2, τ) *is given by*

$$f(x) = \frac{1}{(-x-\tau)\sigma\sqrt{2\pi}} \exp(-\frac{(\ln(-x - \tau) - \mu)^2}{2\sigma^2}), x < -\tau.$$

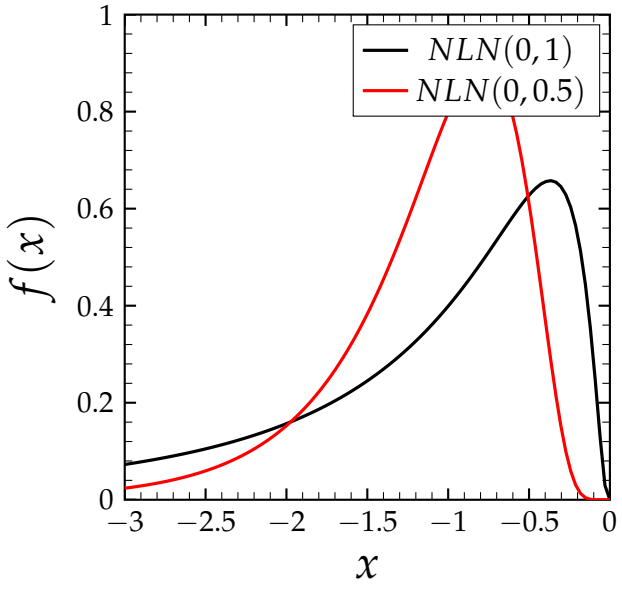


Figure 13.1.3: Density of $NLN(0,1)$ and $NLN(0,0.5)$. Note the negative skewness.

Lemma 13.1.13. *Let* $X \sim LN(\mu, \sigma^2)$*. It follows that*

- *Let $Y = -X$. Then $Y \sim NLN(\mu, \sigma^2)$.*
- *Let $Z = X + \tau$. Then $Y \sim NLN(\mu, \sigma^2, \tau)$.*
- *Let $W = -X - \tau$. Then $Y \sim NSLN(\mu, \sigma^2, \tau)$.*

Proof. Straight forward from definition and transformation. □

Lemma 13.1.14 (basic properties of shifted lognormal distribution). *Let $X \sim SLN(\mu, \sigma^2, \tau)$. Then*

- $$E[X] = \tau + \exp(\mu + \frac{1}{2}\sigma^2).$$
- $$E[X^2] = \tau^2 + 2\tau \exp(\mu + \frac{1}{2}\sigma^2) + \exp(2\mu + 2\sigma^2).$$
- $$E[X^3] = \tau^3 + 3\tau^2 \exp(\mu + \frac{1}{2}\sigma^2) + 3\tau \exp(2\mu + 2\sigma^2) + \exp(3\mu + \frac{9}{2}\sigma^2).$$

Proof. Note that from Lemma 13.1.12, we have if $Y \sim LN(\mu, \sigma^2)$, then $E[Y^m] = \exp(m\mu + \frac{1}{2}m^2\sigma^2)$. Then, we use

$$E[X] = E[Y + \tau] = E[Y] + \tau,$$
$$E[X^2] = E[(Y + \tau)^2] = E[Y^2] + 2\tau E[Y] + \tau^2,$$
$$E[X^3] = E[(Y + \tau)^3] = E[Y^3] + 3\tau E[Y^2] + 3\tau^2 E[Y] + \tau^3.$$

□

13.1.6.3 *Moment matching approximation*

Lemma 13.1.15 (2 parameter Log-normal approximation via moment matching). *Suppose we have a random variable X having moments given by*

$$E[X] = M_1, E[X^2] = M_2.$$

Let Y be a log-normal random variable defined by

$$Y = M_1 \exp(-\frac{1}{2}v^2 + vZ), Z \in N(0, 1),$$

where

$$v^2 = \log(M_2/M_1^2)$$

Then Y has the same first two moments as X; that is

$$E[Y] = M_1, E[Y^2] = M_2.$$

Proof. Using moment generating function of Z, we know that

$$E[Y] = M_Z(v)M_1 \exp(-\frac{1}{2}v^2) = M_1.$$

and

$$E[Y^2] = M_Z(2v)M_1^2 \exp(-v^2) = \exp(v^2)M_1^2 = \frac{M_2}{M_1^2}M_1^2 = M_2.$$

□

Lemma 13.1.16 (3 parameter shifted lognormal approximation via moment matching). *Suppose we have a random variable X having moments given by*

$$E[X] = M_1, E[X^2] = M_2, E[X^3] = M_3.$$

Let Y be a shifted log-normal random variable with parameter $SLN(\mu, \sigma^2, \tau)$ such that

$$E[Y] = \tau + \exp(\mu + \frac{1}{2}\sigma^2),$$

$$E[Y^2] = \tau^2 + 2\tau \exp(\mu + \frac{1}{2}\sigma^2) + \exp(2\mu + 2\sigma^2),$$

$$E[Y^3] = \tau^3 + 3\tau^2 \exp(\mu + \frac{1}{2}\sigma^2) + 3\tau \exp(2\mu + 2\sigma^2) + \exp(3\mu + \frac{9}{2}\sigma^2).$$

If we can find (μ, σ, τ) such that

$$E[X] = E[Y], E[X^2] = E[Y^2], E[X^3] = E[Y^3],$$

then X and Y have matched moments.

Proof. For moments of Y, see Lemma 13.1.14. □

Note 13.1.2 (choice of approximating distribution). [2] We can based on the target distribution's location and skewness to choose the type of lognormal distribution we want to use. The table is good summary.

skewness	$\eta > 0$	$\eta > 0$	$\eta < 0$	$\eta < 0$
location	$\tau \geq 0$	$\tau < 0$	$\tau \geq 0$	$\tau < 0$
choice of approximation	regular	shifted	negative	negative shifted

13.1.6.4 *Multivariate lognormal distribution*

Definition 13.1.8 (multivariate lognormal distribution). *If* $X = (X_1, X_2, ..., X_n) \sim MN(\mu, \Sigma)$, *then* $Y = \exp(X) = (\exp(X_1), \exp(X_2), ..., \exp(X_n)) \sim MLN(\mu, \Sigma)$, *i.e.,* Y *has multivariate lognormal distribution*

Lemma 13.1.17 (basic properties of multivariate lognormal distribution). *Let* $X = (X_1, X_2, ..., X_n) \sim MN(\mu, \Sigma)$ *and* $Y = \exp(X) = (\exp(X_1), \exp(X_2), ..., \exp(X_n))$. *Then*

- $E[Y_i] = \exp(\mu_i + \frac{1}{2}\Sigma_{ii})$.
- $E[Y_iY_j] = \exp(\mu_i + \mu_j + \frac{1}{2}(\Sigma_{ii} + \Sigma_{jj} + 2\Sigma_{ij})) = E[Y_i][Y_i]\exp(\Sigma_{ij})$.
- $Var[Y_i] = \exp(2\mu_i + \Sigma_{ii})(\exp(\Sigma_{ii}) - 1)$.
- $Cov[Y_iY_j] = \exp(\mu_i + \mu_j + \frac{1}{2}(\Sigma_{ii} + \Sigma_{jj}))(\exp(\Sigma_{ij}) - 1)$.

Proof. (1)Note that $M_X(t) = \exp(t^T\mu + \frac{1}{2}t^T\Sigma t), t \in \mathbb{R}^n$, and $E[Y_i] = M_X(e_i)$. (2) Let $t = e_i + e_j$. Then

$$E[Y_iY_j] = M_X(t) = \exp(\mu_i + \mu_j + \frac{1}{2}(\Sigma_{ii} + \Sigma_{jj} + 2\Sigma_{ij})).$$

(3)

$$Var[Y_i] = E[Y_iY_i] - E[Y_i]E[Y_i].$$

(4)

$$Cov[Y_i, Y_j] = E[Y_iY_j] - E[Y_i]E[Y_j].$$

□

13.1.7 Exponential distribution

Definition 13.1.9 (exponential distribution). *A random variable X is said to have an exponential distribution $Exp(\lambda)$ with parameter λ if it has pdf given as*

$$p(x|\lambda) = \lambda \exp(-\lambda x)$$

with $x \in [0, \infty)$.

Lemma 13.1.18 (basic properites). *Let X be a random variable with exponential distribution with parameter λ, then we have*

- $E[X] = 1/\lambda$
- $Var[X] = 1/\lambda^2$
- *memoryless:*

$$P(X > s + t | X > s) = P(X > t)$$

(even though $P(X > s + t) < P(X > t)$)

Proof. (1)(2) are straightforward. (3) The cmf is given as

$$F(t) = \int_0^t \lambda \exp(-\lambda\tau) d\tau = 1 - \exp(-\lambda t)$$

$$P(X > s+t|X > s) = \frac{P(X > s+t \cap X > s)}{P(X > s)} = \frac{P(X > s+t)}{P(X > s)} = \frac{\exp(-\lambda(s+t))}{\exp(-\lambda t)} = \exp(-\lambda t)$$

□

Remark 13.1.4 (interpretation of memorylessness). Supppose we are waiting for an event to occur, and we model the waiting time as a random variable X with $Exp(\lambda)$. If we already wait for s time, the distribution that we need to wait an extra of t time is the same as the distribution of the waiting time at time o.

Remark 13.1.5. Exponential distribution is the only memoryless continuous distribution.[3].

Lemma 13.1.19 (Normal approximate sum of Exponential). *Let $X_1, ..., X_n$ be independent iid random variable of $Exp(\lambda)$, then*

$$Y = \sum_{i=1}^{n} X_i$$

can be approximated(when $n \to \infty$) by

$$\frac{Y - n\mu}{\sqrt{n\sigma}} \sim N(0,1),$$

where $\mu = n/\lambda$,and $\sigma = n/\lambda^2$.

Proof. Directly from Central Limit Theorem [Theorem 12.13.3]. Also see Gamma distribution properties, since exponential distribution is a special case of Gamma distribution. □

13.1.8 Poisson distribution

Definition 13.1.10 (Poisson distribution). *A discrete random variable X is said to have a Poisson distribution $Poisson(\lambda)$ with parameter λ if it has pmf given as*

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

with $x \in \{0, 1, 2, ...\}$.

Lemma 13.1.20 (basic property of Poisson distribution). *[1, p. 154] Let X be a random variable with distribution $Poisson(\lambda)$. Then*

- $M(t) = \exp(\lambda(e^t - 1))$.
- $E[X] = \lambda, Var[X] = \lambda$.

Proof. (1)

$$\begin{aligned} M_X(t) &= E[e^{tX}] \\ &= \sum_{n=0}^{\infty} e^{tn} \frac{\lambda^n}{n!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} \\ &= e^{-\lambda} e^{\lambda e^t}. \end{aligned}$$

(2) $E[X] = M_X'(0) = \lambda, E[X^2] = M_X''(0) = \lambda^2 - \lambda$. □

Lemma 13.1.21 (sum of Poisson distribution). *Assume $X_1, ..., X_n$ to be independent random variables, and $X_i \sim Poisson(\theta_i), i = 1, ..., n$. Then*

$$Y = \sum_{i=1}^{n} X_i \sim Poisson(\sum_{i=1}^{n} \theta_i).$$

Proof. Note that

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \exp \sum_{i=1}^{n} \theta_i (e^t - 1).$$

□

Lemma 13.1.22 (Normal approximate sum of Poisson). *Let $X_1, ..., X_n$ be independent iid random variable of $Poisson(\theta)$, then*

$$Y = \sum_{i=1}^{n} X_i$$

can be approximated by

$$\frac{Y - n\theta}{\sqrt{n\theta}} \sim N(0, 1),$$

or equivalently

$$Y \sim N(n\theta, n\theta).$$

Proof. Directly from Central Limit Theorem [Theorem 12.13.3]. □

13.1.9 Gamma distribution

Definition 13.1.11 (Gamma distribution). *[4, p. 42] A random variable X is said to have a Gamma distribution $Gamma(a, b)$ with parameter a, b if it has pdf given as*

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

with support $x \in (0, \infty)$.

Remark 13.1.6 (exponential distribution is a special case). An exponential distribution with parameter b is a Gamma distribution $Gamma(1, b)$with

$$f(x) = be^{-bx}.$$

Remark 13.1.7 (Application in arrival times of Poisson process). If $N(t)$ is a Poisson process with rate λ, then the arrival time $T_1, T_2, ...$ have $T_n \sim Gamma(n, \lambda)$ distribution.(See Lemma 19.6.4)

Caution! Gamma distribution is different from Gamma function $\Gamma(t)$, which is given as

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$$

Remark 13.1.8 (conjugate prior for Poisson distribution). Gammad distribution conjugate prior for the parameter of Poisson distribution. When integrate out x in $\Gamma(t)$, we have

$$\int_0^\infty x^{a-1} e^{-bx} dx = \Gamma(a)/b^2$$

Lemma 13.1.23 (mean and variance). *The Gamma distribution $Gamma(a, b)$ has mean a/b and variance a/b^2.*

Proof. Using the property of

$$\int_0^\infty x^{a-1} e^{-bx} dx = \Gamma(a)/b^a,$$

we can show the result. □

Theorem 13.1.3 (sum of Gamma random variables). *[1, p. 163] Let $X_1, ..., X_n$ be independent random variables. Suppose $X_i \sim Gamma(a_i, b), \forall i = 1, ..., n$. Then*

$$Y = \sum_{i=1}^n X_i \sim Gamma(\sum_{i=1}^n a_i, b)$$

Proof. This can be proved using moment generating functions. □

Lemma 13.1.24 (Normal approximate sum of Gamma). *Let $X_1, ..., X_n$ be independent iid random variables of $Gamma(a, b)$, then*

$$Y = \sum_{i=1}^n X_i$$

can be approximated(when $n \to \infty$) by

$$\frac{Y - n\mu}{\sqrt{n\sigma}} \sim N(0,1),$$

where $\mu = na/b$,and $\sigma = na/b^2$.

Proof. Directly from Central Limit Theorem [Theorem 12.13.3]. □

13.1.10 Geometric distribution

Definition 13.1.12 (geometric distribution). *A discrete random variable X is said to have a geometric distribution $Geo(\theta)$ with parameter θ if it has pmf given as*

$$p(X = k) = (1 - \theta)^{k-1}\theta$$

with $k \in \{1, 2, ...\}$.

Example 13.1.4 (number of trials needed to succeed in Bernoulli trias). The geometric distribution The probability distribution of the number X of Bernoulli trials needed to get one success, supported on the set $\{1, 2, 3, ...\}$

Lemma 13.1.25 (basic statistics of geometric distribution). *The expected value of a geometrically distributed random variable X with parameter p is $1/p$ and the variance is $(1 - p)/p^2$.*

Proof. (1)

$$E[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p$$

$$(1-p)E[X] = \sum_{k=1}^{\infty} k(1-p)^{k}p$$

subtract and get $pE[X] = 1$.
(2)

$$Var[X] = \sum_{k=1}^{\infty} (k - 1/p)^2(1-p)^{k-1}p$$

can be proved similarly. □

Example 13.1.5 (coupon collection problem, subsection 12.12.4). Consider the **coupon collection problem** where there is an urn of m different coupons. How many coupons do you expect you need to draw **with replacement** before having drawn each coupon at least once?

Let Z_i denote the number of additional samples needed to go from $i-1$ distinct coupons to i distinct coupons. Let W_k denote the number of samples needed to get k distinct coupons. Then $Z_j, j = 1, ..., m$ is a sequence of independent random variables has the geometric distribution with parameter $p_i = \frac{m-i+1}{m}$.

When $i = 1$, Z_1 has a geometric distribution with parameter $p_1 = 1$. Similarly, Z_2 has a geometric distribution with parameter $p_2 = (m-1)/m$; Z_3 has a geometric distribution with parameter $p_3 = (m-2)/m$. Then, we can generalize to Z_i has a geometric distribution with parameter $p_i = (m-(i-1))/m$.

Further, we have

- $W_k = \sum_{i=1}^{k} Z_i$.
- $E[W_k] = \sum_{i=1}^{k} \frac{m}{m-i+1}$.

Example 13.1.6 (number of visits in a Markov chain, Lemma 22.2.1). Consider a state i in a Markov chain. Let f_{ii} denote the probability that a trajectory starting from state i will *ever* revisit i.

Then the probability of of n visit is $f_{ii}^{n-1}(1-f_{ii})$, which is the product of the probability visiting state i $n-1$ times and then never visit again.

The expected total visit is

$$\sum_{n=1}^{\infty} n f_{ii}^{n-1}(1-f_{ii}) = \frac{1}{1-f_{ii}}.$$

13.1.11 Binomial distribution

Definition 13.1.13 (binomial distribution). *A discrete random variable X is said to have a Binomial distribution $Binomial(n,p)$ with parameter n, p if it has pmf given as*

$$f(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

with $x \in \{0, 1, 2, ..., n\}$.

Remark 13.1.9 (interpretation). Binomial distribution represents the probability distribution of the number of successes in a sequence of n independent binary experiments, each of which yields 1 with probability p.

Remark 13.1.10 (relation to Bernoulli distribution). Let X_i be iid random variables with Bernoulli distribution of parameter p, then

$$Y = \sum_{i=1}^{n} X_i$$

is a random variable of binomial distribution with parameter (n, p).

Lemma 13.1.26 (sum of independent binomial random variable). *Let $X_1, X_2, ..., X_K$ be the independent binomial random variables with parameter $(n_1, p), (n_2, p), ..., (n_K, p)$. Let $Y = \sum_{i=1}^{K} X_i$. Then*

- $M_{X_i}(t) = (1 - p + pe^t)^{n_i}, i = 1, ..., K.$
- $M_Y(t) = (1 - p + pe^t)^{\sum_{i=1}^{K} n_i}$
- $Y \sim Binomial(\sum_{i=1}^{K} n_i, p)$.

Proof. (1) Use the mgf of Bernoulli distribution [Lemma 13.1.1]. (2)(3) Consider $X_1 \sim Binomial(n_1, p)$ and $X_2 \sim Binomial(n_2, p)$, each has momemt generating function of $(1 - p + pe^t)^{n_1}$ and $(1 - p + pe^t)^{n_2}$. $X_1 + X_2$ will have mgf of $(1 - p + pe^t)^{n_1+n_2}$ [Lemma 12.7.2], corresponding to $Binomial(n_1 + n_2, p)$. It is straight forward to extend multiple cases. □

Lemma 13.1.27 (convergence of binomial distribution to Poisson distribution). *Suppose that $p_n \in (0, 1)$ for $n \in \mathcal{N}_+$ and $np_n \to \lambda$ as $n \to \infty$. Then the binomial distribution with parameters n and p_n converges to the Poisson distribution with parameter λ **in distribution** as $n \to \infty$. That is, for fixed $k \in \mathcal{N}$,*

$$\binom{n}{k} p_n^k (1 - p_n)^{n-k} \to e^{-\lambda} \frac{\lambda^k}{k!}$$

as $n \to \infty$.

Proof. (direct method)Note that

$$\begin{aligned}\binom{n}{k} p_n^k (1-p_n)^{n-k} &= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}(p_n)^k(1-p_n)^{n-k} \\ &= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k} \\ &\approx \frac{n^k}{k!}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k} \\ &= \frac{\lambda^k}{k!}(1-\frac{\lambda}{n})^{n-k} \\ &\to e^{-\lambda\frac{n-k}{n}}\frac{\lambda^k}{k!} \\ &\approx e^{-\lambda}\frac{\lambda^k}{k!}\end{aligned}$$

(use generating function) Note that binomial distribution has probability generating function [Definition 12.7.6]

$$((1-p_n)+p_n s)^n = (1+(pns-p_n)n/n)^n \to e^{(s-1)a}, n\to\infty$$

where $e^{(s-1)a}$ is the generating function of Poisson distribution. □

Remark 13.1.11 (Poisson distribution as an approximate for large n and small k)**.** Note that the lemma requires that k fixed. In other words, when $n \gg k$, we can use Poisson distribution to approximate binomial distribution.

13.1.12 Hypergeometric distribution

Definition 13.1.14 (hypergeometric distribution distribution). *[1, p. 148] A random variable X is said to have a hypergeometric distribution $HG(N,K,n)$ with parameter N,K,n if it has pmf given as*

$$p(x=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

with support $x \in \{0,1,...,\min(n,K)\}$. Note that the parameters should be non-negative integers and satisfying

$$N \geq K, N \geq n.$$

Remark 13.1.12 (interpretation)**.** $p(x=k)$ describes the probability of k successes in n draws, without replacement, from a finite population of size N that contains exactly K successes.

Lemma 13.1.28 (combinatorial identities). *Assuming $K \geq n$, we have*

$$\sum_{0 \leq k \leq n} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} = 1$$

Lemma 13.1.29 (mean of a hypergeometric distribution). *[1, p. 148] Let X be a random variable with $HG(N, K, n)$, then its mean is*

$$E[X] = n\frac{K}{N}$$

13.1.13 Beta distribution

Definition 13.1.15 (Beta distribution). *[4, p. 43] A random variable X is said to have a Beta distribution $B(a, b)$ with parameter a, b if it has a pdf given as*

$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

with support $x \in [0, 1]$.

Remark 13.1.13.

- $$\int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$
- Beta distribution is commonly **used as the conjugate prior for binomial distribution**, where
  $$p(y_1, ..., y_n|\theta) = \theta^{\sum_i y_i}(1-\theta)^{n-\sum_i y_i}, y_i \in \{0, 1\}$$
  then the posterior distribution will also be Beta.

Lemma 13.1.30 (basic property). *Let X be a random variable with distribution $B(a, b)$.*

- $$E[X] = \frac{a}{a+b}.$$
- $$E[X^2] = \frac{a(a+1)}{(a+b)(a+b+1)}.$$

- $$E[X^r] = \frac{a(a+1)\cdots(a+r-1)}{(a+b)(a+b+1)\cdots(a+b+r-1)}.$$
- $$Var[X] = \frac{ab}{(a+b)^2(a+b+1)}.$$
- *The mode of X, i.e., the value x that has the maximum probability is*

$$x^* = \frac{a-1}{a+b-2}.$$

Proof. (1)This can be proved using properties of Gamma distribution.

$$\begin{aligned}
E[X] &= \int_0^1 xf(x)dx \\
&= \int_0^1 \frac{x^a(1-x)^{b-1}}{B(a,b)} \\
&= \frac{B(a+1,b)}{B(a,b)} \\
&= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \Big/ \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \\
&= \frac{\Gamma(a+1)\Gamma(a+b)}{\Gamma(a+b+1)\Gamma(a)} \\
&= \frac{a}{a+b}
\end{aligned}$$

(2)

$$\begin{aligned}
E[X^2] &= \int_0^1 x^2 f(x)dx \\
&= \int_0^1 \frac{x^{a+1}(1-x)^{b-1}}{B(a,b)} \\
&= \frac{B(a+2,b)}{B(a,b)} \\
&= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} \Big/ \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \\
&= \frac{\Gamma(a+2)\Gamma(a+b)}{\Gamma(a+b+2)\Gamma(a)} \\
&= \frac{a(a+1)}{(a+b)(a+b+1)}
\end{aligned}$$

(3) Use $Var[X] = E[X^2] - E[X]^2$. (4) To find the maximizer for $x^{a-1}(1-x)^{b-1}$, we take the log and maximize it. We have

$$\ln f(x) = (a-1)\ln x + (b-1)\ln(1-x).$$

Take the derivative with respect to x and set to 0, we have

$$\begin{aligned} \frac{a-1}{x} &= \frac{b-1}{1-x} \\ (a-1)(1-x) &= x(b-1) \\ \implies x^* &= \frac{a-1}{a+b-2}. \end{aligned}$$

□

13.1.14 Multinomial distribution

Definition 13.1.16. *[4, p. 35] A discrete random vector $X = (X_1, ..., X_n)$ is said to have multinomial distribution with parameters $(p_1, ..., p_n)$ and m if its pmf is given as*

$$f(x_1, x_2, ..., x_n) = \frac{m!}{x_1!...x_n!} p_1^{x_1} ... p_n^{x_n}$$

where we require $x_i \in \{0, ..., m\}, \sum x_i = m, \sum p_i = 1$.

Remark 13.1.14. Consider m independent experiment, each has n outcomes with probability p_i to occur. The outcome distribution is given as[5]

$$f(x_1, x_2, ..., x_n) = \frac{m!}{x_1!...x_n!} p_1^{x_1} ... p_n^{x_n}$$

where $\sum x_i = m, \sum p_i = 1$.

Lemma 13.1.31 (basic properties of Multinomial Distribution). *Let $X = (X_1, ..., X_n)$ discrete random vector with multinomial distribution with parameters $p = (p_1, ..., p_n)$ and m.*

-
- $$E[X_i] = np_i.$$
- $$Var[X_i] = np_i(1-p_i), Cov(X_i, X_j) = np_i(1-p_i),$$

or in vector form

$$Var[X] = n(diag(p) = pp^T).$$

Proof. (1)This can be proved using properties of Gamma distribution.

$$\begin{aligned} E[X_i] &= \int_0^1 x_i f(x) dx \\ &= \frac{\prod_{k=1}^K \Gamma(a_k + \delta_{ik})}{\Gamma(a_0 + 1)} / \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(a_0)} \\ &= \frac{a_i}{a_0} \end{aligned}$$

(2)

$$\begin{aligned} E[X_i^2] &= \int_0^1 x_i^2 f(x) dx \\ &= \frac{\prod_{k=1}^K \Gamma(a_k + 2\delta_{ik})}{\Gamma(a_0 + 2)} / \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(a_0)} \\ &= \frac{a_i(a_i + 1)}{(a_0 + 1)a_0} \end{aligned}$$

(3) Use $Var[X] = E[X^2] - E[X]^2$. (4) To find the maximizer for $f(x)$, we take the log and maximize it. The optimality condition requires that $x_i^* \propto a_i - 1$ and $\sum_{i=1}^K a_i = 1$. □

13.1.15 Dirichlet distribution

Definition 13.1.17. *[4, p. 49] A random vector $X = (X_1, ..., X_K)$ is said to have a Dirichlet distribution distribution with parameter $a = (a_1, ..., a_K)$ if it has pdf given as*

$$f(x_1, ..., x_K) = \frac{1}{B(a)} \prod_{k=1}^K x_k^{a_k - 1}$$

with support $x \in \{x : 0 \leq x_k \leq 1, \sum_k x_k = 1, \forall k = 1, 2..., K\}$, and $B(a)$ is a normalization constant given as

$$B(a) = \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(\sum_k a_k)},$$

where $\Gamma(\cdot)$ is the Gamma function.

Remark 13.1.15.

- Dirichlet distribution can be viewed as multivariate generalization of Beta distribution.
- Dirichlet distribution is usually **used as the conjugate prior** for multinomial distribution.

Lemma 13.1.32 (basic properties of Dirichlet Distribution). *Let* $X = (X_1, X_2, ..., X_K), x_i \in (0,1), \sum_{i=1}^K x_i = 1$, *be a random vector with distribution* $B(a), a \in \mathbb{R}^K$. *Let* $a_0 = \sum_{i=1}^K a_i$.

- $$E[X_i] = \frac{a_i}{\sum_{i=1}^K a_i}.$$
- $$E[X_i^2] = \frac{a_i(a_i+1)}{(a_0)(a_0+1)}.$$
- $$Var[X_i] = \frac{a_i(a_0-a_i)}{a_0^2(a_0+1)}.$$
- *The mode of X, i.e., the value x that has the maximum probability is*

$$x_i^* = \frac{a_i - 1}{a_0 - K}.$$

Proof. (1)This can be proved using properties of Gamma distribution.

$$\begin{aligned} E[X_i] &= \int_0^1 x_i f(x) dx \\ &= \frac{\prod_{k=1}^K \Gamma(a_k + \delta_{ik})}{\Gamma(a_0+1)} / \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(a_0)} \\ &= \frac{a_i}{a_0} \end{aligned}$$

(2)

$$\begin{aligned} E[X_i^2] &= \int_0^1 x_i^2 f(x) dx \\ &= \frac{\prod_{k=1}^K \Gamma(a_k + 2\delta_{ik})}{\Gamma(a_0+2)} / \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(a_0)} \\ &= \frac{a_i(a_i+1)}{(a_0+1)a_0} \end{aligned}$$

(3) Use $Var[X] = E[X^2] - E[X]^2$. (4) To find the maximizer for $f(x)$, we take the log and maximize it. The optimality condition requires that $x_i^* \propto a_i - 1$ and $\sum_{i=1}^K a_i = 1$. □

13.1.16 χ^2-distribution

13.1.16.1 *Basic properties*

Definition 13.1.18. *A random variable X is said to have a $\chi^2(n)$ distribution with parameter $n \in \mathbb{Z}_+$ if it has pdf given as*

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-y/2}$$

with $x \in (0, +\infty)$

Remark 13.1.16 (special case of Gamma distribution). $\chi^2(n)$ has the same distribution of Gamma(n/2,2).

Definition 13.1.19 (alternative). *The χ^2-distribution with k degrees of freedom is the distribution of a sum of squares of k independent standard normal random variables. Mathematically, if $X_1, X_2, ..., X_k$ are iid random variable with $X_i \sim N(0,1)$, the random variable*

$$Q = \sum_{i=1}^{k} X_i^2$$

is distributed according to the χ^2 distribution with k degrees of freedom, writen as $Q \sim \chi^2(k)$.

Lemma 13.1.33 (basic property). *[1, pp. 161–163] Let X_1, X_2 be independent random variables. Suppose $X_1 \sim \chi^2(a_1), X_2 \sim \chi^2(a_2)$. Then*

- $Y = X_1 + X_2 \sim \chi^2(a_1 + a_2)$
- $\lambda X_1 \sim \lambda^2 \chi^2(a_1)$
- *The moment generating function is given by*

$$M(t) = (1 - 2t)^{-r/2}.$$

Proof. (1)This can be proved using properties of Gamma distribution. (2) λX_1 can be viewed as the sum of squares of normal random variables Y_i with $N(0, \lambda^2)$. Then $\sum_{i=1}^{n} (Y_i/\lambda)^2 \sim \chi^2(n)$.

□

Lemma 13.1.34 (expectation and variance). *Let random variable X has distribution of $\chi^2(n)$, then*

$$E[X] = n, Var[X] = 2n$$

In particular,

$$E[X/n] = 1, Var[X/n] = 0 \text{ as } n \to \infty.$$

that is the random variable X/n becomes deterministic constant as $n \to \infty$.

Proof. (1)Let $Z \sim \chi^2(1), Z = Y^2, Y \sim N(0,1)$, then $E[Z] = Var[Y] + (E[Y])^2 = 1$. $Var[Z] = E[Z^2] - (E[Z])^2 = E[Y^4] - 1 = 3 - 1 = 2$. (2) Use linearity of expectation that $E[X/n] = E[X]/n = 1$. Use $Var[X/n] = Var[X]/n^2 = 2/n$. □

13.1.16.2 *Quadratic forms and chi-square distribution*

Definition 13.1.20 (quadratic forms of random vectors). *[1, p. 485] Let $X = (X_1, X_2, ..., X_n)^T$ be a random vector, we called*

$$Q = X^T \Sigma X, \Sigma \in \mathbb{R}^{n \times n},$$

a quadratic form of random vector X.

Note that Q is also a random variable.

Lemma 13.1.35. *Let X be a m-dimensional random vector with multivariate Gaussian distribution, i.e., $X \sim N(\mu, \Sigma)$. It follows that*

- $$\Sigma^{1/2}(x - \mu) \sim N(0, I).$$
- $$(x - \mu)\Sigma^{-1}(x - \mu) \sim \chi^2(m).$$

Proof. (1) Directly from affine transformation property of multivariate Gaussian random variable [Theorem 15.1.1]. (2) Use the definition that sum of iid normal random variable square is chi-square random variable. □

Theorem 13.1.4 (chi-square orthogonal decomposition). *Let $X_1, X_2, ..., X_n$ be independent standard normal variables such that*

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n).$$

Denote $X = (X_1, ..., X_n)^T$. If there exists an orthogonal projector $P \in \mathbb{R}^{n\times n}$ such that $Y = PX, Z = (I-P)X$, then

- *$Y \sim MN(0, P), Z \sim MN(0, I-P)$, and Y, Z are independent of each other.*
- *$Y^TY \sim \chi^2(r), r = rank(P)$; or equivalently, the quadratic form $Q = X^TPX \sim \chi^2(r)$.*
- *$Z^TZ \sim \chi^2(n-r)$;or equivalently, the quadratic form $Q = X^T(I-P)X \sim \chi^2(n-r)$*

In summary, for a quadratic form $Q = X^T\Sigma X$, if Σ is idempotent and symmetric, then $Q \sim \chi^2(rank(\Sigma))$.

Proof. (1) From affine transform of multivariate normal [Theorem 15.1.1],

$$Y \sim MN(0, P\Sigma_X P^T) = MN(0, P^2) = MN(0, P).$$

To show independence, we have $E[YZ^T] = E[PXX^T(I-P)^T] = E[P(I-P)] = 0$.

(2) Let U be the eigen-decomposition of P such that $P = UU^T$. Let $Z = U^TX, Z \in \mathbb{R}^r, Z \sim MN(0, I_r)$. Let V be the eigen-decomposition of $I-P$ such that $I-P = VV^T$. Let $W = V^TX, W \in \mathbb{R}^{n-r}, W \sim MN(0, I_{n-r})$. We want to show that the characteristic function of the random quantity Y^TY is the same as the characteristic function of $\chi^2(r)$.

$$\begin{aligned}
&E[\exp(itY^TY)] \\
=&\frac{1}{(2\pi)^{n/2}} \int\int\cdots\int \exp(it(U^TX)^T(U^TX)\exp(-\frac{1}{2}X^T(I-P+P)X))dx_1dx_2\cdots dx_n \\
=&\frac{1}{(2\pi)^{n/2}} \int\int\cdots\int \exp(itZ^TZ)\exp(-\frac{1}{2}(Z^TZ+W^TW))dz_1\cdots dz_rdw_{r+1}\cdots dw_n \\
=&\frac{1}{(2\pi)^{r/2}} \int\int\cdots\int \exp(itZ^TZ)\exp(-\frac{1}{2}(Z^TZ))dz_1\cdots dz_r
\end{aligned}$$

where we change the integral variable such that

$$[dz_1\cdots dz_rdw_{r+1}\cdots dw_n]^T = [U\ V](dx_1dx_2\cdots dx_n)^T$$

. The last line is the characteristic function of $\chi^2(r)$. (3) similar to (2). □

Lemma 13.1.36 (moment generating functions for Gaussian quadratic forms). *[1, p. 523] Let $X = (X_1, X_2, ..., X_n)^T$ where $X_1, X_2, ..., X_n$ are iid $N(0,1)$. Consider the quadratic form $Q = X^TAX$ for a symmetric matrix A of rank $r \leq n$. It follows that*

- *Q has the moment generating function $M(t) = \prod_{i=1}^{r}(1-2t\lambda_i)^{-1/2} = |I-2tA|^{-1/2}$, where $\lambda_1, \lambda_2, ..., \lambda_r$ are the nonzero eigenvalues of A, $|t| < \frac{1}{\max|\lambda_i|}$.*
- *If A is an orthogonal projector such that $\lambda_1 = \lambda_2 = \cdots = \lambda_r = 1$, then*

$$M(t) = M_{\chi^2(r)}.$$

Proof. (1)Let the eigen-decomposition of A be

$$A = U\Lambda U^T, U \in \mathbb{R}^{n\times r}, \Lambda \in \mathbb{R}^{r\times r}.$$

Then

$$Q = X^TAX = X^TU\Lambda U^TX = X^T(\sum_{i=1}^{r}\lambda_i u_i u_i^T)X = \sum_{i=1}^{r}\lambda_i(u_i^TX)^T.$$

Let $Y_i = u_i^TX, i = 1, 2, ..., r$. It can be shown that $Y_i \sim N(0,1), E[Y_iY_j] = u_i^TE[XX^T]u_j^T = \delta_{ij}$; that is $Y_1, Y_2, ..., Y_r \sim MN(0, I_r)$. Therefore, $Y_i^2 \sim \chi^2(1)$.

The moment generating function is given by

$$\begin{aligned}
M(t) &= E[\exp(tQ)] \\
&= E[\exp(t\sum_{i=1}^{r}\lambda_i Y_i^2)] \\
&= \prod_{i=1}^{r}E[\exp(t\lambda_i Y_i^2)] \\
&= \prod_{i=1}^{r}M_{\chi^2(1)}(\lambda_i t) \\
&= \prod_{i=1}^{r}(1-2\lambda_i t)^{-1/2}
\end{aligned}$$

where we use the moment generating function of $\chi^2(1)$ from Lemma 13.1.33. (2) straight forward. □

Lemma 13.1.37 (independence of quadratic forms). *[1, p. 528] Let $X = (X_1, X_2, ..., X_n)$ be a random vector where $X_1, X_2, ..., X_n$ are iid $N(0,1)$. For real symmetric matrices $A, B \in \mathbb{R}^{n\times n}$, let $Q_1 = X^TAX$ and $Q_2 = X^TBX$. Then Q_1 and Q_2 are independent if and only if $AB = 0$.*

Proof. Let $rank(A) = r, rank(B) = s$. Let the eigendecomposition of A, B be such that

$$A = \sum_{i=1}^{r}\lambda_i u_i u_i^T, A = \sum_{i=1}^{s}\beta_i v_i v_i^T.$$

If $AB = 0$, then $u_1, ..., u_r, v_1, ..., v_r$ will be orthogonal to each other. Then

$$Q_1 + Q_2 = \sum_{i=1}^{r+s} \lambda_i u_i u_i^T,$$

where $u_{r+i} = v_i, \lambda_{r+i} = \beta_i$.

It is easy to see that [Lemma 13.1.36]

$$M_{Q_1,Q_2}(t_1, t_2) = M_{Q_1}(t_1) M_{Q_2}(t_2).$$

Then from independence-from-mgf [Lemma 12.4.5], we can prove Q_1 and Q_2 are independent.

□

13.1.16.3 *Noncentral chi-squared distribution*

Definition 13.1.21 (noncentral chi-squared distribution). *Let $(X_1, X_2, ..., X_k)$ be k independent, normally distributed random variables with mean μ_i and unit variances. Then the random variable*

$$Y = \sum_{i=1}^{k} X_i^2$$

*is distributed according to the **noncentral chi-squared distribution** with parameter k specifying the degree of freedom and λ, known as the **noncentrality parameter**, given by*

$$\lambda = \sum_{i=1}^{k} \mu_i^2.$$

13.1.17 Wishart distribution

Definition 13.1.22 (Wishart distribution). *Let $X_1, ..., X_n$ be independent p dimensional multivariate normal random vector with distribution $MN(0, V)$. Let $X = [X_1, ..., X_n]$. Then $M = XX^T$ is said to have Wishart distribution with parameter (n, p, V).*

Definition 13.1.23 (Wishart distribution). *A random matrix $M \in \mathbb{R}^{p \times p}$ is said to have the Wishart distribution with parameters $W_p(n, V)$ if it has pdf*

$$f(M) = \frac{1}{2^{np/2} \Gamma_p(\frac{n}{2}) |V|^{n/2}} |M|^{n-p-1/2} \exp(\frac{1}{2} Tr[V^{-1} M]),$$

with the support M be the set of all symmetric positive definite matrices. Here $\Gamma_p(\alpha)$ is the multivariate gamma function.

Lemma 13.1.38 (basic properties).

- *(reduction to χ^2) If $M \in \mathbb{R}^{1\times 1}$, then*

$$M \sim W_1(n, \sigma^2) = \sigma^2 \chi^2(n).$$

- *For $M \sim W_p(n, V)$, then $B^T M B \sim W_m(n, B^T V B)$, where $B \in \mathbb{R}^{p\times m}$.*
- *For $M \sim W_p(n, V)$, then $V^{-1/2} M V^{-1/2} \sim W_m(n, I)$.*
- *If M_i are independent $W_p(n_i, V)$, then $\sum_{i=1}^k \sim W_p(\sum_{i=1}^k n_i, V)$.*
- *If $M \sim W_p(n, V)$, then $E[M] = nV$.*
- *If M_1, M_2 are independent and $M_1 + M_2 = M \sim W_p(n, V)$. Further if $M_1 \sim W_p(n_1, V)$, then $M_2 \sim W_p(n - n_1, V)$.*

Lemma 13.1.39 (sample covariance). *The sample covariance*

$$\hat{Cov} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^T$$

where X_i are iid $MN(0, V)$, and $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, has the property of

$$E[\hat{Cov}] = V.$$

13.1.18 t-distribution

13.1.18.1 *Standard t distribution*

Definition 13.1.24 (t distribution). *[1, p. 192] A random variable X is said to have a $t(n)$ distribution with parameter $n \in \mathbb{Z}_+$ if it has pdf given as*

$$f(x) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}} (1 + \frac{t^2}{n})^{-(n+1)/2}$$

with $x \in (-\infty, +\infty)$

Definition 13.1.25 (alternative). *Let random variable $W \sim N(0,1)$, Let random variable $V \sim \chi^2(n)$* ***independent of*** *V. Define a new random variable T as*

$$T = \frac{W}{\sqrt{V/n}}$$

Then T has a t-distribution with degree of freedom n, denoted by T_n or t_n.

Remark 13.1.17 (comparison with normal distribution).

- t distribution generally have shorter peak and fatter tails than normal distribution.
- $t_n \to N(0,1)$ as $n \to \infty$.

Lemma 13.1.40 (mean and variance of t-distribution). *The mean for a t-distribution with degree of n is given by*

$$E[t_n] = \begin{cases} 0, n > 1 \\ \infty(undefined), n = 1 \end{cases}.$$

The variance for a t-distribution with degree of n is given by

$$Var[t_n] = \begin{cases} \frac{n}{n-2}, n > 2 \\ \infty, n = 1,2 \end{cases}.$$

13.1.18.2 *classical t distribution*

Definition 13.1.26. *[6, p. 95] If Y has a standard t_n distribution, then*

$$Z = \mu + \lambda Y$$

is said to have a $t_n(\mu, \lambda^2)$ distribution.

Lemma 13.1.41 (mean and variance of classical t-distribution). *Let Z be a random variable $t_n(\mu, \lambda^2)$. Then*

$$E[Z] = \begin{cases} \mu, n > 1 \\ \infty(undefined), n = 1 \end{cases},$$

and

$$Var[Z] = \begin{cases} \lambda^2 \frac{n}{n-2}, n > 2 \\ \infty, n = 1, 2 \end{cases}.$$

13.1.18.3 *Multivariate t distribution*

Definition 13.1.27 (multivariate t distribution). *[6]*

- *Let Z be a d dimensional multivariate Gaussian $MN(0, \Sigma)$, and $\mu \in \mathbb{R}^d$. The d dimensional random vector Y, defined as,*

$$X = \mu + \sqrt{\frac{n}{W}} Z,$$

  *where $W \sim \chi^2(n)$ and W is **independent** of Z, has a $t_n(\mu, \Sigma)$ multivariate distribution.*
- *Let $X \sim t_n(\mu, \Sigma)$. Then X has the density given by*

$$f(x) = \frac{\Gamma((n+d)/2)}{\Gamma(n/2) n^{d/2} \pi^{d/2} |\Sigma|^{1/2}} (1 + \frac{1}{n}(x-\mu)^T \Sigma^{-1} (x-\mu))^{-(n+d)/2}.$$

Lemma 13.1.42 (mean and variance of multivariate t-distribution). *Let Z be a random variable $t_n(\mu, \Sigma)$. Then*

$$E[Z] = \begin{cases} \mu, n > 1 \\ \infty(undefined), n = 1 \end{cases},$$

and

$$Cov[Z] = \begin{cases} \Sigma \frac{n}{n-2}, n > 2 \\ \infty, n = 1, 2 \end{cases}.$$

13.1.18.4 *Student's Theorem*

Theorem 13.1.5 (Student's Theorem). *[1, p. 194] Let $X_1, X_2, ..., X_n$ be iid random variables each having a normal distribution with mean μ and variance σ^2. Define random variables as:[1]*

$$\bar{X} = \frac{1}{n}\sum_{i=0}^{n} X_i, S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

1. *$\bar{X}$ has a $N(\mu, \sigma^2/n)$ distribution*
2. *$\bar{X}$ and S^2 are independent.*
3. *$(n-1)S^2/\sigma^2$ has a $\chi^2(n-1)$ distribution*
4. *The random variable*

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has t-distribution with $n-1$ degrees of freedom.

Proof. (1)From Lemma 13.1.3. (2) We can prove $\overline{X}$ and the random vector $Y = (X_1 - \overline{X}, ..., X_n - \overline{X})$ are independent. Note that

$$\overline{X} = \frac{1}{n}\mathbf{1}^T X, Y = (I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)X,$$

and hence $\overline{X}$ and Y are both normal.

$$\begin{aligned} Cov(\overline{X}, Y) &= X^T(\frac{1}{n}\mathbf{1}^T(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T))X \\ &= X^T\frac{1}{n}(\mathbf{1}^T - \frac{1}{n}\mathbf{1}^T\mathbf{1}\mathbf{1}^T)X \\ &= X^T\frac{1}{n}(\mathbf{1}^T - \mathbf{1}^T))X = 0 \end{aligned}$$

where we use the fact that $\mathbf{1}^T\mathbf{1} = n$.

Then $S^2 = \frac{1}{n-1}Y^TY$ will be independent of $\overline{X}$ because S^2 is a function of Y [Lemma 12.3.3]. (3) See reference and Corollary 13.3.1.1. (4) From the definition of the t distribution, we have

$$Y = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is the $N(0,1)$. $W = (n-1)S^2/\sigma^2$ has a $\chi^2(n-1)$ distribution. Then

$$\frac{Y}{\sqrt{W/n-1}}$$

has $t(n-1)$ distribution. □

13.1.19 F-distribution

Definition 13.1.28 (F distribution). *[1, p. 192] A random variable X is said to have a $F(n_1, n_2)$ distribution with parameter $n_1, n_2 \in \mathbb{Z}_+$ if it has pdf given as*

$$f(x) = \frac{\Gamma((n_1+n)2)/2)(n_1/n_2)^{n_1/2}y^{n_1/2-1}}{\Gamma(n_1/2)\Gamma(n_2/2)[1+(n_1x/n_2)]^{(n_1+n_2)/2}}$$

with $x \in (0, +\infty)$

Definition 13.1.29 (alternative). *Given two **independent** chi-squared random variables W and V having r_1 and r_2 degrees of freedom. We define a new random variable*

$$W = \frac{U/r_1}{V/r_2}$$

Then W has a F-distribution with parameter (r_1, r_2).

Lemma 13.1.43 (inverse relationship). *Let X be a random variable with distribution $F(n_1, n_2)$, then $1/X$ is a random variable with distribution $F(n_2, n_1)$.*

Proof. Directly from definition. □

Lemma 13.1.44 (relationship to t distribution). *Let X be a random variable with standard t distribution with n degrees of freedom. Then*

$$X^2 \sim F(1, n).$$

That is, X^2 has the distribution of $F(1, n)$.

Proof. Directly from definition. □

Definition 13.1.30 (noncentral F distribution). *Given two chi-squared random variables W and V such that V is a noncentral chi-squared random variable with non-centrality parameter λ and degree of freedom r_1, and W is a chi-squared random variable having r_1 r_2 degrees of freedom. We define a new random variable*

$$W = \frac{U/r_1}{V/r_2}$$

Then W has a noncentral F-distribution with parameter (λ, r_1, r_2). .

13.1.20 Empirical distributions

Definition 13.1.31 (empirical cumulative distribution function(CDF)). *Given N iid random variables* $Y_1, Y_2, ..., Y_N$ *with common cdf* $F(t)$*, the empirical CDF is defined by*

$$\hat{F}_N(t) = \frac{number\ of\ elements\ in\ the\ sample \leq t}{N} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{Y_i \leq t}$$

.

Lemma 13.1.45 (basic statistic properties). *Let* $\hat{F}_N(t)$ *be the empirical cdf of a random sample of size N. For a fixed t, we have*

- $N\hat{F}_N(t)$ *is a binomial random variable with parameter* (N, p)*,where* $p = F(t)$.
- $N\hat{F}_N(t)$ *is an unbiased estimator for* $NF(t)$.
- $N\hat{F}_N(t)$ *has variance* $NF(t)(1 - F(t))$.

Proof. (1) Note that based on the definition of $\hat{F}_N(t) \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{Y_i \leq t}$, $\mathbf{1}_{Y_i \leq t}$ is a Bernoulli random variable with parameter $p = F(t)$. Therefore, $N\hat{F}_N(t) = \sum_{i=1}^{N} \mathbf{1}_{Y_i \leq t}$ will follow a binomial distribution of parameter (N, p). (2)

$$E[N\hat{F}_N(t)] = Np = NF(t).$$

(3)

$$Var[N\hat{F}_N(t)] = Np(1-p) = NF(t)(1 - F(t)).$$

□

13.1.21 Heavy-tailed distributions

13.1.21.1 *Basic characterization*

Definition 13.1.32 (Heavy-tailed distribution). *The distribution of a random variable X with distribution function F is said to have a heavy right tail if*

$$\lim_{x \to \infty} e^{\lambda x} Pr(X > x) = \infty, \forall \lambda > 0.$$

Remark 13.1.18 (interpretation). Heavy-tailed distributions have densities decaying slower in the tails than the normal.

13.1.21.2 *Pareto and power distribution*

Definition 13.1.33 (Pareto distribution). *A random variable X is said to have Pareto distribution with scale parameter $x_m > 0$ and shape parameter $\alpha > 0$ if its has pdf*

$$f_X(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, x \geq x_m, \\ 0, x < x_m. \end{cases};$$

or cdf

$$f_X(x) = \begin{cases} 1 - (\frac{\alpha x_m}{x})^\alpha, x \geq x_m, \\ 0, x < x_m. \end{cases}.$$

X has support $[x_m, \infty)$.

Definition 13.1.34 (power law distribution). *A random variable X is said to have power law distribution with parameters K, α if its has probability characterization on its tail given by*

$$Pr(X > x) = Kx^{-\alpha}.$$

Remark 13.1.19 (Pareto distribution and power law distribution are heavy-tailed distribution). Note that since power grows much slower than the exponentialA.2.1, therefore

$$\lim_{x \to \infty} e^{\lambda x} Pr(X > x) = \infty, \forall \lambda > 0.$$

13.1.21.3 *Student t distribution family*

Definition 13.1.35 (Student's t-Distribution family). *The t distribution has a single parameter, $\nu > 0$, known as* degrees of freedom. *The density function is given as*

$$f_\nu(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1 + \frac{x^2}{\nu})^{-\frac{1}{2}(\nu+1)}$$

The first two members of family are

1. $f_1(x) = \frac{1}{\pi(1+x^2)}$

2. $f_2(x) = \frac{1}{2\sqrt{2}}(1 + x^2/2)^{-3/2}$

The $\nu = 1$ density is known as Cauchy's density. *As $\nu \to \infty$, the density distribution tends to the standard normal density.*

Definition 13.1.36 (Cauchy distribution). *The Cauchy distribution with parameter (x_0, γ) has the probability density function*

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{(x - x_0)^2 + \gamma^2}\right),$$

where x_0 is the location parameter, specifying the location of the peak of the distribution, and γ is the scale parameter which specifies the half-width at half-maximum. ***Standard Cauchy distribution*** *is Cauchy distribution with parameter $(0, 1)$.*

Remark 13.1.20 (nonexistence of moments)**.**

- The Cauchy distribution is an example of a distribution which has no mean, variance or higher moments. And therefore the moment generating function does not exist. However, the **mode and median** are well defined and both equal to x_0.
- The nonexistence of expectation is because of the $E[|X|] < \infty$.

Lemma 13.1.46 (sum of Cauchy distribution). *If $X_1, ..., X_n$ are independent and identically distributed random variables, each with a standard Cauchy distribution, then the sample mean*

$$\overline{X} = (X_1 + ... + X_n)/n$$

has the same standard Cauchy distribution.

Proof. Note that we need to use characteristic function to prove, since the moment generating function does not exist. □

13.1.21.4 *Gaussian mixture distributions*

Definition 13.1.37 (normal scale mixture distribution). *[6, p. 99] The normal scale mixture distribution is the distribution of the random variable*

$$Y = \mu + \sqrt{U}Z,$$

where μ is constant equal to the mean, and $Z \sim N(0, 1)$, U is a positive random variable giving the variance of each component, and Z and U are independent.

*If U can assume only a finite number of values, then Y has a **discrete scale mixture distribution**. If U is continuously distributed, then Y has a **continuous scale mixture distribution**.*

Example 13.1.7 (discrete Gaussian mixture distribution). Let $\mu = 0$, and U have the following distribution

$$P(U = 25) = 0.1, P(U = 1) = 0.9.$$

Then

$$Y = \mu + \sqrt{U}Z,$$

is the mixture of 10% of $N(0, 25)$ and 90% of $N(0, 1)$.

Example 13.1.8 (t distribution). The t_n distribution with n degrees of freedom is a continuous Gaussian mixture with

$$\mu = 0, U = \frac{n}{W},$$

where $W \sim \chi^2(n)$.

Definition 13.1.38 (multivariate normal variance mixtures). *The random vector X has a multivariate normal variance mixture distribution if*

$$X \triangleq \mu + \sqrt{W}AZ$$

where

- $Z \sim MN(0, I_k)$
- *W is a **positive** scalar random variable which is independent of Z*
- *$A \in \mathbb{R}^{d \times k}$ and $\mu \in \mathbb{R}^d$ are a matrix and a vector of constants*

Remark 13.1.21 (conditional distribution).

Example 13.1.9 (special case: multivariate t distribution). The t_n distribution with n degrees of freedom is a continuous Gaussian mixture with

$$\mu = 0, U = \frac{n}{W},$$

where $W \sim \chi^2(n)$.

13.2 Characterizing distributions

13.2.1 Skewness and kurtosis

Definition 13.2.1 (skewness). *The skewness of an univariate population for random variable X is defined by*

$$\gamma_1 = E[(\frac{X-\mu}{\sigma})^3] = \frac{E[(X-\mu)^3]}{(E[(X-\mu)^2])^{3/2}} = \frac{\mu_3}{\mu_2^{3/2}}$$

*where μ_2 and μ_3 are the second and the third **central moments**.*

Remark 13.2.1 (interpretation).

- Intuitively, the skewness is a measure of symmetr y[Figure 13.2.1].
- **Negative skewness** indicates that the mean of the data values is less than the median, and the data distribution is **left-skewed**.
- **Positive skewness** indicates that the mean of the data values is greater than the median, and the data distribution is **right-skewed**.

Example 13.2.1. Let $X \sim N(\mu, \sigma^2)$. Then the skewness of X distribution is $\gamma_1 = E[(\frac{X-\mu}{\sigma})^3] = E[Z^3] = 0, Z \sim N(0,1)$, where we use the fact the third moment for a standard normal is zero [Lemma 13.1.4].

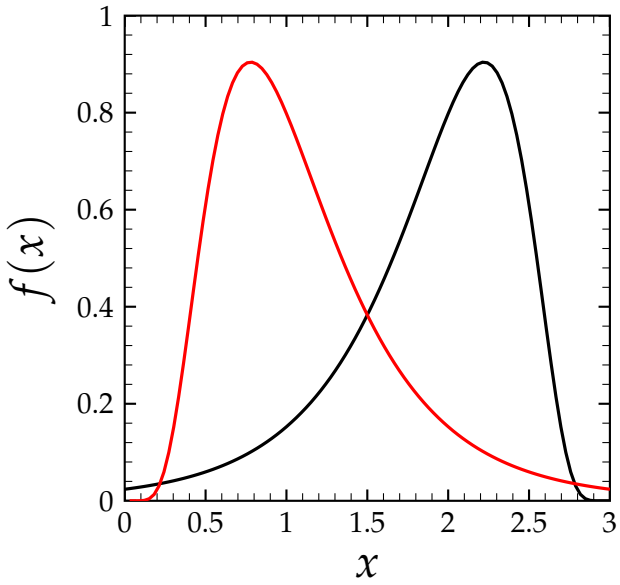


Figure 13.2.1: Distributions with left-skewness (black) and right-skewness (red).

Definition 13.2.2 (Kurtosis, excess Kurtosis).

- *The **kurtosis** of a univariate population is defined by*

$$\gamma_2 = E[(\frac{X-\mu}{\sigma})^4] = \frac{E[(X-\mu)^4]}{(E[(X-\mu)^2])^2} = \frac{\mu_4}{\mu_2^2},$$

  where μ_2 and μ_4 are the second and the fourth central moments.
- *The **excess Kurtosis** of a univariate population is defined by*

$$\gamma_2^{ex} = \gamma_2 - 3.$$

Remark 13.2.2 (interpretation).

- Intuitively, the Kurtosis is a measure of tail shape of a distribution.

Remark 13.2.3 (Three types of Kurtosis).

- **mesokurtic**: zero excess kurtosis as standard normal distribution.
- **leptokurtic**: excess kurtosis greater than 0. This type of distribution is one with extremely thick tails and a very thin and tall peak. t-distributions and Laplace distributions are leptokurtic.
- **platykurtic**:excess kurtosis smaller than 0. This type of distribution has a short and broad-looking peak. Uniform distributions are platykurtic.

See Figure 13.2.2 for illustrations.

Example 13.2.2. Let $X \sim N(\mu, \sigma^2)$. Then the Kurtosis of X distribution is $\gamma_2 = E[(\frac{X-\mu}{\sigma})^4] = E[Z^4] = 3, Z \sim N(0,1)$, where we use the fact the fourth moment for a standard normal is 3 [Lemma 13.1.4].

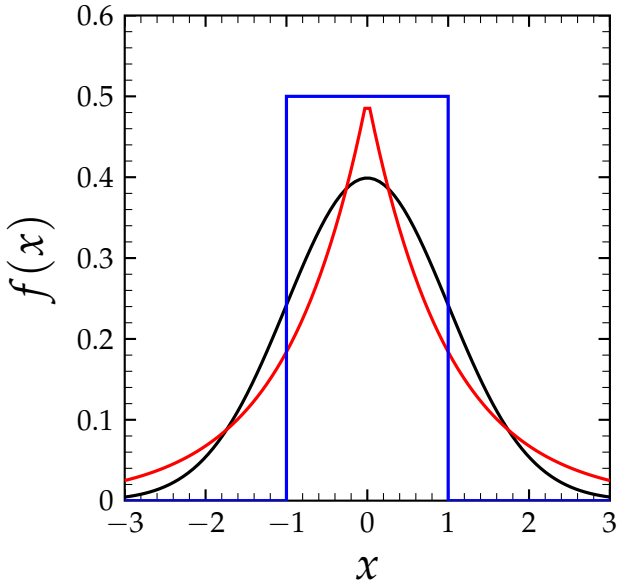


Figure 13.2.2: Distributions with zero excess Kurtosis (Normal distribution, black), positive excess kurtosis (Laplace distribution, red), and negative excess Kurtosis (Uniform distribution, blue).

13.2.2 Quantiles and percentiles

13.2.2.1 *Basics*

Definition 13.2.3 (percentile of a distribution). *The α **percentile**($\alpha \in [0,1]$) of a probability distribution of random number X is a number p in the support D of the support such that*

$$Pr(x < p) = \alpha, Pr(x > p) = 1 - \alpha.$$

Or equivalently, the α percentile is given by

$$p = F_X^{-1}(\alpha).$$

Definition 13.2.4 (percentile in a set of sample values). *The α **percentile**($\alpha \in [0,1]$) of a set of values is a value in $\mathbb{R}$ that divides them so that $100\alpha\%$ of values lie below and $100(1-\alpha)\%$ of the values lie above.*

Definition 13.2.5 (quantiles of a distribution). *Quantiles are the cutpoints dividing the range of a probability distribution into contiguous intervals with equal probabilities.*

Lemma 13.2.1 (linear relationship between percentiles from two distributions). *Let X and Y be two random variables with cdf F_X and F_Y. Let $p_X = F_X^{-1}(\alpha)$ and $p_Y = F_Y^{-1}(\alpha)$ for $\alpha \in [0,1]$. It follows that*

- *If $Y = aX + b$, then*
$$p_Y = ap_X + b$$
- *If $Y = \alpha X^\beta$, then*
$$p_{\ln Y} = \beta p_{\ln X} + \ln \alpha,$$
where $p_{\ln Y} = F_{\ln Y}^{-1}(\alpha), p_{\ln X} = F_{\ln X}^{-1}(\alpha)$,

Proof. (1) We know that
$$\alpha = F_Y(p_Y) = F_X(p_X).$$
From scale-location transformation [Lemma 12.4.8], we have
$$p_X = (p_Y - b)/a.$$
(2) From $Y = \alpha X^\beta$, we have $\ln Y = \beta \ln X + \ln \alpha$. □

Remark 13.2.4 (QQplot and applications).

- When we plotting the percentiles from two samples together, an approximate linear relation suggests $Y = aX + b$.
- When we plotting the percentiles from two log-value samples together, an approximate linear relation suggests a power relationship $Y = \alpha X^\beta$.

13.2.2.2 *Cornish-Fisher expansion*

Theorem 13.2.1 (Cornish-Fisher expansion). *Consider a distribution with mean μ and variance σ^2. Then its α quantile can be approximate by*
$$\mu + \sigma z_\alpha^{cf}$$

where

$$z_\alpha^{cf} = q_\alpha + \frac{(q_\alpha^2 - 1)S(X)}{6} + \frac{(q_\alpha^3 - 3q_\alpha)K(X)}{24} - \frac{(2q_\alpha^3 - 5q_\alpha)S^2(X)}{36},$$

where $S(X)$ is skewness, $K(X)$ is kurtosis, z_α^{cf} is the Cornish-Fisher approximate quantile value for the confidence level α, and q_α is the quantile value for the standard normal distribution with confidence level α.

Remark 13.2.5 (motivation). The Cornish-Fisher expansion enables us to approximate quantiles of a random variable based only on its skewness and cumulants.

13.2.3 Exponential families

Definition 13.2.6. *[5, p. 111] A family of pdfs or pmfs is called an exponential family if it can be expressed as*

$$f(x|\theta) = h(x)c(\theta)\exp(\sum_{i=1}^{k} w_i(\theta)t_i(x)), \theta \in \mathbb{R}^n$$

where $h(x) \geq 0$ and $t_1(x), ..., t_k(x)$ are real-valued functions of the observations, the $c(\theta) \geq 0$ and $w_1(\theta), ..., w_k(\theta)$ are real-valued functions of the vector θ.

Theorem 13.2.2 (mean and variance of exponential family). *[5, p. 112] The mean and variance of the exponential family is*

$$E[\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(X)] = -\frac{\partial \log(c(\theta))}{\partial \theta_j} \tag{8}$$

$$Var[\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(X)] = -\frac{\partial^2 \log(c(\theta))}{\partial \theta_j^2} - E[\sum_{i=1}^{k} \frac{\partial^2 w_i(\theta)}{\partial \theta_j^2} t_i(X)] \tag{9}$$

Proof: see [5, p. 132] See [5, p. 622] for a complete review.

Definition 13.2.7 (alternative representation of exponential families). *A class of distributions is in the exponential family if it can be written in the form*

$$p(y;\eta) = b(\eta)\exp(\eta^T T(y) - a(\eta))$$

where η is called the natural parameter of the distribution; $T(y)$ is the sufficient statistic; and $a(\eta)$ is the log partition function.

13.3 Cochran's theorem

Lemma 13.3.1. *Let $X \sim MN(\mu, \Sigma)$ be a n dimensional random vector, then*

$$(X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi^2(n)$$

Proof. Let $Y = \Sigma^{-1/2}(X-\mu)$, then $Y \sim MN(0, I)$. Then

$$Y^TY \sim \chi^2(n).$$

□

Lemma 13.3.2. *Let $X_1, X_2, ..., X_n$ be real numbers. Suppose that $\sum_{i=1}^n X_i^2$ can be decomposed into a sum of positive semi-definite quadratic forms, that is*

$$\sum_{i=1}^n X_i^2 = Q_1 + ... + Q_k$$

where $Q_i = X^TA_iX$ with $rank(A_i) = r_i$. If $\sum_{i=1}^k r_i = n$, then there exists an orthonormal matrix C such that $X = CY$ and

$$\begin{aligned} Q_1 &= Y_1^2 + ... + Y_{r_1}^2 \\ Q_2 &= Y_{r_1+1}^2 + ... + Y_{r_1+r_2}^2 \\ &\dots \end{aligned}$$

Proof. (informal)Note that when we decompose a matrix, its sum of rank of the decomposed matrix will increase [Theorem 5.4.1], i.e.,

$$rank(A+B) \leq rank(A) + rank(B)$$

and the equality only holds when $\mathcal{R}(A) \cap \mathcal{R}(B) = \emptyset$.
Since in our case $rank(\sum A_i) = \sum rank(A_i)$, then we must have $\mathbb{R}^n = \mathcal{R}(A_1) \oplus \mathcal{R}(A_2) ... \oplus \mathcal{R}(A_k)$. Take the basis of each $\mathcal{R}(A_i)$ and make it to be orthonormal matrix C. Then Y_i are just the orthonormal projection to subspace $\mathcal{R}(A_j)$. An accessible proof is at Theorem 13.1.4 □

Theorem 13.3.1 (Cochran's theorem). *Let $X_1, X_2, ..., X_n$ be iid $N(0, \sigma^2)$ random variables. Suppose that $\sum_{i=1}^n X_i^2$ can be decomposed into a sum of positive semi-definite quadratic forms, that is*

$$\sum_{i=1}^n X_i^2 = Q_1 + ... + Q_k$$

where $Q_i = X^T A_i X$ with $rank(A_i) = r_i$. If $\sum_{i=1}^k r_i = n$, then there exists an orthonormal matrix C such that $X = CY, Y = C^T X$ ($Y_1, Y_2, ..., Y_n$ are independent random variables with $N(0, \sigma^2)$)and

$$\begin{aligned} Q_1 &= Y_1^2 + ... + Y_{r_1}^2 \\ Q_2 &= Y_{r_1+1}^2 + ... + Y_{r_1+r_2}^2 \\ &\dots \end{aligned}$$

Moreover, we have

- *$Q_1, Q_2, ..., Q_k$ are independent*
- *$Q_i \sim \sigma^2 \chi^2(r_i)$.*

Proof. (1) Use above lemma. Note that $Y_1, Y_2, ..., Y_n$ are still independent normal because of Lemma 15.1.2. (2) Since Q_i and Q_j have non-overlapping Y_is, they are independent to each other. (3) From properties of χ^2 distribution [Lemma 13.1.33]. □

Corollary 13.3.1.1 (distribution of sample variance). *Let $Y_1, ..., Y_n$ be iid random variable with $N(\mu, \sigma^2)$, then*

$$\sum_{i=1}^n (Y_i - \overline{Y})^2 \sim \sigma^2 \chi^2(n-1)$$

and

$$\sum_{i=1}^n (Y_i - \mu)^2 / n \sim \sigma^2 \chi^2(1)$$

Proof.

$$\sum_{i=1}^n (Y_i - \mu)^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2 + \sum_{i=1}^n (Y_i - \mu)^2 / n$$

And

$$(Y - \mu)^T (Y - \mu) = (Y - \mu)^T (I - \frac{1}{n} J)(Y - \mu) + (Y^T - \mu)^T (\frac{1}{n} J)(Y - \mu)$$

and $rank(\frac{1}{n} J)$ has rank 1 and $rank(I - \frac{1}{n} J) = n - 1$. □

Remark 13.3.1.

- The matrix $\frac{1}{n}J$ has rank 1 is because it only has one linearly independent column.
- The matrix $I - \frac{1}{n}J$ is because $rank(I - \frac{1}{n}J) \geq rank(I) - rank(\frac{1}{n}J) = n - 1$ [Theorem 5.4.1]. Also $I - \frac{1}{n}J$ has eigenvector 1 associated with eigenvalue 0. Therefore, $rank(I - \frac{1}{n}J) < n$. In summary, we have $rank(I - \frac{1}{n}J) = n - 1$.
- The matrix $I - \frac{1}{n}J$ has rank $n - 1$ because it is orthogonal projector($P^T = P, P^2 = P$) and $rank(I - \frac{1}{n}J) = Tr(I - \frac{1}{n}J) = n - 1$. [Theorem 5.5.7]

13.4 Notes on bibliography

For an extensive discussion on statistical distribution, see [7][8].

BIBLIOGRAPHY

1. Hogg, R. V., McKean, J. & Craig, A. T. Introduction to Mathematical Statistics, 7 ed (2012).
2. Borovkova, S., Permana, F. J. & Weide, H. V. A closed form approach to the valuation and hedging of basket and spread options. *Journal of Derivatives* **14,** 8 (2007).
3. Brzezniak, Z. & Zastawniak, T. *Basic stochastic processes: a course through exercises* (Springer Science & Business Media, 1999).
4. Murphy, K. P. *Machine learning: a probabilistic perspective* (MIT press, 2012).
5. Casella, G. & Berger, R. L. *Statistical inference* (Duxbury Pacific Grove, CA, 2002).
6. Ruppert, D. *Statistics and data analysis for financial engineering, 2ed* (Springer, 2015).
7. Forbes, C., Evans, M., Hastings, N. & Peacock, B. *Statistical distributions* (John Wiley & Sons, 2011).
8. Krishnamoorthy, K. *Handbook of statistical distributions with applications* (CRC Press, 2016).

14

STATISTICAL ESTIMATION THEORY

14.1 Estimator theory

14.1.1 Overview

Given observations of random variable X, the **ultimate goal** of statistical inference is to infer the distribution of X. Limited observation data usually make direct inference on the distribution of X impractical. We subjectively **assume** the distribution of X is given by some parameterized statistical model(e.g. Gaussian, Binomial,Poisson). More commonly, a statistical model is written as a set of distributions for X, $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$, where Ω is the parameter space containing all possible values of θ. Note that a statistical model is a hypothesis, which might be correct or incorrect. With the statistical model proposed, we estimate the model parameter θ from the data. Once estimated, we have a way to describe the distribution of X, which finishes the inference task.

There are two major components in statistical inference: **statistical models** and **estimating model parameters**. We usually restrict ourselves to **mathematically convenient models**, such as exponential families, such that useful properties of the distribution, such as mean and We design statistic δ, which is a function of random sample X, such that δ is closed to θ or $g(\theta)$. In this way, we use statistic to relate data X to model parameter θ.

- Not all statistics are equally good. Some are biased: $E\delta_\theta \neq \theta$. Some are more efficient in terms of using information to reduce uncertainty.
- We can use mean-square-error(MSE), or more general risk functions to evaluate of a statistic.

14.1.2 Statistic

Definition 14.1.1 (statistic). *Let $X_1, X_2, ..X_n$ denote a random sample on a random variable X. Let $T = T(X_1, X_2, ..., X_n)$ be a function of the sample. Then T is called a statistic.*

Remark 14.1.1. T is also a random variable.

Definition 14.1.2 (common statistic). *Given a random sample $X_1, ..., X_n$ from X, we have following definitions:*

- *Sample mean:*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- *Sample variance:*
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$
- *Sample standard deviation:*
$$S = \sqrt{S^2}.$$

Remark 14.1.2 (another equivalent form of sample variance). Note that $\sum_{i=1}^{n}(X_i - \overline{X})^2$ can also be written by
$$\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2$$
.

We have
$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2\overline{X}X_i + \overline{X}^2)\\ &= \sum_{i=1}^{n} X_i^2 - 2\overline{X}n\overline{X} + n\overline{X}\\ &= \sum_{i=1}^{n} X_i^2 - n\overline{X}\overline{X}.\end{aligned}$$

We have
$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2 &= \sum_{i=1}^{n} 2nX_i^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} 2X_iX_j\\ &= \sum_{i=1}^{n} 2nX_i^2 - \sum_{i=1}^{n} 2X_in\overline{X}\\ &= \sum_{i=1}^{n} 2nX_i^2 - 2n^2\overline{X}^2\\ &= 2n(\sum_{i=1}^{n} X_i^2 - n\overline{X}\overline{X})\end{aligned}$$

14.1.3 Estimators properties

14.1.3.1 *Basic concepts*

Definition 14.1.3 (unbiased Estimator). *Let $X_1, X_2, ..X_n$ denote a random sample on a random variable X with pdf $f(x;\theta), \theta \in \Omega$. Let $T = T(X_1, X_2, ..., X_n)$ be a statistic. We say that T is an unbiased estimator of θ if $E(T) = \theta$.*

Definition 14.1.4 (consistent Estimator). *Let X be a random variable with cdf $F(x, \theta)$. Let $X_1, X_2, ..., X_n$ be a sample from the distribution of X, and let T_n denote a statistic. T_n is a consistent estimator of θ if*

$$T_n \xrightarrow{P} \theta$$

Definition 14.1.5 (bias of an estimator). *The bias of an estimator $\hat{\theta}$ is*

$$Bias(\hat{\theta}) = E[\hat{\theta}] - \theta,$$

where θ is the true value. If $Bias(\hat{\theta}) = 0$, then estimator $\hat{\theta}$ is said to be unbiased.

Definition 14.1.6 (variance of an estimator). *The variance of an **unbiased** estimator $\hat{\theta}$ is*

$$Var(\hat{\theta}) = E[\hat{\theta}] - \theta)^2],$$

and the covariance is

$$Var(\hat{\theta}) = E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T],$$

where θ is the true value.

Definition 14.1.7 (mean squared error of an estimator). *The mean squared error(MSE) of an estimator is*

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2].$$

14.1.3.2 *Variance-bias decomposition*

Theorem 14.1.1 (variance bias decomposition). *The MSE of an estimator is related to its variance and bias via*

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var[\hat{\theta}] + (Bias(\hat{\theta}))^2 \tag{10}$$

where $Var[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$. *Particularly, if the estimator is unbiased (i.e.* $Bias(\hat{\theta}) = 0$*), we have*

$$MSE(\hat{\theta}) = Var[\hat{\theta}]$$

Proof. Make $(\hat{\theta} - \theta) = (\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)$ and note that $\hat{\theta}$ is a random variable. Specifically,

$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\ &= E[((\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta))^2] \\ &= E[((\hat{\theta} - E[\hat{\theta}])^2] + 2E[((\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta))] + E[(E[\hat{\theta}] - \theta))^2] \\ &= Var[\hat{\theta}] + 2E[((\hat{\theta} - E[\hat{\theta}])](E[\hat{\theta}] - \theta)) + Bias[\hat{\theta}]^2 \\ &= Var[\hat{\theta}] + 0 + Bias[\hat{\theta}]^2 \end{aligned}$$

□

Remark 14.1.3 (biasedness can be useful). At first glance, it may seem that biasedness is always undesired. However, biased estimator might have smaller variance (Figure Figure 14.1.1). As a consequence, biased estimator can have smaller MSE than unbiased estimator. Also consider the following example.

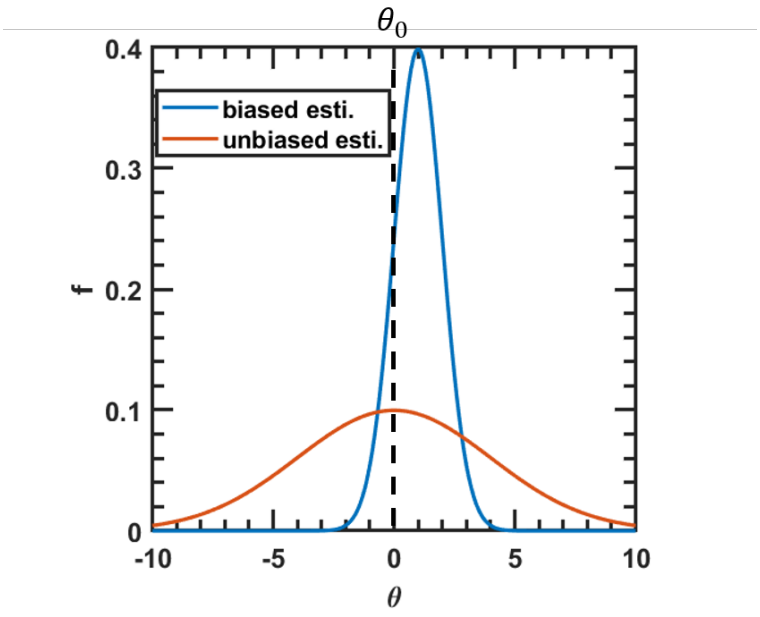


Figure 14.1.1: An example of biased estimator with smaller variance than unbiased estimator

Example 14.1.1. Consider a sample $X_1, X_2, ..., X_n$ of iid normal random variable with unknown mean and variance. Consider two variance estimator

$$S_1^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n-1}, S_2^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n}.$$

Then

- The MSE for S_1^2 is

$$\begin{aligned} MSE[S_1^2] &= Var[S_1^2] + [Bias]^2 \\ &= \frac{2\sigma^4}{n-1} + 0 = \frac{2\sigma^4}{n-1} \end{aligned}$$

  where we use the fact that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

  from Theorem 13.1.5 and

$$Var[\frac{(n-1)S^2}{\sigma^2}] = \frac{(n-1)^2}{\sigma^4} Var[S^2] = 2(n-1),$$

  where use $Var[\chi^2(n)] = 2n$ in Lemma 13.1.33.
- The MSE for S_2^2 is

$$\begin{aligned} MSE[S_2^2] &= (Var[S_2^2] + [Bias])^2 \\ &= \frac{2(n-1)\sigma^4}{n} + (E[S_2^2] - \sigma^2)^2 \\ &= \frac{2(n-1)\sigma^4}{n} + (\frac{(n-1)\sigma^2}{n} - \sigma^2)^2 \\ &= \frac{2n-1}{n^2}\sigma^4 \end{aligned}$$

- $MSE[S_2^2] < MSE[S_1^2]$. That is, the maximum-likelihood estimator has smaller MSE than the unbiased estimator.

14.1.3.3 *Consistence*

Definition 14.1.8 (consistent estimator). *We say $\hat{\theta}$ is a consistent estimator of θ if $\hat{\theta}$ converges to θ in probability, i.e.,*

$$\lim_{n\to\infty} P(\left|\hat{\theta}(X_1, X_2, ..., X_n) - \theta\right| < \epsilon) = 1, \forall \epsilon > 0.$$

Theorem 14.1.2 (MSE criterion for consistent estimator). *An **unbiased** estimator $\hat{\theta}$ is consistent if*

$$\lim_{n\to\infty} Var(\hat{\theta}(X_1, X_2, ..., X_n)) = 0$$

More generally, an estimator $\hat{\theta}$ is consistent if

$$\lim_{n\to\infty} MSE(\hat{\theta}(X_1, X_2, ..., X_n)) = 0$$

Proof. Overall, we can use Theorem 12.11.3 (convergence in mean square implies convergence in probability)

□

Remark 14.1.4 (consistence vs. unbiasedness).

- A consistent estimator is at least **asymptotically unbiased**.However, some unbiased estimators can be inconsistent(i.e. the variance does not converge to 0).
- If the sample size is larger, consistent estimators are consider better than unbiased estimator because consistent estimator ensure that estimator variance is sufficiently smaller.
- Inconsistent estimator usually should be avoided, since increasing the number of samples will not necessarily reduce the variance.

Theorem 14.1.3 (sample mean estimator is consistent). *[1, p. 1160] Let $X_1, ..., X_n$ be a random sample from any population with finite mean μ and finite variance σ^2. Let $\overline{X}_n$ be the sample mean. It follows that*

- *$\overline{X}_n$ is the consistent estimator of μ.*
- *For any function $g(x)$, if $E[g(x)]$ and $Var[g(x)]$ are finite, then the quantity*

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

is the consistent estimator of $E[g(X)]$.

Proof. Because $E[\overline{X}_n] = \mu$ and $Var[\overline{X}_n] \to 0$, then $\overline{X}_n$ converges to μ in mean square and thus in probability [Theorem 12.11.2 and Theorem 12.11.3]. □

14.1.3.4 *Efficiency*

Definition 14.1.9 (relative efficiency). *The relative efficiency of two **unbiased** estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ is the ratio of their variance*

$$\frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_2)}$$

.

Remark 14.1.5. The more efficient estimator is that: **given fixed number of samples**, it has a lower MSE/variance.

Example 14.1.2 (sample mean is the most efficient linear estimator of population mean). Consider the linear estimator

$$\widehat{\theta}_n = \sum_{i=1}^{n} a_i X_i$$

where $E(X_i) = \theta, Var(X_i) = \sigma^2$ for $1 \leq i \leq n$.

$$E\left(\widehat{\theta}_n\right) = \sum_{i=1}^{n} a_i E(X_i) = \theta \sum_{i=1}^{n} a_i,$$

so the estimator is unbiased provided

$$\sum_{i=1}^{n} a_i = 1.$$

For i.i.d. random variables,

$$Var\left(\widehat{\theta}_n\right) = \sum_{i=1}^{n} a_i^2 \sigma^2.$$

Now from constrained optimization theory, we know that minimum is achieved iff $a_i = \frac{1}{n}$, $1 \leq i \leq n$.

The conclusion then is that, if $\widehat{\theta}_n$ is a linear unbiased estimator of the form $\sum_{i=1}^n a_i X_i$ and if $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, then

$$Var\left(\overline{X}\right) \leq Var\left(\widehat{\theta}_n\right).$$

Therefore, among all linear unbiased estimators, $\overline{X}$ *is most efficient* estimator of the population mean.

Definition 14.1.10. *Consider a statistic θ as a function of iid random samples $X_1, X_2, ..., X_n$ with pdf $f(x, \theta)$. The estimator is a* ***uniformly minimum-variance unbiased estimator(UMVUE)*** *if*

- *it is unbiased, i.e., $\hat{\theta} = \theta$*
- *$\forall \theta \in \Theta$, we have*

$$var(\hat{\theta}) \leq var(\hat{\theta}')$$

for any other unbiased estimator $\hat{\theta}'$.

Remark 14.1.6 (interpretation)**.** In terms of efficiency in using data to reduce uncertainty, UMVUE has the optimal estimation efficiency.

Remark 14.1.7 (How to find UMVUE?)**.** There is no simple, general procedure for finding the MVUE estimator. Here are some several approaches:

- Find a sufficient statistic and apply the Rao-Blackwell theorem.
- Determine the so-called Cramer-Rao Lower Bound (CRLB) and verify that the estimator achieves it.
- Further restrict the estimator to a class of estimators (e.g., linear or polynomial functions of the data)
- The existence of UMVUE is in discussed [2, p. 62].

14.1.4 Robust statistics

Definition 14.1.11 (breakdown point). *The finite sample* ***breakdown point*** *of an estimator is the smallest fraction α of data points such that if $[n\alpha]$ points approach ∞, then the estimator approach ∞.*

Example 14.1.3. The sample mean $x_1, x_2, ..., x_n$ is

$$\begin{aligned}\overline{x}_n &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{n}(\sum_{i=1}^{n} x_i + x_n) \\ &= \frac{n-1}{n}(\overline{x}_{n-1}) + \frac{1}{n}x_n\end{aligned}$$

Example 14.1.4.

- (mean) Given sample size n, the breakdown point for the mean using the arithmetic mean is $1/n$; that is one point can ruin the mean.
- (median) The sample median, as an estimate of a population median, can tolerate up to 50% bad values.

Remark 14.1.8. Under the assumption of H_0, we can calculate the p value of the observation, and then decide whether we can reject H_0.

Definition 14.1.12 (α trimmed mean). *Let $k = n\alpha$ rounded to an integer(k is the number of observation removed from both ends for calculation). The* ***α-trimmed mean*** *is defined as*

$$\overline{X}_\alpha = \sum_{i=k+1}^{n-k} \frac{X_i}{n-2k}.$$

Definition 14.1.13 (median absolute deviation). *[3, p. 122] A robust estimator of standard deviation of iid random sample $X_1, X_2, ..., X_n$ is the* ***MAD(median absolute deviation)***

$$\hat{\sigma}^{MAD} = 1.4826 \times median\{|X_i - median(X_i)|\}.$$

Remark 14.1.9 (interpretation).

- For normally distributed data, $median\{|X_i - median(Y_i)|\}$ is the estimator of $\Phi^{-1}(0.75)\sigma = \sigma/1.4826$.
- For a iid normal random sample, as sample size $n \to \infty$, the MAD is the unbiased estimates of σ.

14.2 Method of moments

Definition 14.2.1 (moments of the random sample). *[4, p. 312] Let $X_1, X_2, ..., X_n$ be a sample from a population with pdf or pmf $f(x|\theta_1, ..., \theta_k)$. The first k moments are given by*

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\ &\cdots \\ m_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k \end{aligned}$$

Definition 14.2.2 (moments of the random sample). *[4, p. 312] Let $X_1, X_2, ..., X_n$ be a random sample of X from a population with pdf or pmf $f(x|\theta_1, ..., \theta_k)$. Define $\mu_i = E[X^i], i = 1, 2, ..., k$. The method of moments is aimed at solving $\theta_1, \theta_2, ..., \theta_k$ from the k equations*

$$\begin{aligned} m_1 &= \mu_1(\theta_1, ..., \theta_k) \\ m_2 &= \mu_2(\theta_1, ..., \theta_k) \\ &\cdots \\ m_k &= \mu_k(\theta_1, ..., \theta_k) \end{aligned}$$

where $m_i, i = 1, 2, ..., k$ are the k moments of the sample.

Lemma 14.2.1 (estimating normal distribution parameter via method of moments). *[4, p. 312] Suppose $X_1, X_2, ..., X_n$ are iid random variable with $N(\mu, \sigma^2)$. It follows that*

- $m_1 = \mu, \mu^2 + \sigma^2 = m_2$.
- *The moment of method estimators for (μ, σ) are*

$$\hat{\mu} = m_1, \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Proof. (1) note that $E[X^2] = Var[X] + E[X]^2$. (2) note that

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

□

Lemma 14.2.2 (estimating t distribution parameter via method of moments). *Suppose $X_1, X_2, ..., X_n$ are iid random variable with $t_v(\mu, \sigma^2), v > 2$. It follows that*

- $m_1 = \mu, \mu^2 + \sigma^2 \frac{v}{v-2} = m_2$.
- *The moment of method estimators for (μ, σ) are*

$$\hat{\mu} = m_1, \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \frac{v-2}{v}.$$

Proof. (1)(2)From subsection 13.1.18 note that $E[X] = \mu$, $E[X^2] = Var[X] + E[X]^2 = \sigma^2 \frac{v}{v-2} + \mu^2$. note that

$$\hat{\sigma}^2 \frac{v}{v-2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

□

14.3 Maximum likelihood estimation

14.3.1 Basic concepts

Definition 14.3.1 (estimator). *[4, p. 315] A point estimator is any function $W(\boldsymbol{X}) = W(X_1, X_2, ..., X_n)$ of a random sample; that is, any statistic is a point estimator.*

Definition 14.3.2 (likelihood function). *[5, p. 22] Assuming a statistical model parametrized by a fixed and unknown θ, the likelihood $L(\boldsymbol{x}|\theta)$ is the probability of the observations $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ of iid random samples $X_1, X_2, ..., X_n$ as a function of θ. It can be written as*

$$L(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(X = x_i|\theta)$$

And the corresponding log-likelihood function is defined as

$$\log L(\boldsymbol{x}|\theta) = \sum_{i=1}^{n} f(X = x_i|\theta)$$

Definition 14.3.3 (maximum likelihood estimator and score function). *[4, p. 316] A maximum likelihood estimator(MLE) of the parameter θ based on given observations $\boldsymbol{x}$ is*

$$\hat{\theta} = \max_{\theta} \log L(\boldsymbol{x}|\theta),$$

Or alternatively, $\hat{\theta}$ satisfies

$$s(\theta, \boldsymbol{x}) = \frac{\partial \log L}{\partial \theta} = 0,$$

*where $s(\theta, \boldsymbol{x})$ is called **score function**.*

14.3.2 Examples

Example 14.3.1 (Bernoulli trial MLE). Consider a series of independent Bernoulli trials with success probability θ such that we have probability mass function given by

$$Pr(Y_i = y) = (1 - \theta)^{1-y}\theta^y, y \in \{0, 1\}.$$

- The log-likelihood function based on n observations $Y = \{Y_1, ..., Y_N\}$ can be written by

$$\log L(\theta; Y) = \sum_{i=1}^{n}((1-y_i)\log(1-\theta) + y_i \log\theta) = n((1-\overline{y})\log(1-\theta) + \overline{y}\log(\theta)),$$

  where $\overline{y}$ is the sample mean.
- The MLE is given by

$$\hat{\theta} = \overline{y}.$$

Example 14.3.2 (Normal distribution MLE). The log-likelihood function for n iid observations $x_1, ..., x_n$ drawn from normal distribution is given by

$$\begin{aligned}\log L(\theta_1, \theta_2) &= \prod_{i=1}^{n} f(x_i \theta_1, \theta_2) \\ &= \theta_2^{-n/2}(2\pi)^{-n/2}\exp(-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2) \\ &= -\frac{n}{2}\log\theta_2 - \frac{n}{2}\log(2\pi) - \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2}\end{aligned}$$

where $\theta_1 = \mu, \theta_2 = \sigma^2$. Setting derivatives to zeros, we have

$$\begin{aligned}\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_1} &= \frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2} = 0 \\ \frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_2} &= -\frac{n}{2\theta_2} + \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{\theta_2^2} = 0\end{aligned}$$

which produces

$$\hat{\mu} = \hat{\theta}_1 = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}, \hat{\sigma}^2 = \hat{\theta}_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x}}{n}.$$

Example 14.3.3 (exponential distribution MLE). Consider an exponential distribution with parameter α such that its pdf is given by

$$f(x; \alpha) = \alpha e^{-\alpha x}, x \geq 0.$$

The MLE for α from an iid random sample $X_1, ..., X_n$ is given by $\hat{\alpha} = 1/\overline{X}$ since

$$\log L(\alpha) = n \log \alpha - \alpha \sum_{i=1}^{n} X_i$$

$$\partial \log L(\alpha)/\partial \alpha = \frac{n}{\alpha} - \sum_{i=1}^{n} X_i$$

$$\partial \log L(\alpha)/\partial \alpha = 0 \implies \hat{\alpha} = 1/\overline{X}.$$

14.4 Information and efficiency

14.4.1 Fish information

Assumption 14.1 (Fisher information regularity assumption). *For a pdf $f(x;\theta)$ of random variable X with parameter θ. We make the following regularity assumptions:*

- *The set $A = \{x|p(x;\theta) > 0\}$ does not depend on θ. For all $x \in A, \theta \in \Theta$, $\frac{\partial}{\partial\theta}\log p(x;\theta)$ exists and is finite. Here Θ is the parameter space.*
- *If T is any statistic of X such that $E[|T|] < \infty$ for all $\theta \in \Theta$, then integration and differentiation by θ can be interchanged in the following way:*

$$\frac{\partial}{\partial\theta}[\int T(x)f(x;\theta)dx] = \int T(x)\frac{\partial}{\partial\theta}f(x;\theta)dx,$$

whenever the right-hand side is finite.

Definition 14.4.1 (Fisher information). *For one dimensional parametric family of pdf or pmf $f(x;\theta)$, we define the Fisher information for $\theta \in \mathbb{R}$ as*

$$I(\theta) = E[(\frac{d}{d\theta}\log f(x;\theta))^2],$$

*where the expectation is **taken with respect to** x. In particular, if $\theta \in \mathbb{R}^N$, we have Fisher information matrix defined as*

$$I(\theta)_{ij} = -E[\frac{\partial^2 \log(f(x;\theta))}{\partial\theta_i\theta_j}].$$

Remark 14.4.1. This definition holds both for discrete or continuous random variables, as long as f is differentiable respect to θ.

Theorem 14.4.1 (Basic properties of Fisher information). *Let $f(x;\theta)$ be a pdf parameterized by $\theta \in \mathbb{R}$ with Assumption 14.1 holds, then*

- $$E[\frac{d}{d\theta}\log f(x;\theta)] = 0$$
- $$I(\theta) = Var[\frac{d}{d\theta}\log f(x;\theta)].$$

- *Further assume $f(x;\theta)$ is twice differentiable and interchange between integration and differentiation is permitted. Then*

$$I(\theta) = E[(\frac{d}{d\theta}\log f(x;\theta))^2] = -E[\frac{\partial^2 \log(f(x;\theta))}{\partial\theta^2}].$$

- *For $\theta \in \mathbb{R}^N$,*

$$I(\theta) = E[\frac{\partial \log(f(x;\theta))}{\partial\theta}(\frac{\partial \log(f(x;\theta))}{\partial\theta})^T] = -E[\frac{\partial^2 \log(f(x;\theta))}{\partial\theta\theta^T}].$$

Proof. (1) The equivalence of these two expressions can be showed as:

$$\begin{aligned} E[\frac{d}{d\theta}\log f(x;\theta)] &= \int \frac{1}{f(x;\theta)}\frac{d}{d\theta}f(x;\theta)f(x;\theta)dx \\ &= \int \frac{d}{d\theta}f(x;\theta)dx \\ &= \frac{d}{d\theta}\int f(x;\theta)dx \\ &= \frac{d}{d\theta}1 \\ &= 0 \end{aligned}$$

(2) Based on definition, we have

$$\begin{aligned} Var[\frac{d}{d\theta}\log f(x;\theta)] &= E[(\frac{d}{d\theta}\log f(x;\theta))^2] - E[\frac{d}{d\theta}\log f(x;\theta)]^2 \\ &= E[(\frac{d}{d\theta}\log f(x;\theta))^2] - 0 \\ &= I(\theta) \end{aligned}$$

(3)

$$\begin{aligned} \frac{\partial^2}{\partial\theta^2}\log f(x;\theta) &= \frac{\partial}{\partial\theta}\frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta}f(x;\theta) \\ &= -\frac{\partial}{\partial\theta}\frac{1}{f(x;\theta)^2}\frac{\partial}{\partial\theta}f(x;\theta) + \frac{1}{f(x;\theta)}\frac{\partial^2}{\partial\theta^2}f(x;\theta) \\ &= -(\frac{\partial}{\partial\theta}\log f(x;\theta))^2 + \frac{1}{f(x;\theta)}\frac{\partial^2}{\partial\theta^2}f(x;\theta) \end{aligned}$$

Take expectation with respect to x on both sides and note that

$$E[\frac{1}{f(x;\theta)}\frac{\partial^2}{\partial\theta^2}f(x;\theta)] = \int \frac{\partial^2}{\partial\theta^2}f(x;\theta)dx = \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)dx = \frac{\partial^2}{\partial^2\theta}1 = 0$$

□

14.4.2 Information matrix for common distributions

14.4.2.1 *Bernoulli distribution*

Lemma 14.4.1 (Fisher information for Bernoulli distribution). *Let the pmf of Bernoulli distribution parameterized by $f(x;\theta) = \theta^x(1-\theta)^{1-x}, x \in \{0,1\}$. Then*

$$I(\theta) = \frac{1}{\theta(1-\theta)}.$$

Proof.

$$\begin{aligned} I(\theta) &= -E[\frac{\partial^2 \log(f(x;\theta))}{\partial\theta^2}] \\ &= E[\frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}] \\ &= \theta(1/\theta^2) + (1-\theta)(1/(1-\theta)^2) \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

□

14.4.2.2 *Normal distribution*

Lemma 14.4.2 (Fish information matrix for univariate normal distribution). *[1, p. 548] Let the pdf of normal distribution parameterized by*

$$f(x;\theta) = (2\pi\theta_2)^{-1/2}\exp(-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x-\theta_1)^2).$$

$$\begin{aligned} \frac{\partial^2 \log f}{\partial\theta_1^2} &= -\frac{1}{\theta_2} = -\frac{1}{\sigma^2} \\ \frac{\partial^2 \log f}{\partial\theta_2^2} &= \frac{1}{2\theta_2^2} - \frac{1}{\theta_2^3}(x-\theta_1)^2 \\ \frac{\partial^2 \ln f}{\partial\theta_1\partial\theta_2} &= -\frac{1}{\theta_2^2}(x_i-\theta_1) \end{aligned}$$

Finally, take expectation with respect to x and we have

$$I(\theta_1, \theta_2) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix}$$

14.4.3 Cramer-Rao lower bound

14.4.3.1 *Information inequality*

Theorem 14.4.2 (information inequality for statistic). *Let $T(X)$ be any statistic such that $Var[T(X)] < \infty$ for all θ. Denote $E[T(X)]$ by $\phi(\theta)$. Suppose Assumption 14.1 holds and $0 < I(\theta) < \infty$. Then **for all** θ*

$$Var[T(X)] \geq \frac{[\phi'(\theta)]^2}{I(\theta)}.$$

Proof. Based on the Assumption 14.1, ,we have

$$\phi'(\theta) = \int T(x) \frac{\partial}{\partial \theta} f(x;\theta) dx = \int T(x) \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta) dx.$$

Therefore, we can view

$$\phi'(\theta) = E[T(X) \frac{\partial \log f(x;\theta)}{\partial \theta}] = Cov[T(X), \frac{\partial \log f(x;\theta)}{\partial \theta}],$$

since $E[\frac{\partial \log f(x;\theta)}{\partial \theta}] = 0 \implies E[T(X)]E[\frac{\partial \log f(x;\theta)}{\partial \theta}] = 0.$

Using Cauchy-Schwartz inequality [Theorem 12.10.4], we have

$$\left|\phi'(\theta)\right|^2 = Cov[T(X), \frac{\partial \log f(x;\theta)}{\partial \theta}]^2 \leq Var[T(X)] Var[\frac{\partial \log f(x;\theta)}{\partial \theta}].$$

At last, use the fact [Theorem 14.4.1] that $I(\theta) = Var[\frac{\partial \log f(x;\theta)}{\partial \theta}]$, we can get the final result. □

Corollary 14.4.2.1 (information lower bound for general estimators). *Let $T(X)$ be a (generally biased) estimator of θ such that*

$$\phi(\theta) \triangleq E[T(X)] = \theta + \underbrace{b(\theta)}_{bias}.$$

Then

- *the variance of $T(X)$ is*
$$Var[T(X)] \geq \frac{\left|1 + b'(\theta)\right|}{I(\theta)}.$$
- *the MSE of $T(X)$ is*
$$MSE[T(X)] \geq \frac{\left|1 + b'(\theta)\right|}{I(\theta)} + b(\theta)^2.$$

14.4.3.2 *Cramer-Rao lower bound: univariate case*

Theorem 14.4.3 (Cramer-Rao lower bound in univariate estimation). *Let $\hat{\theta}$ be an arbitrary univariate estimator as a function of iid random samples $X_1, ..., X_n$, whose distribution is parameterized by single parameter θ. Let θ_0 be the true value. Then the variance of the estimator $\hat{\theta}$ is bounded by*
$$Var(\hat{\theta}) \geq \frac{(\frac{d}{d\theta} E[\hat{\theta}])^2}{nI_1(\theta^0)},$$
where $I_1(\theta)$ is the Fish information associated with distribution $f(x;\theta)$ and the expectation is taken with respect to x. Particularly, if the estimator $\hat{\theta}$ is unbiased (that is $E[\hat{\theta}] = \theta$), we have
$$Var(\hat{\theta}) \geq \frac{1}{nI_1(\theta^0)}.$$

Proof. Note that the Fisher information $I(\theta)$ associated with the joint distribution of $(X_1, ..., X_n)$ can be expressed by $I(\theta) = nI_1(\theta)$, where $I_1(\theta)$ is the Fisher information associated with $f(x;\theta)$. This is because under iid assumption,
$$E[\log f(x_1, ..., x_n; \theta)] = nE[\log f(x;\theta)].$$
Then use the information inequality [Theorem 14.4.2], we have
$$Var(\hat{\theta}) \geq \frac{(\frac{d}{d\theta} E[\hat{\theta}])^2}{nI_1(\theta)}.$$
□

Example 14.4.1 (univariate estimation for normal distributions).

- Consider an unbiased mean estimator $\hat{\mu}$ and an unbiased variance estimator $\hat{\sigma}^2$ for normal distribution with unknown mean μ and variance σ^2. Because the information matrix is given by [Lemma 14.4.2]

The mean estimator has bounded variance given by (using information matrix from

$$Var[\hat{\mu}] \geq \frac{1}{nI_1(\theta)} = \sigma^2/n.$$

- Consider anfor normal distribution with known mean μ. The mean estimator has bounded variance given by

$$Var[\hat{\sigma}^2] \geq \frac{1}{nI_1(\theta)} = 2\sigma^4/n.$$

- It is clear that
  - Increasing sample size n will reduce the estimator variance.
  - Mean/variance estimators of random samples drawn from small-variance distributions have inherent smaller variances.

14.4.3.3 *Cramer-Rao lower bound: multivariate case*

Theorem 14.4.4 (information inequality for statistic: multivariate case). *Let $T(X)$ be any statistic such that $Var[T(X)] < \infty$ for all θ. Denote $E[T(X)]$ by $\phi(\theta)$. Suppose Assumption 14.1 holds and $0 < I(\theta) < \infty$. Then **for all** θ*

$$Var[T(X)] \geq [\nabla_\theta \phi]^T [I(\theta)]^{-1} [\nabla_\theta \phi].$$

Proof. Based on the Assumption 14.1, ,we have

$$\phi'(\theta) = \int T(x) \frac{\partial}{\partial \theta} f(x;\theta) dx = \int T(x) \frac{\partial \log f(x;\theta)}{\partial \theta} dx.$$

Therefore, we can view

$$\phi'(\theta) = E[T(X) \frac{\partial \log f(x;\theta)}{\partial \theta}] = Cov[T(X), \frac{\partial \log f(x;\theta)}{\partial \theta}],$$

since $E[\frac{\partial \log f(x;\theta)}{\partial \theta}] = 0 \implies E[T(X)]E[\frac{\partial \log f(x;\theta)}{\partial \theta}] = 0.$

Using Cauchy-Schwartz inequality [Theorem 12.10.4], we have

$$\left|\phi'(\theta)\right|^2 = Cov[T(X), \frac{\partial \log f(x;\theta)}{\partial \theta}]^2 \leq Var[T(X)] Var[\frac{\partial \log f(x;\theta)}{\partial \theta}].$$

At last, use the fact [Theorem 14.4.1] that $I(\theta) = Var[\frac{\partial \log f(x;\theta)}{\partial \theta}]$, we can get the final result. □

Proof. Similar to Theorem 14.4.4, we can show that

$$\frac{\partial \phi(\theta)}{\partial \theta_j} = Cov(T(X), \frac{\partial \log f(x;\theta)}{\partial \theta_j}).$$

For constants $c_1, c_2, ..., c_p$, note that

$$Var[T(X) - \sum_{j=1}^{p} c_j \frac{\partial \log f(x;\theta)}{\partial \theta_j}] = Var[T(X)] + c^T I(\theta) c - 2c^T [\nabla_\theta \phi]$$
$$\geq 0$$

Particularly, the minimum is achieved at $c* = [I(\theta)]^{-1}\delta$. Then

$$Var[T(X) - \sum_{j=1}^{p} c_j \frac{\partial \log f(x;\theta)}{\partial \theta_j}] = Var[T(X)] - [\nabla_\theta \phi]^T [I(\theta)]^{-1} [\nabla_\theta \phi] \geq 0.$$

□

Theorem 14.4.5 (Cramer-Rao lower bound in multivariate estimation). *Let $\hat{\theta}$ be a p-dimension **unbiased** estimator as a function of iid random samples $X_1, ..., X_n$, whose distribution is parameterized by parameter vector $\theta \in \mathbb{R}^p, \theta = (\theta_1, ..., \theta_p)$. Let θ^0 be the true value.*

Then the variance matrix of the estimator $\hat{\theta}_i$ is bounded by

$$Var(\hat{\theta}) \geq [n[I_1(\theta)]^{-1},$$

where $I_1(\theta^0)$ is the Fish information maxtrix associated with distribution $f(x;\theta)$ and the expectation is taken with respect to x.

Proof. Note that the Fisher information $I(\theta)$ associated with the joint distribution of $(X_1, ..., X_n)$ can be expressed by $I(\theta) = nI_1(\theta)$, where $I_1(\theta)$ is the Fisher information associated with $f(x, \theta)$. This is because under iid assumption,

$$E[\log f(x_1, ..., x_n; \theta)] = nE[\log f(x;\theta)].$$

Then use the information inequality [Theorem 14.4.4], we have

$$Var(\alpha^T \hat{\theta}) = \alpha^T Var[\hat{\theta}]\alpha \geq [\nabla_\theta \alpha^T \hat{\theta}]^T [nI_1(\theta)]^{-1} [\nabla_\theta \alpha^T \hat{\theta}] = \alpha^T [nI_1(\theta)]^{-1} \alpha,$$

where $\alpha \in \mathbb{R}^p$ is an arbitrary vector. □

Example 14.4.2 (multivariate estimation for normal distributions).

- Consider an unbiased mean estimator $\hat{\mu}$ for normal distribution with known variance σ^2. The information matrix is given by (Lemma 14.4.2)

$$I(\theta_1, \theta_2) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix}.$$

  Therefore,

$$Var[\hat{\mu}] \geq \sigma^2/n, Var[\hat{\sigma}^2] \geq 2\sigma^4/n,$$

- It is clear that
  - Increasing sample size n will reduce the estimator variance.
  - Mean/variance estimators of random samples drawn from small-variance distributions have inherent smaller variances.

14.4.3.4 *Efficient estimator*

Definition 14.4.2 (efficient estimator). *An estimator $\hat{\theta}$ if variance achieves equality in the Cramer Rao lower bound for all $\theta \in \Theta$.*

Remark 14.4.2. An efficient estimator is optimal in the sense of using information to reduce uncertainty.

Example 14.4.3. Let the pmf of Bernoulli distribution parameterized by $f(x;\theta) = \theta^x(1-\theta)^{1-x}, x \in \{0,1\}$. Then

$$\begin{aligned} I(\theta) &= -E[\frac{\partial^2 \log(f(x;\theta))}{\partial \theta^2}] \\ &= E[\frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}] \\ &= \theta(1/\theta^2) + (1-\theta)(1/(1-\theta)^2) \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

Consider the estimator $\hat{\theta} = \overline{X}$, then

$$E[\hat{\theta}] = E[\overline{X}] = E[\sum_{i=1}^{n} X_i/n] = \theta$$

and

$$Var[\hat{\theta}] = E[\overline{X}^2] - E[\hat{\theta}]^2 = \theta/n + \theta^2(n-1)/n - \theta^2 = \theta(1-\theta)/n$$

Therefore, the variance of the estimator is achieving the lower bound.

Lemma 14.4.3. *An efficient estimator is UMVUE; that is, an unbiased estimator with variance achieving the Cramer-Rao lower bound is UMVUE. The converse might not be true.*

Proof. If the variance achieve the lower bound, then it has unformly minimum variance, and therefore it is UMVUE. On the other hand, it is likely that for some estimator the lower bounded is not achieved, but it has smaller variance than any other competing estimators. □

Remark 14.4.3 (implications and practical issues).

- The Cramer-Rao lower Bound enables us to judge whether one estimator is efficient: the closer to lower bound, the more efficient.
- Efficient estimators are usually difficult to find in practice.

Theorem 14.4.6 (Rao-Blackwell theorem). *Let $\hat{\theta}(X_1, ..., X_n)$ be an estimator, and T be any sufficient statistic. Then the new estimator $\hat{\theta}^* = E[\hat{\theta}(X_1, ..., X_n)|T(X_1, ..., X_n)]$ will satisfy*

$$MSE(\hat{\theta}^*) \leq MSE(\hat{\theta}), \forall\theta.$$

Particularly, if both estimators are unbiased, we have

$$Var(\hat{\theta}^*) \leq Var(\hat{\theta}), \forall\theta.$$

*The process of creating a new improved estimator via conditioning is called **Rao-Blackwellization**.*

Proof. Since $Var(X) \geq 0$ implies $E[X^2] \geq E[X]^2$, we have

$$E[(\hat{\theta} - \theta)^2|T] \geq (E[\theta \hat{-} \theta|T])^2 = (E[\hat{\theta}|T] - \theta)^2 = (\hat{\theta}^* - \theta)^2$$

Take expectation on both sides and use the Tower property of conditional expectation, we have

$$E[(\hat{\theta}^* - \theta)^2] \leq E[(\hat{\theta} - \theta)^2]$$

□

Lemma 14.4.4 (unbiasedness inheritance). *The improved estimator is unbiased if and only if the original estimator is unbiased.*

Proof.

$$E[E[\hat{\theta}(X_1, ..., X_n)|T(X_1, ..., X_n)]] = E[\hat{\theta}(X_1, ..., X_n)]$$

□

Lemma 14.4.5 (idempotent operation). *Rao-Blackwellization is an idempotent operation. More precisely, let Rao-Blackwellization operator denoted as R, let $\hat{\theta}$ be the original estimator, then*

$$R^2[\hat{\theta}] = R[\hat{\theta}]$$

Proof.

$$E[E[\hat{\theta}|T]|T] = E[\hat{\theta}|T]$$

via the law of iterated expectation. □

Remark 14.4.4 (Rao-Blackwell theorem vs. Cramer-Rao lower bound)**.**

- An estimator achieving the Cramer-Rao lower bound is performing better than any other estimators in all the possible θ.
- Rao-Blackwellization provides a way to improve current estimator such that the improved estimator is no worst than the current estimator in all the possible θ. The improved estimator cannot guarantee to achieve the Cramer-Rao lower bound, i.e., being better than any other estimators.

14.4.4 Fisher information characterization of MLE

14.4.4.1 *Properties of score function*

Lemma 14.4.6.

$$E[s(\theta, \boldsymbol{X})] = 0$$

Proof. This is a restatement of Theorem 14.4.1 □

Lemma 14.4.7 (useful properties of score function). *[6, p. 550] Let t be any vector-valued function $\boldsymbol{X}$ and θ, then*

$$E[s(\theta, \boldsymbol{X})t^T(\theta, \boldsymbol{X})] = \frac{\partial}{\partial\theta}E[t^T] - E[\frac{\partial}{\partial\theta}t^T]$$

And particularly, if $t = s$, we have

$$E[s(\theta, \boldsymbol{X})s^T(\theta, \boldsymbol{X})] = -E[\frac{\partial}{\partial\theta}t^T]$$

Proof. (1) Use the fact that

$$E[t^T] = \int t^T f(x, \theta)dx$$

and differentiate with respect to θ on both sides. (2) use the fact $E[s(\theta, \boldsymbol{X})] = 0$ in above lemma. □

14.4.4.2 *Fisher information and MLE*

Definition 14.4.3 (Fisher information). *The covariance matrix of the score function is called **Fisher information matrix**, denoted by $I(\theta)$, and is given by*

$$I(\theta) = E[s(\theta, \boldsymbol{X})s^T(\theta, \boldsymbol{X})],$$

where

$$s(\theta, \boldsymbol{X}) = \nabla_\theta \log L(\boldsymbol{x}|\theta), \log L(\boldsymbol{x}|\theta) = \sum_{i=1}^{n} f(X = x_i|\theta).$$

14.4.4.3 *MLE efficiency*

In general, with finite samples, an MLE is not necessary efficient. Consider a biased MLE, then

$$s(\theta, X) = 0 \neq I(\theta)(\hat{\theta} - \theta).$$

Therefore, the variance of $\hat{\theta}$ does not achieve the Cramer-Rao lower bound [Theorem 14.4.7]. However, as the sample size becomes sufficiently large, MLE is consistent and asymptotic efficient [Theorem 14.4.8] no matter it is unbiased or not.

Definition 14.4.4 (efficient estimator). *An estimator $\hat{\theta}$ is said to be **efficient** if it is **unbiased** and the covariance of $\hat{\theta}$ achieves the Cramer-Rao lower bound; i.e., it satisfies*

$$E[\hat{\theta}] = \theta, E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] = I^{-1}(\theta).$$

Theorem 14.4.7 (sufficient and necessary condition for MLE be efficient). *[6, p. 552] An unbiased estimator $\hat{\theta}$ is efficient (i.e., achieving the Cramer-Rao lower bound) if and only if*

$$I(\theta)(\hat{\theta} - \theta) = s(\theta, \boldsymbol{X})$$

*where J is the Fisher information matrix. Furthermore, **any unbiased maximum-likelihood estimator is an efficient estimator.***

Proof. (forward) Suppose $I(\theta)(\hat{\theta} - \theta) = s(\theta, \boldsymbol{X})$, then

$$I = E[ss^T] = IE[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T]I$$

by the definition of I. Multiply both sides by I^{-1}, and rearrange we get

$$E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] = I^{-1}.$$

(converse) Assume $\hat{\theta}$ is efficient (also unbiased). First we can show

$$E[s(\theta, X)(\hat{\theta} - \theta)^T] = \mathcal{I},$$

where $\mathcal{I}$ is identity matrix and the expectation is taken with respect to X. This is because $E[s(\theta, X)\theta^T] = E[s(\theta, X)]\theta = 0$, where we use Theorem 14.4.1. In addition,

$$\begin{aligned} E[s(\theta, X)\hat{\theta}^T] &= \int \frac{\partial \log f(x|\theta)}{\partial \theta} \hat{\theta}^T f(x|\theta) dx \\ &= \int \frac{\partial f(x|\theta)}{\partial \theta} \hat{\theta}^T dx \\ &= \frac{\partial}{\partial \theta} E[\hat{\theta}^T] \\ &= \frac{\partial}{\partial \theta} \theta^T \\ &= \mathcal{I}. \end{aligned}$$

Now use Cauchy-Schwartz inequality [Theorem 12.10.4]

$$\begin{aligned} \mathcal{I} &= (E[s(\theta, X)(\hat{\theta} - \theta))^T])^2 \\ &= E[ss^T]E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] \\ &= IE[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] \\ &= II^{-1} \\ &= \mathcal{I} \end{aligned}$$

Because the equality in Cauchy-Schwartz inequality hold if and only if

$$s(\theta, X) = K(\theta)(\hat{\theta} - \theta),$$

for some $K(\theta)$, which can be $I(\theta)$. □

Example 14.4.4. Let the pmf of Bernoulli distribution parameterized by $f(x;\theta) = \theta^x(1-\theta)^{1-x}, x \in \{0,1\}$. Then

$$\begin{aligned} I(\theta) &= -E[\frac{\partial^2 \log(f(x;\theta))}{\partial\theta^2}] \\ &= E[\frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}] \\ &= \theta(1/\theta^2) + (1-\theta)(1/(1-\theta)^2) \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

14.4.5 MLE for normal distribution

Lemma 14.4.8 (mean and variance estimator of normal samples). *[7] Let $X_1, X_2, ..., X_n$ be a random sample from $N(\mu, \sigma^2)$. Define two statistics*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

- *$\overline{X}$ is the **uniformly minimum variance unbiased** (UMVU) estimator of μ. And*

$$E[\overline{X}] = \mu, Var[\overline{X}] = \frac{\sigma^2}{n}.$$

- *S^2 is the uniformly minimum variance unbiased (UMVU) estimator of σ^2. And*

$$E[S^2] = \sigma^2, Var[S^2] = \sigma^4\frac{2}{n-1}.$$

Proof. (1) To prove UMVU, see reference. Note that from Theorem 13.1.5,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

and

$$Var[\frac{(n-1)S^2}{\sigma^2}] = \frac{(n-1)^2}{\sigma^4}Var[S^2] = 2(n-1),$$

where use $Var[\chi^2(n)] = 2n$ in Lemma 13.1.33.

The information matrix is given by

$$I_1()$$

□

Example 14.4.5. Consider a sample $X_1, X_2, ..., X_n$ of iid normal random variable with unknown mean and variance. Consider two variance estimator

$$S_1^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n-1}, S_2^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n}.$$

Then

- The MSE for S_1^2 is

$$\begin{aligned} MSE[S_1^2] &= Var[S_1^2] + [Bias]^2 \\ &= \frac{2\sigma^4}{n-1} + 0 = \frac{2\sigma^4}{n-1} \end{aligned}$$

  where we use the fact that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

  from Theorem 13.1.5 and

$$Var[\frac{(n-1)S^2}{\sigma^2}] = \frac{(n-1)^2}{\sigma^4} Var[S^2] = 2(n-1),$$

  where use $Var[\chi^2(n)] = 2n$ in Lemma 13.1.33.
- The MSE for S_2^2 is

$$\begin{aligned} MSE[S_2^2] &= (Var[S_2^2] + [Bias])^2 \\ &= \frac{2(n-1)\sigma^4}{n} + (E[S_2^2] - \sigma^2)^2 \\ &= \frac{2(n-1)\sigma^4}{n} + (\frac{(n-1)\sigma^2}{n} - \sigma^2)^2 \\ &= \frac{2n-1}{n^2}\sigma^4 \end{aligned}$$

- $MSE[S_2^2] < MSE[S_1^2]$. That is, the maximum-likelihood estimator has smaller MSE than the unbiased estimator.

14.4.6 Asymptotic properties of MLE

Definition 14.4.5 (asymptotic efficiency). *[1, p. 542] An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed, and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimators.*

Theorem 14.4.8 (asymptotic properties of MLE). *[6, p. 553][1, p. 478] Let $\hat{\theta}$ be the MLE of coefficient associated with distribution $f(x;\theta)$. Let θ_0 be the true value of the parameter. Let $I_1(\theta)$ be the Fisher information matrix associated with distribution $f(x;\theta)$. It follows that*

- *Maximum-likelihood estimators are consistent; that is*

$$\operatorname*{plim}_{n\to\infty} \hat{\theta} = \theta_0.$$

- *Maximum-likelihood estimators are asymptotic normal; that is,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to MN(0, [I_1(\theta_0)]^{-1})$$

- *Maximum-likelihood estimators are asymptotic efficient.*

Proof. (sketch) (1) Define a scaled log-likelihood function

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i|\theta).$$

Consider the function $L(\theta) = \int (\log f(x;\theta) f(x;\theta_0) dx$, we can show that the true parameter θ_0 is the maximizer of $L(\theta)$; that is, for any θ, we have

$$L(\theta) \leq L(\theta_0).$$

$$\begin{aligned} L(\theta) - L(\theta_0) &= E_{\theta_0}[\log f(X;\theta) - \log f(X;\theta_0)] \\ &= E_{\theta_0}[\log \frac{f(X;\theta)}{f(X;\theta_0)}] \\ &\leq E_{\theta_0}[\frac{f(X;\theta)}{f(X;\theta_0)} - 1] \\ &= \int (\frac{f(X;\theta)}{f(X;\theta_0)} - 1) f(x;\theta_0) dx \\ &= \int f(x;\theta) dx - \int f(x;\theta_0) dx \\ &= 1 - 1 = 0 \end{aligned}$$

where we use inequality $\log x \leq x - 1$. From the law of large numbers, $L_n(\theta)$ converge to $L(\theta)$ in probability. Since MLE $\hat{\theta}$ is the maximizer for $L_n(\theta)$, $\hat{\theta}$ converges to θ_0 in probability. (2) Since as $n \to \infty$, $\hat{\theta} \to \theta_0$ (becomes unbiased), we can show that $\hat{\theta} - \theta_0$ has the variance reaching the Cramer-Rao lower bound. Further, use central limit theorem can arrive at the conclusion. (3) It is asymptotic efficient because its variance reaches Cramer-Rao lower bound. □

Remark 14.4.5 (interpretation and implications). For finite-sample MLE, only unbiased MLE is efficient [Theorem 14.4.7]. At the large sample limit, MLE is always consistent (asymptotic unbiased) therefore efficient.

Example 14.4.6. Consider an exponential distribution with parameter α such that its pdf is given by

$$f(x;\alpha) = \alpha e^{-\alpha x}, x \geq 0.$$

- The MLE for α from an iid random sample $X_1, ..., X_n$ is given by $\hat{\alpha} = 1/\overline{X}$ since

$$\log L(\alpha) = n \log \alpha - \alpha \sum_{i=1}^{n} X_i$$

$$\partial \log L(\alpha)/\partial \alpha = \frac{n}{\alpha} - \sum_{i=1}^{n} X_i$$

$$\partial \log L(\alpha)/\partial \alpha = 0 \implies \hat{\alpha} = 1/\overline{X}.$$

- The Fisher information is given by $I(\alpha) = \frac{1}{\alpha^2}$ since

$$\log f(x;\alpha) = \log \alpha - \alpha x$$
$$\partial^2 \log L(\alpha)/\partial \alpha^2 = -\frac{1}{\alpha^2}$$
$$I(\alpha) = -E[\partial^2 \log L(\alpha)/\partial \alpha^2] = \frac{1}{\alpha^2}.$$

- The MLE $\hat{\alpha} = 1/\overline{X}$ is asymptotic normal and

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \to N(0, \alpha_0^2).$$

Example 14.4.7 (Fish information matrix for normal distributions). [1, p. 548]

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\sigma^2}$$
$$\frac{\partial^2 \ln L}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2$$
$$\frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu)$$

For the asymptotic variance of the maximum likelihood estimator, we need the expectations of these derivatives. Using $E[X_i] = \mu$, we have

$$[-E_0[\frac{\partial^2 \ln L(\theta_0)}{\partial \theta_0 \partial \theta_0^T}]]^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{bmatrix}.$$

Let $\hat{\theta}$ be the maximum-likelihood estimator, then the gradient of the log-likelihood function equals zero at $\hat{\theta}$; that is,

$$g(\hat{\theta}) = 0.$$

Expand this equation in a second-order Taylor series around θ_0, we have

$$0 = g(\hat{\theta}) = g(\theta_0) + H(\overline{\theta})(\hat{\theta} - \theta_0),$$

where θ_0 is the true value, and $\overline{\theta} = w\hat{\theta} + (1-w)\theta_0$ for some $0 < w < 1$. Rearranging this function and multiply it by n and we get

$$\sqrt{n}(\hat{\theta} - \theta_0) = [-H(\overline{\theta})]^{-1}(\sqrt{n}g(\theta_0))$$

14.5 Sufficiency and data reduction

14.5.1 Sufficient estimators

When we estimate a model parameter θ, **not all the information in the data are relevant** to the estimation procedure. For example, if we want to estimate the mean, then the order of the sample is irrelevant. A **sufficient statistic** for a model parameter θ represents the **summary of all information from the data that are useful** for estimation of θ.

Definition 14.5.1 (sufficient statistics). *Let X be a random sample of size n. A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ; that is*

$$P(X|T,\theta) = P(X|T)$$

In otherwise, X and θ are conditional independent given T.

Remark 14.5.1 (sufficient statistic as a lossless data compression). A statistic is sufficient means that $T(X)$ itself can capture all the information useful in estimating θ; the sample X might contain more information than $T(X)$(since $T(X)$ is usually not 1-1), but this additional information does not provide additional usefulness in estimating θ.

Example 14.5.1 (**trivial sufficient statistic**). The statistic $T(X_1, ..., X_n) = (X_1, ..., X_n)$ is always sufficient for any estimation task.

Example 14.5.2. Suppose $X_1, X_2 \sim B(n, \theta)$ and consider

$$P\left(X_1 = x | X_1 + X_2 = r\right)$$

$$= \frac{P\left(X_1 = x, X_1 + X_2 = r\right)}{P\left(X_1 + X_2 = r\right)}$$

$$= \frac{P\left(X_1 = x, X_2 = r - x\right)}{P\left(X_1 + X_2 = r\right)}$$

$$= \frac{\binom{n}{x}\theta^x (1-\theta)^{n-x}\binom{n}{r-x}\theta^{r-x}(1-\theta)^{n-r+x}}{\binom{2n}{r}\theta^r (1-\theta)^{2n-r}}$$

$$= \frac{\binom{n}{x}\binom{n}{r-x}}{\binom{2n}{r}}.$$

This does not contain θ, so that $X_1 + X_2$ is a sufficient statistic for θ.

14.5.2 Factorization theorem

Theorem 14.5.1 (Neyman-Fisher Factorization theorem). *[4, p. 276] Let $f(\boldsymbol{x}|\theta)$ denote the joint pdf or pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if there exist functions $g(T(\boldsymbol{x})|\theta))$ such that for all sample points $\boldsymbol{x} \in \mathcal{X}$ and all parameter points $\theta \in \Theta$, we have*

$$f(\boldsymbol{x}|\theta) = g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})$$

Proof. (1) Assume $T(\boldsymbol{X})$ is sufficient, then we have $f(\boldsymbol{x}|T(\boldsymbol{x}), \theta) = f(\boldsymbol{x}|T(\boldsymbol{x}))$. Then we have

$$\begin{aligned} f(\boldsymbol{x}|\theta) = f(\boldsymbol{x}|\theta)f(T(\boldsymbol{x})|\boldsymbol{x}, \theta) &= f(\boldsymbol{x}, T(\boldsymbol{x})|\theta) = f(T(\boldsymbol{x})|\theta)f(\boldsymbol{x}|T(\boldsymbol{x}), \theta) \\ &= f(T(\boldsymbol{x})|\theta)f(\boldsymbol{x}|T(\boldsymbol{x}))(\text{use sufficiency}) \\ &= h(\boldsymbol{x})g(T(\boldsymbol{x})|\theta) \end{aligned}$$

(2) Assume the factorization holds. Let $T(\boldsymbol{x}) = a$.

Because $f(\boldsymbol{x};\theta) = g\left(T(\boldsymbol{x})|\theta\right) h(\mathbf{x})$, we have

$$P\left(T(\boldsymbol{X}) = a\right) = \int_{\boldsymbol{y}\in T^{-1}(a)} p(\boldsymbol{y})d\boldsymbol{y} = g(a|\theta) \int_{\boldsymbol{y}\in T^{-1}(a)} h(\boldsymbol{y})d\boldsymbol{y}.$$

Hence

$$P\left(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = a\right) = \frac{h(\boldsymbol{x})}{\int_{\boldsymbol{y}\in s^{-1}(a)} h(\boldsymbol{y})d\boldsymbol{y}}$$

and this does not depend upon θ.

□

Example 14.5.3. $X_1, X_2, \ldots, X_n \sim U(0,\theta)$, so that

$$f(\boldsymbol{x}|\theta) = \theta^{-n}, \quad 0 < x_1, \ldots, x_n < \theta.$$

Or equivalently that

$$f(\boldsymbol{x}|\theta) = \theta^{-n}, \quad \theta > x_{(n)} \triangleq \max_i X_i.$$

We can factorize this as $T(\mathbf{x}) = x_{(n)}$ and $h\left(\mathbf{x}\right) = 1$, so that $X_{(n)}$ is a sufficient statistic for θ.

Example 14.5.4. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Bernoulli distribution. Then

$$f(\mathbf{x};\theta) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}.$$

We can factorize this as $T(\mathbf{x}) = \sum_i x_i$ and $h\left(\mathbf{x}\right) = 1$.

Example 14.5.5. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a $N(\mu,\sigma^2)$ distribution, where $(\mu,\sigma^2)^T$ is a vector of unknown parameters. Then

$$f(\mathbf{x};\mu,\sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right]$$

$$= \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\overline{x})^2 + n(\overline{x}-\mu)^2\right].$$

We can factorize this as $T(\mathbf{x}) = \left(\overline{x}, \sum_i (x_i - \overline{x})^2\right)^T$.

14.6 Bootstrap method

Let $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ be a random sample of size n and let $\hat{\theta}(\boldsymbol{X})$ be a statistic of interest. One central task of statistical estimation is characterize the variance of $\hat{\theta}(\boldsymbol{X})$. In simple cases, we might be able to directly derive the distribution of the estimator. For example, let $\boldsymbol{X}$ be random samples of a normal distribution, the sample variance S^2 will have $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.In complex cases where obtaining standard deviation, confidence interval or even distributions of $\hat{\theta}$ is difficult. The goal of bootstrap methods is to measure the standard deviation, confidence interval, or even distributions of $\hat{\theta}$ by numerical simulation method.

On a high level, a bootstrap method of estimating the variance of an estimator consisting of the following steps

- Draw B bootstrap samples, a bootstrap sample is a set of N sample drawn from the original samples with replacement.
- On each bootstrap sample i, evaluate the estimator $\hat{\theta}(\boldsymbol{X})_i$.
- Estimate the variance of $\hat{\theta}(\boldsymbol{X})$ from $\hat{\theta}(\boldsymbol{X})_i, i = 1, ..., B$.

The intuition of the working mechanism underlying bootstrap method is the fact that we can view the bootstrap sample as a new set of samples drawn from the empirical sample distribution (the joint distribution of $(X_1, ..., X_N)$).

Remark 14.6.1 (resampling property). Given a sample of size n, we re-draw n sample with replacement. Then from probability, we have:

- The probability that ith sample is not being resampled is $(1 - \frac{1}{n})$ at the first time.
- The probability that ith sample is not being resampled is $(1 - \frac{1}{n})^n$ at the new sample of size n.
- The probability that ith sample is not being resampled is e^{-1} at the new sample when $n \to \infty$.
- On average, about ne^{-1} of original samples will not show in the new sample as $n \to \infty$.

Now we can summerize the basic procedure in a bootstrap method.

Methodology 14.6.1 (general bootstrap estimation). *Let $\hat{\theta}$ be a statistic as a function of $(X_1, ..., X_N)$. Let $\hat{\theta}_i$ be the estimation evaluated at bootstrap sample i. Then*

- *The mean estimation*

$$m = \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i \approx E[\hat{\theta}].$$

- *The variance estimation is*

$$s = \frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_i - m)^2 \approx Var[\hat{\theta}].$$

Clearly, there are two sources of error in the bootstrap estimate: The first arises from finite sample size N, and the second arises from finite B. In practice, we usually take as large as possible, say B = 10000 or $\sim N^2$. As $B \approx \infty$, $Var[\hat{\theta}]$ will converge.$Var[\hat{\theta}]$, which is determined by the original sample size N.

By properly modifying the variance estimation procedure, we arrive at the following confidence level estimation procedure.

Methodology 14.6.2 (bootstrap confidence level). *Let $\hat{\theta}$ be a statistic as a function of $(X_1, ..., X_N)$. Let $\hat{\theta}_i$ be the estimation evaluated at bootstrap sample i. Let $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_B$ be sorted. Denote $k_1 = (B \times \frac{\alpha}{2}), k_2 = (B \times (1 - \frac{\alpha}{2}))$. Then $[\hat{\theta}_{k_1}, \hat{\theta}_{k_2}]$ is the α confidence interval such that*

$$Pr(\hat{\theta}_{k_1} \leq \theta \leq \hat{\theta}_{k_2}) = 1 - \alpha.$$

Remark 14.6.2. One variation of bootstrap method is the jackknife where the standard error is estimated by from $N - 1$ leaving-one-out subsamples.

14.7 Hypothesis testing general theory

14.7.1 Basics

In statistical modeling of observed data, one may put forward a hypothesis regarding the specification of statistical models, statistical relationship among data or between different group of data, etc.

For example, a principal of a school claims that the students in his school have an average height at 5 feet. Suppose measurement of heights of 100 students, we get average height of 5.5 feet and standard deviation of 0.5 feet. Is there sufficient evidence to conclude the principal's statement?

In a typical hypothesis testing, we usually propose two **hypotheses**: **null hypothesis** and **alternative hypothesis**, denoted as H_0 and H_1. In general, null hypothesis and the alternative hypothesis are **complementary** to each other. Our goal is to determine which one is statistically correct or incorrect.

The null hypothesis is usually a simple hypothesis **the contradiction** to what we would like to prove.The alternative hypothesis is usually a hypothesis what we would like to prove. Alternative hypothesis can be **two-sided or one-sided**.

Example 14.7.1. Consider a clinical trial of a new drug. Treatment data are collected to compare two treatments.
The *null hypothesis* is usually no difference between treatments.

Depending on the purpose the test, the *alternative hypothesis* might be:

- New drug is different from old drug. (*two-sided*)
- New drug is better than old drug.(*one-sided*),
- Old drug is better than new drug. (*two-sided*).

Example 14.7.2. Given observation sampled from a normal distribution. A null hypothesis regarding the mean μ can be $\mu = \mu_0$.

And the alternative hypothesis can be

- $H_1 : \mu > \mu_0$, which is an **upper-tailed one-sided hypothesis**.
- $H_1 : \mu < \mu_0$, which is a **lower-tailed one-sided hypothesis**.
- $H_1 : \mu \neq \mu_0$, which is a **two-sided hypothesis**.

Mathematically, a hypothesis can be viewed as a statement about a population parameter θ. Let θ denote a population parameter. The general format of the null and

alternative hypothesis is $H_0 : \theta \in \Theta_0$, and $H_1 : \theta \in \Theta_1$, where Θ_0 and Θ_1 are disjoint subsets of the parameter space Θ.

Example 14.7.3. A coin is tossed and we hypothesize that it is fair. Hence Θ_0 is the set $\left\{\frac{1}{2}\right\}$ containing just one element of the parameter space $\Theta = [0, 1]$

Given H_0 and H_1, we need to decide which hypothesis to reject or accept, or which partition in Θ the population coefficient θ lies in. From this perspective, we can view hypothesis testing as a decision making problem with uncertainty.

Usually, decision is based on $p(\theta|x)$, the posterior distribution can be calculation as $p(x|\theta \in H_i)$. Given the observation data we can calculate $p(x|H_0)$ and $p(x|H_1)$. We will determine which, H_0 or H_1, is more appropriate, by either comparing $p(x|H_0)$ and $p(x|H_1)$ or specifying the value range of a test statistics T to accept or reject H_0.

A basic framework of hypothesis testing using test statistic can be summarized in the following method.

Methodology 14.7.1 (hypothesis testing via test statistic). *Suppose we are given random samples drawn from a normal distribution with known various σ^2 and unknow mean μ. A typical hypothesis testing on the μ involves the following procedures.*

- *State the Null hypothesis. For example $H0 : \mu = \mu_0$.*
- *State the Alternate Hypothesis. For example, $H1 : \mu > \mu_0$ or $\mu < \mu_0$ or $\mu \neq \mu_0$.*
- *State the significance level α, say $\alpha = 0.05$.*
- *Select the appropriate test statistic. For example,*

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

- *Determine the rejection region area [Figure 14.7.1]. For test statistic Z and $H_1 : \mu > \mu_0$, one rejection region could be $R = \{Z : Z > \phi^{-1}(1 - \alpha)\}$, where ϕ^{-1} is the inverse cdf of Z. Rejection regions for other cases can be determined similarly.*
- *Calculate the test statistic value and accept or reject H_0 based on Z. If $Z \in R$, we reject the null hypothesis.*

The **significance level** α is a probability threshold below which the **null hypothesis will be rejected under the assumption that H_0 is true**. Common values are 0.05 and 0.01.

There are different ways to specify a decision rule. In general, the decision rule depends on whether the test is an upper-tailed, lower-tailed, or two-tailed test[Figure 14.7.1]. In an upper-tailed or lower-tailed test the rejection region is around the upper or lower

tail whereH_0 will will get rejected when the test statistic is larger or smaller than a critical value, respectively. In a two-tailed test, the rejection region is at both tails where H_0 will get rejected if the test statistic is extreme, either larger than an upper critical value or smaller than a lower critical value.

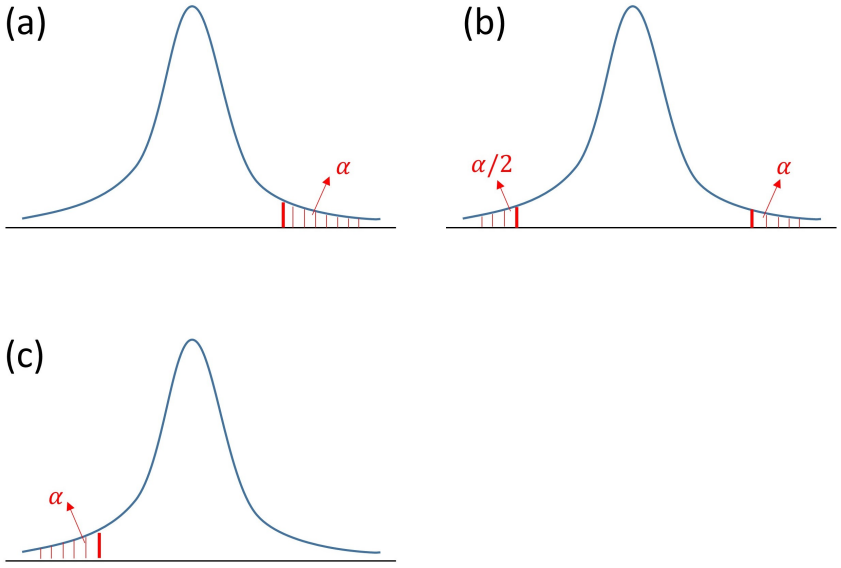


Figure 14.7.1: Demonstration for rejection regions for upper-tailed one-sided hypothesis (a), two-sided hypothesis(b), and lower-tailed one-sided hypothesis (c).

Definition 14.7.1 (p value). *The p value is the probability, assuming the null hypothesis is true, of observing the at least as extreme as (equal to or "more extreme" than) the observed test statistic in the alternative hypothesis direction.*
p value can also be interpreted as the smallest significant value that H_0 will be reject.

Methodology 14.7.2 (p value method). *Given a significance level α:*

- *If $p \leq \alpha$, then reject H_0.*
- *If $p > \alpha$, then accept H_0.*

Example 14.7.4. Consider a hypothesis test of n random samples from normal distribution $N(\mu, \sigma^2)$. Let $H_0 : \mu = \mu_0$. Let the test statistic be

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

- If $H_1 : \mu > 0$, the $p = 1 - \Phi^{-1}(Z)$. If $p \leq \alpha$, we reject H_0.
- If $H_1 : \mu < 0$, the $p = 1 - \Phi^{-1}(-Z)$. If $p \leq \alpha$, we reject H_0.
- If $H_1 : \mu \neq 0$, the $p = 2(1 - \Phi^{-1}(|Z|))$. If $p \leq \alpha$, we reject H_0.

where Φ is the cdf of a standard normal distribution.

14.7.2 Characterizing errors and power

We usually only test if H_0 is true or not and **do not test** the correctness H_1. For a binary hypothesis testing, there could be four types of results.

Definition 14.7.2 (Four types of results in binary hopothesis testing).

1. *Detection: H_0 true, decide H_0*
2. *False alarm/ **type I error:** H_0 true, we reject H_0, decide H_1.*
3. *Miss/ **type II error:** H_1 true, decide H_0(or H_0 is false, we do not reject H_0.)*
4. *Correctly rejection: H_1 true,decide H_1*

There are two types of possible error.

- A **Type I error** is the error of rejecting the null hypothesis H_0 when H_0 is true.
- A **Type II error** is the error of not rejecting the null hypothesis H_0 when H_0 is false.

We have following summary table.

	H_0 not rejected	H_0 rejected
H_0 true	no error	Type I error
H_0 false	Type II error	no error

We usually denote

$$\alpha = Pr(\text{Type I error}), \beta = Pr(\text{Type II error}).$$

Example 14.7.5 (type I, II error in hypothesis test of a normal distribution). Consider a hypothesis test of n random samples from normal distribution $N(0, \sigma^2)$. Let $H_0 : \mu = 0, H_1 : \mu > 0$. Let the test statistic be

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}.$$

Then

- To calculate type I error, we assume $H_0 : \mu = 0$ is correct, then

$$Pr(\text{type I error}) = Pr(\frac{\bar{X}}{\sigma/\sqrt{n}} > z_\alpha | H_0) = \alpha,$$

  where z_α is defined as $\Phi(z_\alpha) = 1 - \alpha$, $\Phi(z)$ is the cdf of a normal distribution.
- To calculate type II error, we assume $H_1 : \mu > 0$ is correct, then $Pr(\text{type II error}) = Pr(Z < z_\alpha | H_1)$ can be calculated in the following way:

$$\begin{aligned}\beta &= P(\frac{\bar{X}}{\sigma/\sqrt{n}} < z_\alpha) \\ &= P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}) \\ &= \Phi(z_\alpha - \frac{\mu}{\sigma/\sqrt{n}})\end{aligned}$$

  There are several critical implications:
  - If μ increases, then β decreases.
  - If n increases, then β decreases.
  - If α increases, then z_α increases and β increases.

Remark 14.7.1 (interpretation on two types of errors).

- Under the null hypothesis H_0, significance level α is the probability measure (i.e., the size) of the rejection region where H_0 will be rejected. α bounds the type I error.
- Given a fixed size of samples, it is generally not possible to minimize both types of error.
- We usually consider type I error to be worse and try to minimize or bound type I error first and then minimize type II error.
- H_0 is usually conservative statement such that reject H_0 when it is true will have **significant bad consequence**.

14.7.3 Power of a statistical test

Hypothesis testing inherently involves two type of test errors, which usually can not be minimized together. In practice, we design hypothesis and choosing signficance level following the principle that

- **Minimize the probability of committing a Type I error.** That, is minimize $\alpha = P(TypeIError)$. Typically, $\alpha \leq 0.1$.
- **Maximize the power, or reduce the type II error** Note that $\beta = P(TypeIIError) = 1 - power$, typically $\beta \leq 0.2$.

In this section, we will give a close look at the statistical power.

Definition 14.7.3 (statistical power of a test). *The power of a hypothesis test is the probability of making the correct decision if the alternative hypothesis is true. That is, the power of a hypothesis test is the probability of rejecting the null hypothesis H_0 is incorrect(or when the alternative hypothesis H_1 is true):*

$$power = P(reject\ H_0|H_1).$$

Let's revisit the previous example. Consider a hypothesis test of n random samples from normal distribution $N(0, \sigma^2)$. Let $H_0 : \mu = 0, H_1 : \mu > 0$. Let the test statistic be

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

To calculate type II error, we assume $H_1 : \mu > 0$ is correct, then $Pr(\text{type II error}) = Pr(Z < z_\alpha|H_1)$ can be calculated in the following way:

$$\begin{aligned}\beta &= P(\frac{\bar{X}}{\sigma/\sqrt{n}} < z_\alpha) \\ &= P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}) \\ &= \Phi(z_\alpha - \frac{\mu}{\sigma/\sqrt{n}})\end{aligned}$$

Since power equals $1 - \beta$, there are several critical implications:

- If μ increases, then β decreases, and power increases
- If n increases, then β decreases, and power increases
- If α increases, then z_α increases, β increases, power decreases

We have the following summary on factors affecting statistical power.

Note 14.7.1 (factors affecting statistical power). Statistical power may depend on a number of factors.

- the statistical significance criterion used in the test, i.e, α.
- the magnitude of the effect of interest in the population, i.e., μ.
- the sample size used to detect the effect, i.e., n.

Power analysis can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size. For example: "how many times do I need to toss a coin to conclude it is unfair?

Example 14.7.6 (calculating required sample size). Following previous example, the power in a normal distribution mean test is given by

$$power = 1 - \Phi(z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}),$$

If we need power to be greater than p_0, then via algebra, we can get

$$n \geq \frac{\sigma^2}{\mu^2}(z_\alpha - \Phi^{-1}(1-p_0)).$$

14.7.4 Common statistical tests

14.7.4.1 *Chi-square goodness-of-fit test*

Theorem 14.7.1 (Pearson's theorem). *Consider r boxes $B_1, ..., B_r$ and throw n balls $X_1, X_2, ..., X_n$ into these boxes independently of each other with probabilities*

$$P(X_1 \in B_1) = p_1, ..., P(X_r \in B_r) = p_r,$$

such that $p_1 + ... + p_r = 1$.

Let v_j be the number of balls in the jth box, i.e. $v_j = \sum_{i=1}^n \mathbf{1}_{X_i = B_j}$.

It follows that

- *The random variable*

$$\frac{v_j - np_j}{\sqrt{np_j}} \to N(0, 1-p_j) \text{ in distribution, as } n \to \infty$$

- *The random vector $Y = (Y_1, Y_2, ..., Y_r)$, $Y_j = \frac{v_j - np_j}{\sqrt{np_j}}$ will converge to $MN(0, \Sigma)$ in distribution, where*

$$\Sigma_{ii} = 1 - p_i, \Sigma_{ij} = -\sqrt{p_i p_j}.$$

- *The random variable*

$$\sum_{j=1}^r \frac{(v_j - np_j)^2}{np_j} \to \chi^2(r-1) \text{ in distribution, as } n \to \infty.$$

Proof. (1) Note that from Bernoulli distribution

$$E[\mathbf{1}(X_1 \in B_j)] = p_j, Var[\mathbf{1}(X_1 \in B_j)] = p_j(1-p_j).$$

By the central limit theorem

$$\frac{v_j - np_j}{\sqrt{np_j(1-p_j)}} \to N(0,1) \text{ in dist } \implies \frac{v_j - np_j}{\sqrt{np_j}} \to N(0,1-p_j) \text{ in dist.}$$

(2)

$$\begin{aligned} E[\frac{v_i - np_i}{\sqrt{np_i}}\frac{v_j - np_j}{\sqrt{np_j}}] &= \frac{1}{n\sqrt{p_ip_j}}(E[v_iv_j] - n^2p_ip_j) \\ E[v_iv_j] &= E[\sum_{l=1}^{n}\mathbf{1}(X_l \in B_i)\sum_{k=1}^{n}\mathbf{1}(X_k \in B_i)] \\ &= E[\sum_{l=1}^{n}\sum_{k=1,k\neq l}^{n}\mathbf{1}(X_l \in B_i)\mathbf{1}(X_k \in B_i)] \\ &= 2E[\sum_{l=1}^{n}\sum_{k>1}^{n}\mathbf{1}(X_l \in B_i)\mathbf{1}(X_k \in B_i)] \\ &= n(n-1)p_ip_j \\ E[\frac{v_i - np_i}{\sqrt{np_i}}\frac{v_j - np_j}{\sqrt{np_j}}] &= -\sqrt{p_ip_j}. \end{aligned}$$

(3) Note that

$$Y^TY = Z^T(I - UU^T)Z, Z \in MN(0, I_r), U = (\sqrt{p_1}, \sqrt{p_2}, ..., \sqrt{p_r}),$$

where UU^T is an rank 1 orthogonal projector($U^TU = p_1 + p_2 + \cdots + p_r = 1$).

From the chi-square decomposition theorem [Theorem 13.1.4], we know that $Y^TY \to \chi^2(r-1).in.dist.$ □

Theorem 14.7.2 (chi-square goodness-of-fit test). *Suppose that we observe an iid sample $X_1, X_2, ..., X_n$ of random variable that take a finite number of values $B_1, B_2, ..., B_r$ with unknown probabilities $p_1, p_2, ..., p_r$. Consider hypotheses*

$$\begin{aligned} H_0 &: p_i = p_i^0, for\ i = 1, 2, ..., r \\ H_1 &: for\ some\ i, p_i \neq p_i^0 \end{aligned}$$

and the test statistic

$$T = \sum_{i=1}^{r}\frac{(v_i - np_i^0)^2}{np_i^0},$$

where $v_j = \sum_{i=1}^{n}\mathbf{1}_{X_i=B_j}$.

It follows that

- *If H_0 is true, then $T \to \chi^2(r) - 1$ in dist.*
- *If H_1 is true, then $T \to \infty$, as $n \to \infty$.*
- *The decision rule is reject H_0 if $T > c$ where $c = \inf\{z : F(z) \geq 0.99\}$.*

Proof. (1) From Pearson's theorem [Theorem 14.7.1]. (2) If we write

$$\frac{(v_i - np_i^0)}{\sqrt{np_i^0}} = \sqrt{\frac{p_i}{p_i^0}} \frac{(v_i - np_i)}{\sqrt{np_i^0}} + \sqrt{n} \frac{(v_i - n(p_i - p_i^0)}{\sqrt{np_i^0}},$$

then the second quantity will diverge as $n \to \infty$. □

Note 14.7.2 (p value method for chi-square test)**.** The p-value for a chi-square test is defined as the **tail area above the calculated test statistic**.

For example, consider an experiment with test statistic result

$$T = \sum_{i=1}^{r} \frac{(v_i - np_i^0)^2}{np_i^0}.$$

Then

$$p - value = Pr(\chi^2(r-1) \geq T).$$

Given a significance level α:

- If $p \leq \alpha$, then reject H_0.
- If $p > \alpha$, then accept H_0.

14.7.4.2 *Chi-square test for statistical independence*

Lemma 14.7.1. *link*

Denote

$$p_i = \sum_{j=1}^{c} \frac{O_{ij}}{N}, q_j = \sum_{i=1}^{r} \frac{O_{ij}}{N}, E_{ij} = N p_i q_j$$

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{ij})^2}{E_{ij}} \sim \chi^2(p),$$

where $p = (r-1)(c-1)$.

The hypothesis is given by

- *H_0: U is independent of V;*

- *H_1: there exists an statistical relationship between U and V.*

14.7.4.3 *Kolmogorov-Smirnov goodness-of-fit test*

Definition 14.7.4 (Kolmogorov-Smirnov(KS) goodness-of-fit test). *The Kolmogorov-Smirnov goodness-of-fit test for a random sample of size N has the following elements:*

- *Hypothesis:*
  - *H_0: the data follow a specified **continuous** distribution with cdf $F(t)$.*
  - *H_1: the data do not follow the specified distribution.*
- *For **ascending ordered** sample $Y_1, Y_2, ..., Y_N$. KS test statistic is defined as*

$$D = \max_{1 \le i \le N} (F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i)).$$

- *The significance level α and critical value K_α.*
- *If $D > K_\alpha$, reject H_0.*

Remark 14.7.2 (interpretation and usage).

- The KS test statistic is measuring the distance of proposed distribution F is the empirical cdf given by $(i-1)/N$ and i/N.
- KS test is used for continuous distribution test. For discrete distribution test, see chi-square goodness-of-fit test [Theorem 14.7.2].
- For the KS critical value table, see link.

14.8 Hypothesis testing on normal distributions

Common notations in this sections:

- sample mean $\overline{X}$
- sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - E[X])^2$

14.8.1 Normality test

Before we proceed to hypothesis testing related to normal distributions, we need to review typical methods used to determine if samples are drawn from a normal distribution. Most straight forwards methods are qualitative graphical methods in where we plot the histogram or QQ plots. In QQ plot, we plot the quantiles of the sample against the theoretical quantiles of a standard normal distribution. For sample truly drawn from a normal distribution, we expect the plot is a perfect straight line; Different deviation from a straight line will be observed when the sample distribution is has non-zero excess Kurtosis (heavy tails or not) or skewness [Figure 14.8.1].

Additional quantitative testing methods include Shapiro–Wilk test, Kolmogorov–Smirnov test, Jarque–Bera test, Pearson's chi-squared test Theorem 14.7.1, D'Agostino's K-squared test, etc.

14.8.2 Sample mean with known variance

Consider we have n samples $X_1, ..., X_n$ for a random variable with $N(\mu, \sigma^2)$ with σ^2 known. The hypothesis testing involving the mean can be obtained by using the fact that

$$Z = \frac{\overline{X} - \mu}{\sigma / n}$$

is $N(0, 1)$. We can summarize the test as:

Table 14.8.1: Test on mean with known variance σ^2

H_0	test statistic	H_1	critical(rejection) region
$\mu \leq \mu_0$	$Z = \frac{\overline{X}-\mu}{\sigma/n}$	$\mu > \mu_0$	$z \geq z_\alpha$
$\mu \geq \mu_0$	$Z = \frac{\overline{X}-\mu}{\sigma/n}$	$\mu < \mu_0$	$z \leq -z_\alpha$
$\mu = \mu_0$	$Z = \frac{\overline{X}-\mu}{\sigma/n}$	$\mu \neq \mu_0$	$\lvert z \rvert \geq z_\alpha$

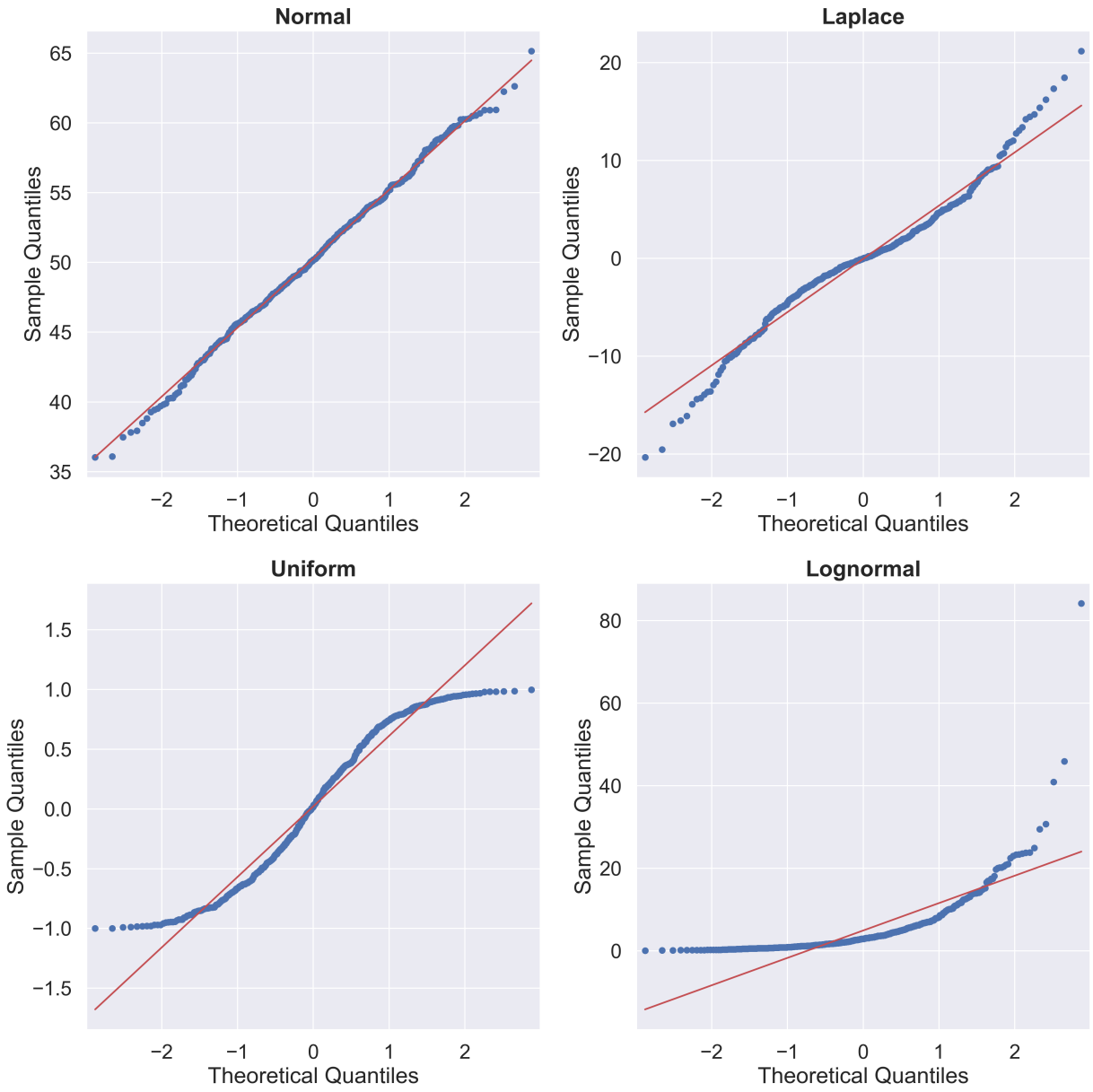


Figure 14.8.1: QQ plot with different sample distributions, including normal, Laplace ($b = 4$), uniform $U([-1, 1])$ (d) Lognormal $LN(0, 1)$. Red solid lines are the fitted linear lines.

14.8.3 Sample mean with unknown variance

Consider we have n samples $X_1, ..., X_n$ for a random variable with $N(\mu, \sigma^2)$ with σ^2 being unknown. The hypothesis testing involving the mean can be obtained by using the fact that

$$T = \frac{\overline{X} - \mu}{\sigma / n}$$

is $t(n-1)$. [Theorem 13.1.5] We can summarize the test as:

Table 14.8.2: Test on mean with unknown variance σ^2

H_0	test statistic	H_1	critical(rejection) region
$\mu \leq \mu_0$	$T = \frac{\overline{X}-\mu}{S/n}$	$\mu > \mu_0$	$t \geq t_\alpha(n-1)$
$\mu \geq \mu_0$	$T = \frac{\overline{X}-\mu}{S/n}$	$\mu < \mu_0$	$t \leq -t_\alpha(n-1)$
$\mu = \mu_0$	$T = \frac{\overline{X}-\mu}{S/n}$	$\mu \neq \mu_0$	$\lvert t \rvert \geq t_\alpha(n-1)$

14.8.4 Variance test

Consider we have n samples $X_1, ..., X_n$ for a random variable with $N(\mu, \sigma^2)$ with σ^2 being unknown. The hypothesis testing involving the mean can be obtained by using the fact that

$$T = \frac{\overline{X} - \mu}{\sigma / n}$$

is $t(n-1)$.[Theorem 13.1.5] We can summarize the test as:

Table 14.8.3: Test on variance

H_0	test statistic	H_1	critical(rejection) region
$\sigma^2 \leq \sigma_0^2$	$T = \frac{(n-1)S^2}{\sigma_0^2}$	$\sigma^2 \leq \sigma_0^2$	$t \geq \chi^2_\alpha$
$\sigma^2 \geq \sigma_0^2$	$T = \frac{(n-1)S^2}{\sigma_0^2}$	$\sigma^2 \leq \sigma_0^2$	$t \leq \chi^2_{1-\alpha}$
$\sigma^2 = \sigma_0^2$	$T = \frac{(n-1)S^2}{\sigma_0^2}$	$\sigma^2 \neq \sigma_0^2$	$t \leq \chi^2_{1-\alpha}$ or $t \geq \chi^2_\alpha$

14.8.5 Variance comparison test

Consider we have n samples $X_1, ..., X_n$ for a random variable with $N(\mu, \sigma^2)$ with σ^2 being unknown. The hypothesis testing involving the mean can be obtained by using the fact that

$$T = \frac{\overline{X} - \mu}{\sigma / n}$$

is $t(n-1)$. [Theorem 13.1.5] We can summarize the test as:

Table 14.8.4: Test on variance comparison between two samples

H_0	test statistic	H_1	critical(rejection) region
$\sigma_1^2 \leq \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$\sigma_1^2 \leq \sigma_2^2$	$t \geq \chi_\alpha^2$
$\sigma^2 \geq \sigma_0^2$	$F = \frac{S_1^2}{S_2^2}$	$\sigma_1^2 \leq \sigma_2^2$	$t \leq \chi_{1-\alpha}^2$
$\sigma_1^2 = \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$\sigma_1^2 \neq \sigma_2^2$	$F \leq \chi_{1-\alpha}^2$ or $F \geq \chi_\alpha^2$

14.8.6 Person correlation t test

Lemma 14.8.1 (Person correlation t test). *Let X and Y be random variable related by $Y = \beta X + \epsilon$, where $\beta \in \mathbb{R}$ and $\epsilon \sim N(\mu, \sigma)$. Let $\hat{\rho}$ be the correlation estimated from n samples of X and Y. Let the statistic*

$$T = \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}\sqrt{n-2}$$

follows t distribution with degree of freedom $n-2$.

Proof. Note that if we construct the linear regression model on $Y \sim \beta X$, then [Theorem 16.1.11]

$$\hat{\rho}^2 = \frac{\hat{\beta}^2 S_{XX}}{S_{YY}}, 1 - \hat{\rho}^2 = \frac{SSE}{S_{YY}}.$$

Therefore

$$\frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}\sqrt{n-2} = \sqrt{n-2}\sqrt{\frac{S_{XX}}{SSE}}\hat{\beta}$$

is the t statistics follows $n-2$ degrees of freedom [Methodology 16.1.2]. □

14.8.7 Two sample tests

14.8.7.1 *Two-sample z test*

Bbasic setup of two-sample z test:

- $X_1, X_2, ..., X_m$ is a random sample from a distribution with mean μ_1 and variance σ_1^2.
- $Y_1, Y_2, ..., Y_n$ is a random sample from a distribution with mean μ_1 and variance σ_1^2.
- X and Y samples are independent of each other.

Lemma 14.8.2 (mean difference estimator). *[8, p. 363]Let $\overline{X}$ and $\overline{Y}$ denote the sample mean.*

- *$E[\overline{X} - \overline{Y}] = \mu_1 - \mu_2$, i.e., $\overline{X} - \overline{Y}$ is the unbiased estimator of $\mu_1 - \mu_2$.*
- $$Var[(\overline{X} - \overline{Y})] = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}.$$

Proof. (1) Straight forwward. (2) Using independence, we have

$$Var[\overline{X} - \overline{Y}] = Var[\overline{X}] + Var[\overline{Y}] = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}.$$

□

14.8.7.2 *Two-sample t test*

Basic setup two-sample t test:

- $X_1, X_2, ..., X_m$ is a random sample from a distribution with mean μ_1 and variance σ_1^2.
- $Y_1, Y_2, ..., Y_n$ is a random sample from a distribution with mean μ_2 and variance σ_2^2.
- X and Y samples are independent of each other.

Lemma 14.8.3 (mean difference estimator). *[8, p. 363] The standardized variable*

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

has approximately a t distribution with degree of freedom v estimated to be (round to the nearest integer)

$$v = \frac{\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right)^2}{\frac{\left(S_1^2/m\right)^2}{m-1} + \frac{\left(S_2^2/n\right)^2}{n-1}}$$

14.8.7.3 *Paired Data*

Basic setup for paired data

- The data consists of n independently selected pairs $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$, with $E[X_i] = \mu_1$ and $E[Y_i] = \mu_2$.
- Let $D_1 = X_1 - Y_1, D_2 = X_2 - Y_2, ..., D_n = X_n - Y_n$ so that $D_i's$ are the differences within pairs.
- The $D_i's$ are assumed to be normally distributed within mean value μ_D and variance σ_D^2.

Because we assume D_i are IID samples from normal distribution, we can apply the two-sample z test or two sample t test to test if D_i has zero mean.

Note 14.8.1 (caution!). X_i could be dependent on Y_i, but pairs are independent of each other. Here we assume D_i follows normal distribution, this is not necessarily true even if X and Y are normal distributions [Corollary 13.1.1.1].

14.8.8 Interval estimation for normal distribution

Definition 14.8.1 (confidence interval). *Let $X_1, X_2, ..X_n$ denote a random sample on a random variable X, where X has pdf $f(x;\theta)$. Let $\alpha (0 < \alpha < 1)$ be given. Let $L = L(X_1, X_2, ..X_n), U = U(X_1, X_2, ..X_n)$ be two statistics. We say that the interval (L, U) is a $(1-\alpha)$ confidence interval for θ if*

$$1 - \alpha = P_\theta(\theta \in (L, U))$$

Lemma 14.8.4 (confidence interval for mean of normal random sample). *Let X be a normal random variable $N(\mu, \sigma^2)$, Let $X_1, ..., X_n$ be the random sample, let S^2 and $\overline{X}$ be the sample variance and sample mean, then*

- *If σ is known, then the $(1-\alpha)$ confidence interval for μ is*

$$(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2})$$

- *If σ is unknown, then the $(1-\alpha)$ confidence interval for μ is*

$$(\overline{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1), \overline{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1))$$

where $z_{\alpha/2}, t_{\alpha/2)}(n-1)$ are the upper critical point of $\alpha/2$ for standard normal distribution and $t(n-1)$ distribution.

Proof. (1) Use the fact that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

(2) Use the fact that

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

□

Remark 14.8.1 (knowing σ reduce uncertainty). Note that t distribution is wider(has big tails) than normal, which suggest larger confidence interval when σ is unknown.

Lemma 14.8.5 (Large sample confidence interval). *[9, p. 220] Let $X_1, ..., X_n$ be the random sample of a random variable with mean μ and variance σ^2. (Note that X is not necessarily normal). Then the $(1-\alpha)$ confidence interval for μ for large sample size is given as*

$$(\overline{X} - \frac{S}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{S}{\sqrt{n}} z_{\alpha/2})$$

Proof. When n is large, $S \approx \sigma$. Based on central limit theorem Theorem 12.13.3.

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

□

14.9 Notes on bibliography

For a graduate level overall treatment on statistical estimation theory, see [4][9]. For likelihood based methods, see [5] For large sample theory(asymptote analysis), see [10].

For introductory level Bayesian statistics, see [11].

For a good treatment on statistical estimation theory, see [12].

For a tutorial on Fisher Information matrix, see [13]

For linear regression models, see [14][15].

For multivariate statistical analysis, see [16].

For mixed models, see [17].

For an informal but deep treatment on robust statistics, see [18].

For an extensive discussion on statistical distribution, see [19][20]. For decision theory, [21][6] For an applied level treatment, see [22]

BIBLIOGRAPHY

1. Greene, W. *Econometric Analysis* ISBN: 9780134461366 (Pearson, 2017).
2. Keener, R. *Theoretical Statistics: Topics for a Core Course* ISBN: 9780387938394 (Springer New York, 2010).
3. Ruppert, D. *Statistics and data analysis for financial engineering, 2ed* (Springer, 2015).
4. Casella, G. & Berger, R. L. *Statistical inference* (Duxbury Pacific Grove, CA, 2002).
5. Pawitan, Y. *In all likelihood: statistical modelling and inference using likelihood* (Oxford University Press, 2001).
6. Moon, T. K. S. & Wynn, C. *Mathematical methods and algorithms for signal processing* **621.39: 51 MON** (2000).
7. Sciacchitano, A. *et al.* Collaborative framework for PIV uncertainty quantification: comparative assessment of methods. *Measurement Science and Technology* **26,** 074004 (2015).
8. Devore, J. L. *Probability and Statistics for Engineering and the Sciences* (Cengage learning, 2015).
9. Hogg, R. V., McKean, J. & Craig, A. T. Introduction to Mathematical Statistics, 7 ed (2012).
10. Lehmann, E. L. *Elements of large-sample theory* (Springer Science & Business Media, 1999).
11. Hoff, P. D. *A first course in Bayesian statistical methods* (Springer Science & Business Media, 2009).
12. Kay, S. M. Fundamentals of statistical signal processing, volume I: estimation theory (1993).
13. Ly, A., Marsman, M., Verhagen, J., Grasman, R. P. & Wagenmakers, E.-J. A tutorial on Fisher information. *Journal of Mathematical Psychology* **80,** 40–55 (2017).
14. Kutner, M., Nachtsheim, C. & Neter, J. *Applied Linear Regression Models* ISBN: 9780072955675 (McGraw-Hill Higher Education, 2003).
15. Seber, G. A. & Lee, A. J. *Linear regression analysis* (John Wiley & Sons, 2012).
16. Johnson, R. & Wichern, D. *Applied Multivariate Statistical Analysis* ISBN: 9780131877153 (Pearson Prentice Hall, 2007).

17. McCulloch, C. E. & Neuhaus, J. M. *Generalized linear mixed models* (Wiley Online Library, 2001).

18. Wilcox, R. R. *Fundamentals of modern statistical methods: Substantially improving power and accuracy* (Springer Science & Business Media, 2010).

19. Forbes, C., Evans, M., Hastings, N. & Peacock, B. *Statistical distributions* (John Wiley & Sons, 2011).

20. Krishnamoorthy, K. *Handbook of statistical distributions with applications* (CRC Press, 2016).

21. Young, G. A. & Smith, R. L. *Essentials of statistical inference* (Cambridge University Press, 2005).

22. Devore, J. L. *Probability and Statistics for Engineering and the Sciences* (Cengage learning, 2011).

15

MULTIVARIATE STATISTICAL METHODS

15.1 Multivariate data and distribution

15.1.1 Sample statistics

notations:

- $\mathbf{1}$ is the vector of all 1.
- J is a square matrix with all 1.

The most basic sample statistics are sample mean, sample covariance, and sample correlation, as we introduce in the following.

Let $\boldsymbol{X}$ be the data matrix such that $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T, X_i \in \mathbb{R}^p$. It follows that

- (sample mean)
$$\bar{X} \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}X^T J$$
- (sample covariance)
$$S \triangleq \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T = \frac{1}{n}X^T(I - \frac{1}{n}J)X.$$
Because the demean operation can be realized by multiplying matrix $I - \frac{1}{n}J$, which is an orthogonal projector, we can also write the sample covariance by
$$S = \frac{1}{n-1}X^T(I - \frac{1}{n}J)X.$$
- (sample correlation)
$$R = D^{-1/2}SD^{-1/2},$$
where $D = diag(S)$.

Applying affine transformation to sample statistics can still have relative simple forms.

Lemma 15.1.1 (affine transformation of sample statistics). *Let $Y = AX$ and $Z = BX$.*

- $$\bar{Y} = A\bar{X}.$$
- $$S_Y = AS_XA^T.$$

- $$S_{Y,Z} = AS_XB^T.$$

Proof. (1)

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} AX_i = A\bar{X}.$$

(2)

$$S_Y = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})^T = AS_XA^T.$$

(3)

$$S_{Y,Z} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})^T = AS_XB^T.$$

□

15.1.2 Multivariate Gaussian distribution

The most basic joint distribution used to model multiple random variables are multivariate Gaussian/normal distribution. This section, we briefly review the basic properties of multivariate normal distribution, for more detailed discussion,see subsection 13.1.5.

Definition 15.1.1 (multivariate Gaussian/normal distribution). *A random vector is said be multivariate Gaussian/normal random variable if it's pdf is multivariate Gaussian/normal distribution, whose support is $\mathbb{R}^n$ and its pdf is*

$$\rho(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|det\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$.

Example 15.1.1 (bivariate Gaussian distribution). Let $f(x,y)$ be the density of a bivariate Gaussian distribution $MN(\mu,\Sigma)$, where

$$\mu = \begin{Bmatrix} \mu_X \\ \mu_Y \end{Bmatrix}, \Sigma = \begin{Bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{Bmatrix}.$$

Then,

$$f(x,y) = \frac{\exp(-\frac{1}{2(1-\rho^2)}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\frac{\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}]).$$

Theorem 15.1.1 (affine transformation for Multivariate normal distribution). *[1, p. 183] Let X be a n dimensional random vector with $MN(\mu, \Sigma)$ distribution. Let $Y = AX + b, A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$. Then Y is an m dimensional random vector having a $MN(A\mu + b, A\Sigma A^T)$ distribution.*

Proof. Use moment generating function to prove. Let $Y = AX + b$, then from Lemma 12.7.5

$$M_Y(t) = e^{t^T b} M_X(A^T t) = e^{t^T(A\mu+b)+\frac{1}{2}t^T A\Sigma A^T t}$$

which suggesting $Y \sim MN(A\mu + b, A\Sigma A^T)$ □

Lemma 15.1.2 (orthonormal transformation maintains independence). *Let X be a n dimensional random vector with $MN(0, I)$. If C is an orthonormal matrix, then $Y = CX$ has distribution $MN(0, I)$. That is, orthonormal transformation will preserve independence.*

Proof. $Cov(Y) = C^T IC = I$. □

Lemma 15.1.3 (marginal distribution). *The multivariate Gaussian distribution $\rho(x; \mu, \Sigma)$ on $\mathbb{R}^n$ has marginal distribution on $\mathbb{R}^k, k \leq n$ given as $\rho(x_1; \mu_1, \Sigma_{11}), x_1 \in \mathbb{R}^k$ where we decompose*

$$\mu = [\mu_1, \mu_2]^T, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Proof. Use above Theorem 15.1.1. Let

$$A = \begin{bmatrix} I & 0 \end{bmatrix}$$

Then $X_1 = AX$. □

Theorem 15.1.2 (conditional distribution). *The multivariate Gaussian distribution $\rho(x;\mu,\Sigma)$ on $\mathbb{R}^n$ has marginal distribution on $\mathbb{R}^k, k \leq n$ given as*

$$\frac{\rho(x_1,x_2)}{\rho(x_2)} = \rho(x_1;\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2),\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

where we decompose

$$\mu = [\mu_1^T,\mu_2^T]^T, \Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix},$$

with $\mu_1 \in \mathbb{R}^k, \mu_2 \in \mathbb{R}^{n-k}$.

Proof. See link □

Lemma 15.1.4. *Let a p dimensional random vector $X \sim MN(\mu,\Sigma)$ with Σ being nonsingular. Then*

- *$(X-\mu)^T\Sigma^{-1}(X-\mu)$ is distributed as χ_p^2, where χ_p^2 denote the chi-square distribution with p degrees of freedom.*
- *The $MN(\mu,\Sigma)$ distribution assigns probability $1-\alpha$ to the solid ellipsoid $\{x : (x-\mu)^T\Sigma^{-1}(x-\mu) \leq F(1-\alpha)\}$, where $F(x)$ denote the cdf for chi-square distribution with p degrees of freedom.*

Proof. Straight forward. Note that $(X-\mu)^T\Sigma^{-1}(X-\mu)$ is the sum of squares of independent Gaussian standard random variables. □

15.1.3 Estimation methods

15.1.3.1 *Maximum likelihood estimation*

Lemma 15.1.5 (likelihood function). *Assume $x_1, x_2, ..., x_n, \in x_i \in \mathbb{R}^p$ are independent random samples drawn from a multivariate normal distribution (μ,Σ). Then likelihood function is given by*

$$L(\mu,\Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\exp(-Tr(\Sigma^{-1}(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T))).$$

Proof.

$$\begin{aligned} L(\mu,\Sigma) &= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp(-((\sum_{i=1}^{n}(x_i-\mu)\Sigma^{-1}(x_i-\mu)^T))) \\ &= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp(-Tr(\Sigma^{-1}(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T))) \end{aligned}$$

Note that we use $Tr(x^TAx) = Tr(Axx^T)$ [Lemma A.8.8]. □

Lemma 15.1.6 (maximum likelihood estimator). *[2, p. 172] Let $X_1, X_2, ..., X_n$ be a random sample from a multivariate normal distribution with mean μ and covariance Σ. Then,*

$$\hat{\mu} = \bar{X}, \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T = \frac{n-1}{n}S,$$

*are the **maximum likelihood estimator** of μ and Σ. That is*

$$\begin{aligned} (\hat{\mu},\hat{\Sigma}) &= \arg\max L(\mu,\Sigma) \\ &= \arg\max \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp(-Tr(\Sigma^{-1}(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T))). \end{aligned}$$

Proof. (1) Note that

$$\begin{aligned} \sum_{i=1}^{n}(X_i-\mu)(X_i-\mu)^T &= \sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-\mu)(X_i-\bar{X}+\bar{X}-\mu)^T \\ &= \sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T + \sum_{i=1}^{n}(\bar{X}-\mu)(\bar{X}-\mu)^T \\ &= \sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T + n(\bar{X}-\mu)(\bar{X}-\mu)^T \end{aligned}$$

Note that we use

$$\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-\mu)^T = (\sum_{i=1}^{n}X_i - n\bar{X})(\bar{X}-\mu)^T = 0$$

to eliminate the cross terms. Then, for the exponent in L, we have

$$Tr(\Sigma^{-1}(\sum_{i=1}^{n}(X_i-\mu)(X_i-\mu)^T)) = Tr(\Sigma^{-1}(\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T)) + n(\bar{X}-\mu)\Sigma^{-1}(\bar{X}-\mu).$$

It is easy to see that when $\mu = \bar{X}$, L is maximized **given any positive definite matrix** Σ. (2) Note that after replacing μ with $\bar{X}$, we have

$$-\ln L \approx \frac{n}{2}\ln|\Sigma| + \frac{1}{2}Tr(\Sigma^{-1}(\sum_{i=1}^{n}(x_i - \bar{X})(x_i - \bar{X})^T))$$

Taking the derivative w.r.t. Σ^{-1}, using

$$\frac{\partial}{\partial A}\ln|A| = A^{-T}, \frac{\partial}{\partial A}Tr(AB) = \frac{\partial}{\partial A}Tr(BA) = B^T,$$

we obtain

$$\frac{\partial}{\partial \Sigma^{-1}}\ln L = \frac{N}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{X})(x_i - \bar{X})^T).$$

Set the derivative to zero and we will get the results. □

15.1.3.2 *Weighted estimation*

Maximum likelihood method is sensitive to outliers. A more robust estimation method can be achieved by assigning smaller weight to samples with large deviation to the estimated mean[3, p. 97]. A typical iterative algorithm is given by algorithm 21.

Algorithm 21: Iterative reweighed estimation for multivariate normal distribution

Input: Sample set consists of $X_1, X_2, ..., X_N$

Initial estimate $\hat{\mu}(1) = \bar{X}, \hat{\Sigma}^{(1)} = S$.

Set $k = 1$.

repeat

For $i = 1, 2, ..., N$, set

$$D_i^2 = (X_i - \hat{\mu}^{(1)})^T [\hat{\Sigma}^{(1)}]^{-1} (X_i - \hat{\mu}).$$

Update location estimation using

$$\hat{\mu}^{(k+1)} = \frac{\sum_{i=1}^N w_1(D_i) X_i}{\sum_{i=1}^N w_1(D_i)},$$

where w_1 is a weight function.

Update dispersion matrix estimation using

$$\hat{\Sigma}^{(k+1)} = \frac{1}{N-1} \sum_{i=1}^N w_2(D_i^2)(X_i - \hat{\mu}^{(k+1)})(X_i - \hat{\mu}^{(k+1)})^T$$

where w_2 is a weight function.

set $k = k + 1$

until *terminal condition is met*;

Output: $\hat{\mu}, \hat{\Sigma}$

Additionally, we have a robust estimation procedure of correlation via Kendall's tau[3, p. 97]. From Lemma 15.4.22, the linear correlation is connected to the Kendall's tau via

$$\rho_\tau(X_1, X_2) = \frac{2}{\pi} \arcsin \rho.$$

The quantity $\rho_\tau(X_1, X_2)$ can be calculated via Definition 15.4.8.

15.2 Principal component analysis (PCA)

15.2.1 Statistical fundamentals of PCA

15.2.1.1 *PCA for random vectors*

Given a **zero mean** D-dimensional real-valued random vector $X = (X_1, X_2, ..., X_D)$ with covariance matrix Σ_X. PCA aims to find $d, d \ll D$ orthonormal vectors $u_i, u_i \in \mathbb{R}^D, i = 1, 2, ..., d$ such that the linearly transformed random variables $Y_i = u_i^T \Sigma_X, ...Y_D$, also known as **principal components**, has the maximum variances. Intuitively, the goal is to construct a new set of random variables or principal components, as a linear combinations of $X_1, ..., X_D$, that capture the most significant variations. Such new principal components can be further used in classification, regression and optimal control.

Mathematically, we can solve $u_1, ..., u_d$ by a series of optimization problems given by

$$
\begin{aligned}
u_1 &= \arg\max_u u^T \Sigma_X u, \ s.t. \ u^T u = 1 \\
u_2 &= \arg\max_u u^T \Sigma_X u, \ s.t. \ u^T u = 1, u^T u_1 = 0 \\
&\dots \\
u_d &= \arg\max_u u^T \Sigma_X u, \ s.t. \ u^T u = 1, u^T u_1 = 0, u^T u_2, ..., u^T u_{d-1} = 0
\end{aligned}
$$

It turns out that the optimization problems are indeed the top d eigenvectors of Σ_X, as we show in the following theorem.

Theorem 15.2.1 (principal components are top eigenvectors). *The principal components vectors $u_1, ..., u_d$ are given by the top d eigenvectors of Σ_X.*

Proof. (1) Use Reyleigh quotient theorem [Theorem 5.8.4] the top eigenvector of Σ_X maximize $u^T \Sigma_X u$ under the constraint $u^T u = 1$. (2) Use quadratic form maximization theorem [Theorem 5.12.4], we know that $u_1, ..., u_d$ are indeed the top eigenvectors. □

There are a number of critical properties regarding principal components.

Theorem 15.2.2 (principal components property).

- *Principal components $Y_i = u_i^T X, i = 1, 2, ..., d$ are uncorrelated to each other; that is*

$$cov(Y_i, Y_j) = u_i^T \Sigma_X u_j = 0, i \neq j;$$

- *Total **variance** preserved in $Y_1, ..., Y_d$ is*

$$Var[Y_1] + \cdots + Var[Y_d] = \lambda_1 + \cdots + \lambda_d,$$

where the total variance of $X_1, X_2, ..., X_D$ is

$$Var[X_1] + \cdots + Var[X_D] = \lambda_1 + \cdots + \lambda_D.$$

Proof. (1) Since u_i are eigenvector, such that

$$cov(y_i, y_j) = u_i^T \Sigma_X u_j = \lambda u_i^T u_j = 0, i \neq j.$$

(2)

$$Var[Y_1] + \cdots + Var[Y_d] = \sum_{i=1}^{d} u_i^T \Sigma_X u_i = \sum_{i=1}^{d} \lambda_i u_i^T u_i = \sum_{i=1}^{d} \lambda_i.$$

Use the property that trace of a matrix equals the sum of all its eigenvalues [Theorem 5.7.3], we have

$$Var[X_1] + \cdots + Var[X_D] = Tr(\Sigma_X) = \lambda_1 + \cdots + \lambda_D.$$

□

15.2.1.2 *Sample principal components*

In practice, we are given a set of sample point $x_1, ..., x_n, x_i \in \mathbb{R}^D$, and our goal is to find low-dimensional projected sample points $y_1, ..., y_n, y_i \in \mathbb{R}^d, , d \ll D$ via $y_i = U^T x_i, U \in \mathbb{R}^{D \times d}$ such that the variations of in $\{x_1, ..., x_n\}$ are maximally preserved [Figure 31.1.2]. The column vectors $u_1, ..., u_d \in \mathbb{R}^D$ are known as **principal component directions or principal components**, and the transformed sample point (a vector) y_i is known as **principal component scores**, with k component given by $u_k^T x_i$.

The goal of preserving sample variation can be formulated as the following optimization problem. Given a set of multi-dimensional sample point $x_1, ..., x_N \in \mathbb{R}^D$ with sample covariance matrix S. The d sample principal components are d unit vectors $u_i, u_i \in \mathbb{R}^D, i = 1, 2, ..., D, U = [u_1, ..., u_D]$ satisfying

$$\begin{aligned}
u_1 &= \arg\max_u u^T S u, \ s.t. \ u^T u = 1 \\
u_2 &= \arg\max_u u^T S u, \ s.t. \ u^T u = 1, u^T u_1 = 0 \\
&\cdots \\
u_d &= \arg\max_u u^T S u, \ s.t. \ u^T u = 1, u^T u_2 = 0, ..., u^T u_{d-1} = 0
\end{aligned}$$

Figure 15.2.1: Principal components for 2D samples.

Similar to Theorem 15.2.2, principal components $u_1, ..., u_d$ are top d eigenvectors of S defined by

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})^T.$$

In addition, there are a number of critical properties regarding principal components and principal component scores.

Theorem 15.2.3. *Let $x_1, x_2, ..., x_N \in \mathbb{R}^D$ be a set of random samples. Let U of the matrix whose columns are the top d principal components of sample covariance matrix S. Let $\lambda_1, ..., \lambda_d$ be the top d eigenvalues. The principal component scores are given by $y_i = U^T x_i$. It follows that*

- *The sample covariance matrix Σ_y of $y_1, ..., y_N, y_i \in \mathbb{R}^d$ are diagonal. The diagonal terms are given by λ_i.*
- *The total variance of principal component scores is*

$$\sum_{i=1}^{N}(y_i - \overline{y})^T(y_i - \overline{y}) = (N-1)(\lambda_1 + \lambda_2 + \cdots \lambda_d),$$

where total variance of the original sample is

$$\Delta^2 = \sum_{i=1}^{N} (x_i - \overline{x})^T (x_i - \overline{x}) = (N-1)(\lambda_1 + \lambda_2 + \cdots \lambda_D).$$

Proof. (1) First note that

$$\begin{aligned}
\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \overline{y})(y_i - \overline{y})^T &= \sum_{i=1}^{N} (U^T x_i - U^T \overline{x})(U^T x_i - U^T \overline{x})^T \\
&= \frac{1}{N-1} \sum_{i=1}^{N} U(x_i - \overline{x})^T (x_i - \overline{x})^T U^T \\
&= USU^T
\end{aligned}$$

To see the off diagonal terms are zero, we have $e_i^T USU^T e_j = u_i^T S u_j = \lambda_j u_i^T u_j = 0, i \neq j$. To see the diagonal terms, we have To see the off diagonal terms are zero, we have $e_i^T USU^T e_i = u_i^T S u_i = \lambda_i u_i^T u_i = \lambda_i$. To see the diagonal terms, we have

(2)

$$\begin{aligned}
&\sum_{i=1}^{N} (y_i - \overline{y})^T (y_i - \overline{y}) \\
&= \sum_{i=1}^{N} (U^T x_i - V^T \overline{x})^T (U^T x_i - U^T \overline{x}) \\
&= \sum_{i=1}^{N} (x_i - \overline{x})^T UU^T (x_i - \overline{x}) \\
&= \sum_{i=1}^{N} Tr((x_i - \overline{x})^T UU^T (x_i - \overline{x})) \\
&= \sum_{i=1}^{N} Tr(UU^T (x_i - \overline{x})(x_i - \overline{x})^T) \\
&= Tr(UU^T \sum_{i=1}^{N} (x_i - \overline{x})(x_i - \overline{x})^T) \\
&= (N-1) Tr(UU^T S) \\
&= Tr(U^T U \Lambda U^T U) \\
&= Tr(\Lambda) \\
&= \lambda_1 + \lambda_2 + \cdots \lambda_j
\end{aligned}$$

where we use matrix trace cyclic property [Lemma A.8.8]. □

Remark 15.2.1 (rank deficiency for high dimensional input data). When the input data $x_i \in \mathbb{R}^D$ is high dimensional such that $D \gg N$, the scattering matrix

$$S^2 = \sum_{i=1}^{N} (x_i - \overline{x})(x_i - \overline{x})^T,$$

will have rank at most $N-1$ (when columns are linearly independent) or smaller (when there are linearly dependent columns).

15.2.2 Geometric fundamentals of PCA

15.2.2.1 *Optimization approach*

Consider a set of points $X = \{x_1, x_2, ..., x_N\}, x_i \in \mathbb{R}^D$ and assume they are given by

$$x_i = \mu + U_d y_i + \epsilon_i,$$

where $\mu \in \mathbb{R}^D, U_d \in \mathbb{R}^{D \times d}$ is a matrix with independent columns, $y_i \in \mathbb{R}^d$ is the linear combination coefficients, and $\epsilon_i \in \mathbb{R}^D$ is the additional noise.

The geometric picture of such representation is that data points are approximately lying on a low-dimensional affine space, characterized by shift μ and subspace basis U_d. The goal principal component analysis is to find $\mu, U_d, \{y_i\}$, when d is given, such that the sum of squared errors is minimized.

Remark 15.2.2 (redundancy in the representation). [4, p. 19]

- They are redundancy in the above representation because of the arbitrariness in the choice of μ and U. For example, $x_i = \mu + U_d y_i = (\mu + U_d y_0) + U_d(y_i - y_0)$. We can remove this translational ambiguity by requiring $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$; therefore we are effectively dealing with demeaned data.
- Another ambiguity is due to the arbitrariness in the choice of basis spanning the subspace. We can remove this ambiguity by enforcing orthonormality in the columns of U_d.

The principal component problem can be solved by the following optimization framework.

Lemma 15.2.1 (geometric PCA in the optimization framework). *Consider a set of points $X = \{x_1, x_2, ..., x_N\}, x_i \in \mathbb{R}^D$ and further assume they are demeaned such that $\frac{1}{N}\sum_{i=1}^{N} x_i = 0$.*

- *Then finding d orthonormal basis (i.e. U_d) and $y_i, i = 1, ..., N$ that minimizes the sum of squared errors is given by*

$$\min_{U_d, \{y_i\}} \|X - U_d Y\|_F^2 = \sum_{i=1}^{N} \|x_i - U_d y_i\|^2, s.t., U_d^T U_d = I_d.$$

- *The optimization problem can be alternatively formulated as*

$$\min_{U_d} \left\|X - U_d U_d^T X\right\|_F^2, s.t., U_d^T U_d = I_d.$$

- *The necessary condition for $y_i, i = 1, ..., N$ to achieve optimality is*

$$\hat{y}_i = U_d^T x_i.$$

Proof. The Lagrangian function is given by

$$L = \sum_{i=1}^{N} \|x_i - U_d y_i\|^2 + Tr((I_d - U^T U)\Lambda),$$

where $\Lambda \in \mathbb{R}^{d \times d}$ is the matrix of Lagrange multipliers. Then we have

$$\frac{\partial L}{\partial y_i} = 0 \implies -2U_d^T(x_i - U_d y_i) = 0 \implies y_i = U_d^T x_i,$$

where we use the fact that $U_d^T U_d = I_d$. □

Theorem 15.2.4 (PCA via SVD). *[4, p. 21]Let $X = [x_1, x_2, ..., x_N] \in \mathbb{R}^{D \times N}$ be the data matrix. Let $X = U\Sigma V^T$ be the singular value decomposition (SVD) of X. Then for any given $d < D$, a solution to PCA is the first d columns of$U$, given as $U_d = [u_1, u_2, ...u_d]$ and $\{y_i\}$ is the top $d \times N$ sub matrix $\Sigma_d V_d^T$ of the matrix ΣV^T (each column of length d is one y_i and in y_i each row is scaled in $\sqrt{\lambda_i}$).*

Proof.

$$\begin{aligned}
\left\|X - UU^T X\right\|_F^2 &= Tr((X - UU^T X)^T (X - UU^T X)) \\
&= Tr(X^T X) - Tr(XUU^T X) - Tr(UU^T X^T X) + Tr(X^T UU^T UU^T X) \\
&= Tr(X^T X) - 2Tr(XUU^T X) + Tr(X^T UU^T X) \\
&= Tr(X^T X) - Tr(X^T UU^T X)
\end{aligned}$$

Note that $Tr(XUU^T X) = Tr(U^T XX^T U) = \sum_{i=1}^{d} u_i^T XX^T u_i$. From Rayleigh quotient theorem [Theorem 5.8.4], we know that maximum of $Tr(U^T XX^T U$ is attained when u_i are the top d eigenvectors. □

Remark 15.2.3 (pitfalls for statistical approach and geometrical approach). Let X be the data matrix $X \in \mathbb{R}^{p\times N}$. In statistical approach, we calculate principal components by eigen-decomposition or SVD from sample covariance matrix of X, in which X **is demeaned**.

In the geometrical approach, if we directly perform SVD on X without demeaning X, we will get different results.

15.2.2.2 *Properties*

Theorem 15.2.5 (representation in the principal component space). *Let $X = [x_1, x_2, ..., x_N] \in \mathbb{R}^{D\times N}$ be the data matrix. Let $X = U\Sigma V^T$ be the singular value decomposition (SVD) of X.*

- *Then the coordinate vector of x_i in the basis of $U \in \mathbb{R}^{D\times p}$ is given by*

$$y_i = U^T x_i, y_i \in \mathbb{R}^p.$$

- *The matrix $P = UU^T$ is an orthonormal projection matrix that projects a vector in $\mathbb{R}^D$ to a subspace $S \subseteq \mathbb{R}^D$, where S has the basis U. We have recovery representation*

$$Uy_i = UU^T x_i = x_i.$$

- *If X has D independent columns, then $U \in \mathbb{R}^{D\times D}$, then $P = I$*

Proof. (2) Since $X = U\Lambda V^T$, then $x_i \in S$. (x_i can be written as a linear combination of columns vectors of U, and the coefficients are the i column of the matrix ΛV^T). Therefore $x_i = Px_i$. □

Theorem 15.2.6 (PCA distance and inner product preserving properties). *Let $X = [x_1, x_2, ..., x_N] \in \mathbb{R}^{D\times N}$ be the data matrix. Let $X = U\Sigma V^T$ be the singular value decomposition (SVD) of X. It follows that*

- *(distance preservation) For each data point x_i, let $y_i = U^T x_i$ be the (new) coordinates projected on the eigen-basis.*

$$\left\|y_i - y_j\right\|^2 = \left\|x_i - x_j\right\|^2.$$

- *Let U_d be the matrix whose columns are the top d eigenvectors, and let $y_i = U_d^T x_i$. Denote $P_d = U_d U_d^T$. Then*

$$\left\|y_i - y_j\right\|^2 = \left\|P_d(x_i - x_j)\right\|^2 \leq \left\|P_d(x_i - x_j)\right\|^2.$$

- *(beat d rank approximation to inner product matrix) Let $y_i = U_d^T x_i$. Denote $Y = [y_1, ..., y_N]$. Then $Y^T Y$ is the best d rank approximation to $X^T X$ Moreover, $\left\|Y^T Y\right\|_F^2 = \sum_i = 1^d \sigma_i^2$, $\left\|X^T X\right\|_F^2 = \sum_i = 1^D \sigma_i^2$*

Proof. (1)

$$\begin{aligned}\left\|y_i - y_j\right\|^2 &= (y_i - y_j)^T (y_i - y_j) \\ &= (U^T x_i - U^T x_j)^T (U^T x_i - U^T x_j) \\ &= (x_i - x_j)^T U U^T (x_i - x_j) \\ &= (x_i - x_j)^T (x_i - x_j)\end{aligned}$$

where we used the factor that $x_i, x_j, x_i - x_j$ are all lying inside the subspace S spanned by U, and the orthogonal projection matrix to S is $P = UU^T$. Therefore $P(x_i - x_j) = (x_i - x_j)$. (2) Similar to (1). We then use matrix norm inequality [Theorem 5.13.2] that gives

$$\left\|A(x_i - x_j)\right\|^2 \leq \|A\|^2 \left\|x_i - x_j\right\|^2 = \sigma_1(P_d) \left\|x_i - x_j\right\|^2 = \left\|x_i - x_j\right\|^2,$$

where we use the fact that largest eigenvalue of orthogonal projector matrix P_d is 1. (3) Note that $Y^T Y = X^T U_d U_d^T X = V_d^T \Sigma_d^2 V_d$, and $X^T X$ has SVD $X^T X = V^T \Sigma^2 V$. Therefore, $Y^T Y$ is the best d rank approximation to $X^T X$.

The conclusion that $\left\|Y^T Y\right\|_F^2 = \sum_i = 1^d \sigma_i^2$ is from the relationship between SVD and Frobenius norm [Theorem 5.9.2].

□

Remark 15.2.4. It is not possible to derive a lower bound for the ration $\frac{\left\|y_i - y_j\right\|^2}{\left\|x_i - x_j\right\|^2}$ because if $x_i - x_j$ lies in the null space of P_d, then the ratio can be zero.

15.2.3 Probabilistic PCA

In the previous sections, we have addressed two approaches to PCA, one is statistical approach that seeks principal components that preserve variation and another is geometry

approach that view principal components as the orthogonal basis of a low-dimensional affine subspace which the data primarily lie in. This section, we introduce probabilistic PCA [5], which aims to characterize data structure from probabilistic data generation perspective.

Probabilistic PCA offers several flexibility in practical use of PCA, including

- Multiple probabilistic PCA models can be combined as a probabilistic mixture.
- Maximum-likelihood estimates can be computed for elements associated with principal components.
- Probabilistic PCA can handle missing data or incomplete data in a more formal probabilistic framework, rather than using ad-hoc methods.
- Probabilistic PCA can be used to generate new data samples.

Suppose we are given N data points $\{x_1, ..., x_N\}, x_i \in \mathbb{R}^D$, **probabilistic PCA** specifies a data generation model for X, given by

- $X \sim WY + \mu + \epsilon$, where $W \in \mathbb{R}^{D\times d}$, $Y \in MN(0, I_d), \epsilon \sim MN(0, \sigma^2 I), \mu \in \mathbb{R}^n$.
- $X \sim MN(\mu, C_Y), C_Y = WW^T + \sigma^2 I$

Usually, we call $Y = (Y_1, ..., Y_d)$ **latent factors**.

Remark 15.2.5 (interpretation).

- Probabilistic PCA can be viewed a latent factor model with X being the latent variable.
- Normal PCA is a limiting case of probabilistic PCA, taken as the limit as the covariance of the noise becomes infinitesimally small ($\sigma^2 \to 0$).

One goal in probabilistic PCA is to estimate parameters W, μ, σ.

Lemma 15.2.2 (likelihood function). *Assume $x_1, x_2, ..., x_n, \in x_i \in \mathbb{R}^D$ are independent random samples drawn from a multivariate normal distribution (μ, Σ). Then likelihood function is given by*

$$L(\mu, \Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp(-Tr(\Sigma^{-1}(\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T))).$$

Proof.

$$\begin{aligned} L(\mu, \Sigma) &= \frac{1}{(2\pi)^{nD/2}|\Sigma|^{n/2}} \exp(-((\sum_{i=1}^{n}(x_i - \mu)\Sigma^{-1}(x_i - \mu)^T))) \\ &= \frac{1}{(2\pi)^{nD/2}|\Sigma|^{n/2}} \exp(-Tr(\Sigma^{-1}(\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T))) \end{aligned}$$

Note that we use $Tr(x^T Ax) = Tr(Axx^T)$ [Lemma A.8.8]. □

Similar to Lemma 15.1.6 and follow **[??]**, we can show

$$\hat{W} = U_d \left(\Lambda_d - \sigma^2 \mathbf{I} \right)^{1/2} R$$

where U_d consists the top d eigenvectors of S, and the top d corresponding eigenvalues are in the diagonal matrix Λ_q, and R is an arbitrary $d \times d$ orthogonal rotation matrix.

And

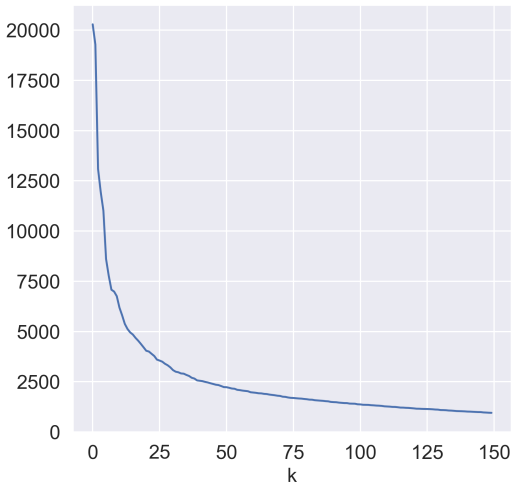
$$\hat{\sigma}^2 = \frac{1}{D-d} \sum_{j=d+1}^{D} \lambda_j.$$

15.2.4 Applications

15.2.4.1 *Eigenfaces and eigendigits*

We first apply PCA to the face image from Labeled Faces in the Wild to extract top 'eigenfaces', which are faces can be used to synthesize the majority of human faces by linear combination. We perform PCA on around one thousand randomly selected face images. Each image has 50×37 pixels. The PCA results are in Figure 15.2.2. Note that the first several components capture the major variations of image data. On the other hand, Components associated with smaller singular values mostly captures noise.

(a) Raw face image examples.

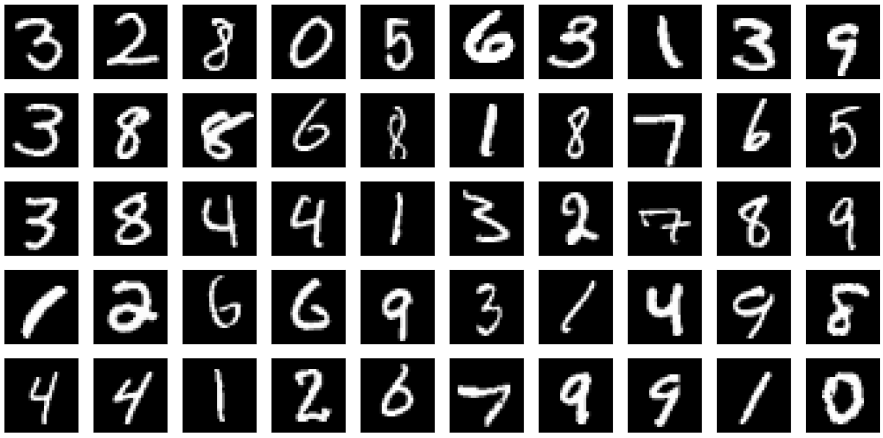


(b) Singular value spectrum.

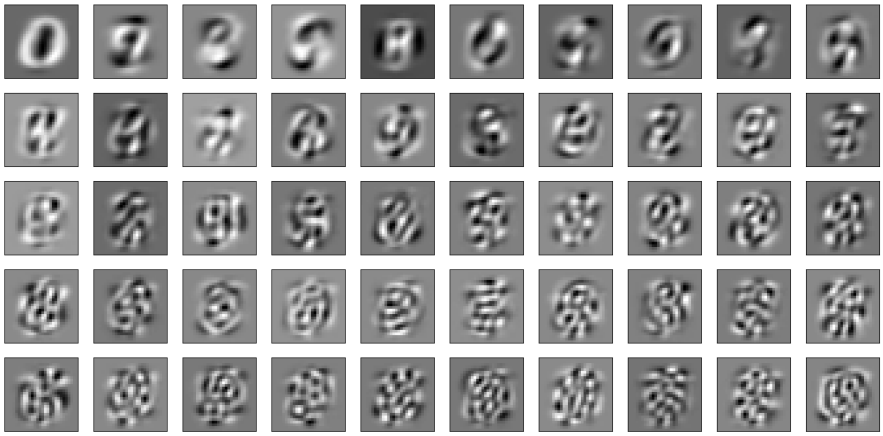
(c) Top eigenfaces.

Figure 15.2.2: PCA eigenface analysis.

Similar PCA on the MNIST data set is showed in Figure 15.2.3.



(a) Raw digit image examples.



(b) Top 50 eigen-digits.

Figure 15.2.3: PCA eigen-digit analysis for MNIST dataset.

15.2.4.2 *Interest rate curve dynamics modeling*

PCA method has found important application in financial industry. One example is to modeling the interest rate curve dynamics. It is common knowledge that in borrowing and lending business, the interest rate we borrow or lend is not a constant independent of

the length of loan. Instead, the interest rate is a function of life of the loan, which takes into account the economy, credit risk and other factors.

At every date t, the interest rates can be represented by a curve, as showed in Figure 15.2.4, whose x is swap tenor approximating the meaning the loan life and y axis is the borrowing or lending rate.

At different t, we observe different curve shape. Usually, the interest rate curve is represented by a vector of points on the curve, denoted by $y \in \mathbb{R}^N$, where N is the number of different tenors.

One brute force way is to model the curve dynamics is to model $y_1, ...y_N$ altogether, leading to a N variable model like

$$y_i(t + dt) = y_i(t) + \sigma_i z_i, i = 1, ..., N,$$

where z_i is zero mean random shock, and $z_1, ..., z_N$ has certain joint distribution to capture the correlation structure in $y_1, ..., y_N$. The full model is usually impractical because the difficulties in accurately estimating covariance structure of $y_1, ..., y_N$ when N is large.

Another low-dimensional modeling approach is to exploit that the high correlation nature among $y_1, ..., y_N$ and seek some model like

$$y_i(t + dt) = y_i(t) + \sum_{j=1}^{k} b_i^{(j)} z_j, i = 1, ..., N,$$

where $z_1, ..., z_k$ are k zero mean random shock **with zero correlation**, and $b^{(1)}, ..., b^{(k)} \in \mathbb{R}^N$ are top k eigenvectors obtained for historical covariance matrix of $y_i(t + dt) - y_i(t)$. Using PCA can lead to much more robust estimation that capture the covariance structure.

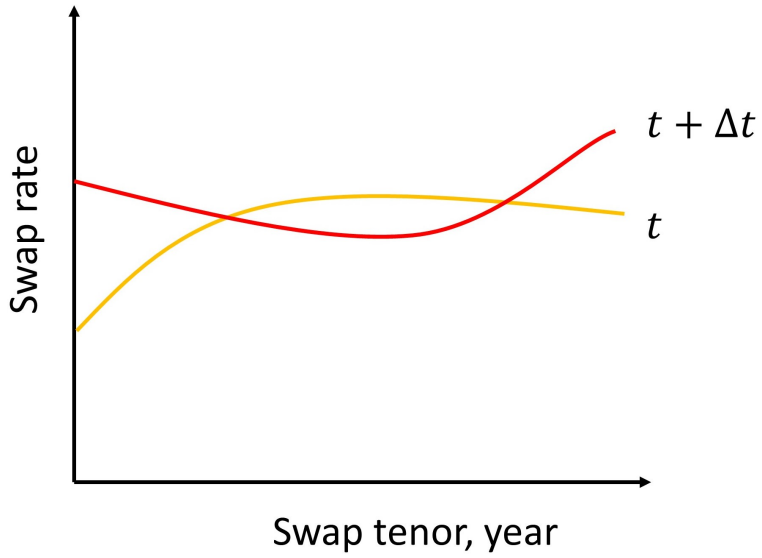


Figure 15.2.4: Demonstration of interest rate curve dynamics.

As a demonstration, we now analyze **daily interest rates** with maturities of 1 year, 2 years, 3 years, 4 years, 5 years, 7 years, 10 years, and 30 years observing between 2000 and 2011. Table 15.2.1 shows the PCA factors and the associated eigenvectors. Figure 15.2.5 plots the first three dominating PCA factors. The first eigenvector/factor has deeper interpretations, they represent the parallel shift mode, the steepening model, and the bending mode that dominating curve changing dynamics.

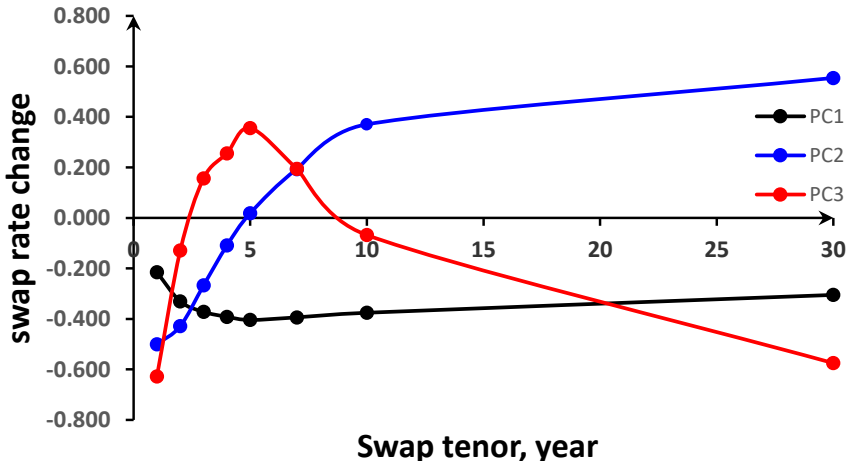


Figure 15.2.5: Demonstration of first three dominating PCA factor in the swap rate curve daily change.

Table 15.2.1: Eigenvectors and eigenvalues for swap rate daily change

(a) Eigenvectors for swap rate daily change

	PC1	PC2	PC3	···	PC8
1Y	0.216	-0.501	0.627	···	-0.034
2Y	0.331	-0.429	0.129	···	0.236
3Y	0.372	-0.267	-0.157	···	-0.564
4Y	0.392	-0.110	-0.256	···	0.512
5Y	0.404	0.019	-0.355	···	-0.327
7Y	0.394	0.194	-0.195	···	0.422
10Y	0.376	0.371	0.068	···	-0.279
30Y	0.305	0.554	0.575	···	0.032

(b) Eigenvalues for swap rate daily change

	PC1	**PC2**	**PC3**	···	**PC8**
eigenvalue	4.77	2.08	1.29	···	-0.034

15.3 Canonical correlation analysis

15.3.1 Basics

CCA looks for linear combinations of variables in the two groups that are highly correlated with each other. Consider an example that we want to study the factors affecting student's academic performance. Let X be the random vector of activities from N students. Each one quantifies the time spents on sports, electronics, reading, and homework. Let Y be the random vector of academic performance from N students. Each one quantifies the performance on reading, math, and language. In CCA, we seek vectors a and b such that $Corr(a^T X, b^T Y)$ is maximized. The results are useful in the following ways:

- Examine the overall relationship between the two set of variables. Suppose $\max Corr(a^T X, b^T Y)$ is low, then it suggests that the two set of variables are statistically irrelevant, i.e., activities are not affecting academic performance.
- Examine the coefficients a and b to understand relationships between individual components. Suppose the result is $0.6 * Reading + 0.9 * Math + 0.7 * Logic$ is highly correlated with $0.5 * Sports - 2 * Electronics + 0.8 * Reading + 5 * Homework$, then we can understand that spending too much time in electronics will negatively affect academic performance and spending more time on homework will have positive impact.
- Seek low dimensional representations. Following previous point, we can view $0.6 * Reading + 0.9 * Math + 0.7 * Logic$ as a new quantity that characterize academic performance and $0.5 * Sports - 2 * Electronics + 0.8 * Reading + 5 * Homework$ as a new quantity characterizing activities.

More formally, given two column vector of random variables $X = (X_1, X_2, ..., X_p)$ and $Y = (Y_1, Y_2, ..., Y_q)$ with finite second moments. **Canonical correlation analysis** seeks vector $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^n$ such that the random variable $a^T X$ and $b^T Y$ has the maximum correlation. More formally, we want to solve

$$\max_{a \in \mathbb{R}^p, b \in \mathbb{R}^q} corr(a^T X, b^T Y) = \max_{a \in \mathbb{R}^p, b \in \mathbb{R}^q} \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_{XX} a} \sqrt{b^T \Sigma_{XX} b}}.$$

To understand the maximization problem better, we first examine its equivalent optimization problems before we proceed to its solution.

Lemma 15.3.1 (equivalent optimization problems). *Let $c = \Sigma_{XX}^{1/2} a, d = \Sigma_{YY}^{1/2} b$, then the original optimization problem can be written by*

$$\max_{c \in \mathbb{R}^p, d \in \mathbb{R}^q} \frac{c^T \Sigma_{XX}^{-1/2} \Sigma_{XY} \sigma_{YY}^{-1/2} d}{\sqrt{c^T c}\sqrt{d^T d}}.$$

We can equivalently write the unconstrained optimization as

$$\max_{c \in \mathbb{R}^p, d \in \mathbb{R}^q} c^T \Sigma_{XX}^{-1/2} \Sigma_{XY} \sigma_{YY}^{-1/2} d$$
$$\text{subject to } c^T c = 1, d^T d = 1$$

Proof. Straight forward. □

Theorem 15.3.1 (solution to canonical correlation problem).

- *The optimizer of the transformed optimization problem is given by*
  - *c is eigenvector of $\Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1/2}$ associated with the largest eigenvalue.*
  - *d is proportional to $\Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1/2} c$*

  Or reciprocally,
  - *d is eigenvector of $\Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1/2}$ associated with the largest eigenvalue.*
  - *c is proportional to $\Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} d$*
- *The optimizer of the original optimizer problem is*
  - *a is eigenvector of $\Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX}$ associated with the largest eigenvalue.*
  - *b is proportional to $\Sigma_{YY}^{-1} \Sigma_{YX} c$*

  Or reciprocally,
  - *b is eigenvector of $\Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}$ associated with the largest eigenvalue.*
  - *a is proportional to $\Sigma_{XX}^{-1} \Sigma_{XY} b$*

Proof. Using Cauchy-Schwartz inequality [Theorem 12.10.4],

$$\begin{aligned}&(c^T \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2}) d \\ \leq &(c^T \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX})^{1/2} (d^T d)^{1/2}\end{aligned}$$

Therefore,

$$\rho \leq \frac{(c^T \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX})^{1/2}}{(c^T c)^{1/2}}$$

Base on Rayleigh quotient [Theorem 5.8.4], ρ will take the maximum value when c is eigenvector of $\Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1/2}$ associated with the largest eigenvalue.

□

Corollary 15.3.1.1 (solution when components are standardized). *Suppose $\Sigma_{XX} = I$ and $\Sigma_{YY} = I$. Then the solution to the following optimization problem is given by*

$$\max_{a\in\mathbb{R}^m, b\in\mathbb{R}^n} corr(a^T X, b^T Y) = \max_{a\in\mathbb{R}^m, b\in\mathbb{R}^n} \frac{a^T\Sigma_{XY}b}{\sqrt{a^T\Sigma_{XX}a}\sqrt{b^T\Sigma_{XX}b}}.$$

- *a is eigenvector of $\Sigma_{XY}\Sigma_{YX}$ associated with the largest eigenvalue.*
- *b is eigenvector of $\Sigma_{YX}\Sigma_{XY}$ associated with the largest eigenvalue.*

In practice, we collect n random samples. Let $X \in \mathbb{R}^{n\times p}, Y \in \mathbb{R}^{n\times q}$ be the **centered** data matrix. Then the optimization will be maximizing sample correlations, which is given by

$$\begin{aligned}&\max_{c\in\mathbb{R}^p, d\in\mathbb{R}^q} c^T(X^TX)^{-1/2}(X^TY)(Y^TY)^{-1/2}d\\ &subject\ to c^Tc = 1, d^Td = 1\end{aligned}$$

15.3.2 Sparse CCA

In some applications such as bioinfomatics [6], we often encounter the situation where $\min(p, q) \gg n$. This leads to the case where $(X^TX)^{-1}$ and $(Y^TY)^{-1}$ do not exist. As a result, we need to use the original optimization form. Further, we might want to induce sparsity in a, b. Eventually, our optimization problem becomes

$$\begin{aligned}&\max_{a\in\mathbb{R}^p, b\in\mathbb{R}^q} a^TX^TYb\\ &subject\ to\ a^TX^TXa = 1, b^TY^TYb = 1,\\ &\|a\|_1 \le c_1, \|b\|_1 \le c_2.\end{aligned}$$

We can relax the constraint to yield a convex optimization given by

$$\begin{aligned}&\max_{a\in\mathbb{R}^p, b\in\mathbb{R}^q} a^TX^TYb\\ &subject\ to\ a^TX^TXa \le 1, b^TY^TYb \le 1,\\ &\|a\|_1 \le c_1, \|b\|_1 \le c_2.\end{aligned}$$

Note that this optimization problem is bi-convex. That is, when we fix a, the optimization is convex with respect to b, and vice versa.

This optimization is derived from the well-known as penalized matrix factorization problem[7, 8], and have a general iterative algorithm below [algorithm 22].

Algorithm 22: Alternating sparse canonical correlation analysis algorithm

Input: Data matrix X, Y
Initialize a to have $a^T XXa = 1$.
repeat
    Solve $b = \arg\max_b a^T X^T Yb$, subject to $b^T Y^T Yb \leq 1, \|b\|_1 \leq c_2$.
    Solve $a = \arg\max_a a^T X^T Yb$, subject to $a^T X^T Xa \leq 1, \|a\|_1 \leq c_1$.
until *terminal condition is met*;
Output: a, b

15.4 Copulas and dependence modeling

15.4.1 Definitions and properties

Definition 15.4.1. *A **copulas** $C : [0,1]^d \to [0,1]$ is a multivariate CDF of a $d-$dimensional random vector on the unit cube whose univariate marginal distributions are all $U(0,1)$.*

*The **density of a copulas**, denoted by c, is given by,*

$$c(u_1, u_2, ..., u_d) = \frac{\partial^d}{\partial u_1 \partial u_2 ... \partial u_d} C(u_1, u_2, ..., u_d).$$

Note 15.4.1 (marginal distribution function from joint distribution function)**.**

$$F_1(y_1) = \lim_{y_2,...,y_d \to \infty} F(y_1, y_2, ..., y_d)$$

Lemma 15.4.1 (basic properties of Copulas).

- $$C(u_1, ..., u_{i-1}, 0, u_{i+1}, ..., u_d) = 0$$
- $$C(1, ..., 1, u, 1, ..., 1) = u$$
- *C is non-decreasing in each of the d variables.*
- *(marginalization property) Let $C_{1:d}(u_1, u_2, ..., u_d)$ be copula(i.e., joint cdf) associated with uniform random variables $U_1, U_2, ..., U_d$, then*

$$C_{1:k}(u_1, u_2, ..., u_k) \triangleq C_{1:d}(u_1, u_2, ..., u_k, 1, ..., 1), k < d,$$

is the marginal cdf associated with uniform random variables $U_1, U_2, ..., U_k$.

Proof. Directly from the property of multivariate cdf. □

Corollary 15.4.0.1. *Let $C : [0,1] \times [0,1] \to [0,1]$ be a bivariate copula. Then*

$$C(0,u) = C(u,0) = 0, C(1,u) = C(u,1) = u$$

and

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0, \forall 0 \leq u_1 \leq u_2 \leq 1, 0 \leq v_1 \leq v_2 \leq 1.$$

Lemma 15.4.2 (probability integral transform for a random variable). *Let X be a random variable with support $\mathbb{R}$ and let $F_X : \mathbb{R} \to [0,1]$ be its cdf.*

It follows that

- *The new random variable $Y = F_X(X)$ has an uniform distribution.*
- *Let U be a uniform random variable on $[0,1]$, then*

$$X = F_X^{-1}(U).$$

- *Let ϕ be the cdf of a standard normal, then the random variable defined by*

$$Z = \phi^{-1}(F_X(X)).$$

is a Gaussian random variable.

Proof. (1)Note that

$$P(Y \leq y) = Pr(F_X(X) \leq y) = Pr(X \leq F_X^{-1}(y)) = F_X[F_X^{-1}(y)] = y.$$

(2) Note that $Pr(X < x) = Pr(F_X^{-1}(U) < x) = Pr(U < F_X(x)) = F_X(x)$. where we use the fact that $Pr(U < y) = y, \forall y \in [0,1]$. (3) use (1) and (2). □

Example 15.4.1. If X has an exponential distribution with unit mean, then

$$F_X(x) = 1 - \exp(-x).$$

It follows that the random variable Y, defined as

$$Y = 1 - \exp(-X)$$

has a uniform distribution.

Lemma 15.4.3 (probability transform for a random vector and its properties). *Consider a random vector $(X_1, X_2, ..., X_d)$. Suppose its marginal cdfs $F_i(x) = P(X_i \leq x)$ are continuous functions. Then the random vector*

$$(Y_1, Y_2, ..., Y_d) \triangleq (F_1(X_1), F_2(X_2), ..., F_d(X_d))$$

has uniformly distributed marginals. Moreover, the joint cdf of $(Y_1, Y_2, ..., Y_d)$ *is the copulas associated with the joint cdf of* $(X_1, X_2, ..., X_d)$.

Proof. (1) Note that from the definition of marginal cdf, we have

$$P(Y_1 < y_1, Y_2 < \infty, ..., Y_d < \infty) = F_{Y_1, Y_2, ..., Y_d}(y_1, \infty, ..., \infty) = F_{Y_1}(y_1).$$

$F_{Y1}(y_1)$ is a uniform distributed cdf, as showed in Lemma 15.4.2. (2) (a)To show the copula is the same we can use monotone transformation invariance Lemma 15.4.7 since F_i is increasing;(b)

$$\begin{aligned} &Pr(F_1(x_1) < u_1, ..., F_d(x_d) < u_d) \\ =&Pr(x_1 < F_1^{-1}(u_1), ..., x_d < F_d^{-1}(u_d)) \\ =&F_X(F_1^{-1}(u_1), ..., F_d^{-1}(u_d)) \end{aligned}$$

That is, the joint cdf of $(Y_1, Y_2, ..., Y_d)$is the copulas associated with the joint cdf of $(X_1, X_2, ..., X_d)$ [Theorem 15.4.2]. □

Lemma 15.4.4 (Frechet-Hoeffding bounding). *[3, p. 189]*

$$W(u_1, u_2, ..., u_d) \leq C(u_1, u_2, ..., u_d) \leq M(u_1, ..., u_d)$$

where

$$W(u_1, u_2, ..., u_d) = \max(1 - d + \sum_{i=1}^{d} u_i, 0)$$

and

$$M(u_1, u_2, ..., u_d) = \min(u_1, u_2, ..., u_d)$$

In particular, for $d = 2$*, we have*

$$\max(u + v - 1, 0) \leq C(u, v) \leq \min(u, v).$$

Proof. (1)

(2) For the lower bound, we have

$$\begin{aligned} C(u_1, ..., u_d) &= Pr(\cap_{1\leq i\leq d}\{U_i \leq u_i\}) \\ &= 1 - Pr(\cup_{1\leq i\leq d}\{U_i > u_i\}) \\ &= 1 - \sum_{i=1}^{d} Pr(U_i > u_i) \\ &= 1 - \sum_{i=1}^{d}(1 - u_i) \\ &= 1 - d + \sum_{i=1}^{d} u_i \end{aligned}$$

where we use Demorgan's law [Lemma 1.1.2] and union bound. □

Corollary 15.4.0.2. *For a multivariate joint cdf F with margins $F_1, F_2, ..., F_d$, we have*

$$\max(1 - d + \sum_{i=1}^{d} F_i(x_i), 0) \leq F(x_1, ..., x_d) \leq \min(F_1(x_1), ..., F_d(x_d))$$

Proof. Note that

$$W(F_1(x_1), ..., F_d(x_d)) \leq C(F_1(x_1), ..., F_d(x_d)) \leq M(F_1(x_1), ..., F_d(x_d)),$$

and

$$C(F_1(x_1), ..., F_d(x_d)) = F(x_1, ..., x_d).$$

□

Example 15.4.2 (Gaussian copula).

$$C_\rho(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)})dxdy$$

where ρ is the linear correlation coefficient.

Example 15.4.3 (Student's t copula).

$$C_{\rho,\nu}(u, v) = \int_{-\infty}^{t^{-1}(u)} \int_{-\infty}^{t^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp(1 + \frac{x^2 - 2\rho xy + y^2}{\nu(1-\rho^2)})^{-(\nu+2)/2}dxdy$$

where ρ is the linear correlation coefficient.

15.4.2 Copulas and distributions

Note 15.4.2. In this section we define inverse function for a cdf as

$$F^{-1}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t\}.$$

15.4.2.1 *Fundamentals*

Definition 15.4.2 (copula associated with a distribution). *[9, p. 3] Let $F(x_1, x_2, ..., x_n)$ be a cdf of a random vector $(X_1, X_2, ..., X_n)$ with support $\mathbb{R}^n$. Let $F_1, F_2, ..., F_n$ be the corresponding marginal cdf. The copula $C(u_1, u_2, ..., u_n)$ associated with F is such that*

$$F(x_1, x_2, ..., x_n) = C(F_1(x_1), F_2(x_2), ..., F_n(x_n)).$$

Theorem 15.4.1 (construct a multivariate cdf from a copulas and margins). *Consider d random variables $X_1, X_2, ..., X_d$ with $F_1, F_2, ..., F_d$ being the univariate cdf. Let C be a d dimensional copulas, then we can construct a d-dimensional multivariate cdf as*

$$F(x_1, x_2, ..., x_n) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d))$$

such that the marginal cdf is given by $F_1, F_2, ..., F_d$.

Proof. (1)Use the definition of marginal cdf and Lemma 15.4.1, we have

$$\begin{aligned} F_1(x_1) &\triangleq F(x_1, \infty, ..., \infty) \\ &= C(F_1(x_1), F_2(\infty), ..., F_n(\infty)) \\ &= F_1(x_1). \end{aligned}$$

□

Remark 15.4.1 (back out the copula from constructed cdf). Note that we construct a new cdf F from a given copula $C(u_1, u_2, ..., u_d), u_i \in [0, 1]$ via

$$F(x_1, x_2, ..., x_n) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d)).$$

On the other hand, Theorem 15.4.2 gives a way to back out the original copula from joint cdf via

$$\begin{aligned} C_n(u_1, u_2, ..., u_d) &= F(F_1^{-1}(u_1), F_2^{-1}(u_2), ..., F_d^{-1}(u_d)) \\ &= C_o(F_1F_1^{-1}(u_1), F_2F_2^{-1}(u_2), ..., F_dF_d^{-1}(u_d)) \\ &= C_o(F_1F_1^{-1}(u_1), F_2F_2^{-1}(u_2), ..., F_dF_d^{-1}(u_d)) \\ &= C_o(u_1, u_2, ..., u_d) \end{aligned}$$

where $u_1, u_2, ..., u_d \in [0, 1]$ and F is given by

$$F(x_1, x_2, ..., x_n) = C_o(F_1(x_1), F_2(x_2), ..., F_d(x_d)).$$

Note 15.4.3 (caution!).

- It is **incorrect** that univariate margins and correlation matrix will allow the construction of the multivariate joint cdf.
- It is **incorrect** that univariate Gaussian margins and correlation matrix will allow the construction of the multivariate joint cdf.
- It is **only correct** that, **if we know the joint distribution is multivariate Gaussian**, univariate Gaussian margins and correlation matrix will allow the construction of the multivariate joint cdf.

Theorem 15.4.2 (construct a copulas for a joint distribution; any continuous multivariate cdf has a copula). *Consider a random vector $(X_1, X_2, ..., X_d)$ with continuous cdf F and marginal cdf $F_1, ..., F_d$. The copula for this random vector $(X_1, X_2, ..., X_d)$ is defined as*

$$C(u_1, u_2, ..., u_d) = F(F_1^{-1}(u_1), F_2^{-1}(u_2), ..., F_d^{-1}(u_d))$$

where $u_1, u_2, ..., u_d \in [0, 1]$.

Proof. Based on the definition of a copula associated with a cdf, we want to show that $F(x_1, x_2, ..., x_d) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d))$. We have

$$\begin{aligned} C(u_1, u_2, ..., u_d) &= F(F_1^{-1}(u_1), F_2^{-1}(u_2), ..., F_d^{-1}(u_d)) \\ \implies C(F_1(u_1), F_2(u_2), ..., F_d(u_d)) &= F(F_1^{-1}(F_1(u_1)), F_2^{-1}(F_2(u_2)), ..., F_d^{-1}(F_d(u_d))) \\ &= F(F_1^{-1}(F_1(u_1)), F_2^{-1}(F_2(u_2)), ..., F_d^{-1}(F_d(u_d))) \\ &= F(u_1, u_2, ..., u_n). \end{aligned}$$

□

Lemma 15.4.5 (copula of a uniform distribution). *For d-dimensional uniform cdf F, its copulas $C = F$.*

Proof. Note that

$$C(u_1, u_2, ..., u_d) = F(F_1^{-1}(u_1), F_2^{-1}(u_2), ..., F_d^{-1}(u_d)) = F(u_1, u_2, ..., u_d).$$

since for all marginals, $F_i(x) = x$. □

Remark 15.4.2. Given a copula, we can obtain different multivariate cdf by selecting different marginal distribution.

Theorem 15.4.3 (Sklar's theorem, construct joint cdf and pdf from copula and margins). *For every multivariate cumulative distribution function*

$$H(x_1, ..., x_d) \triangleq P(X_1 \leq x_1, ..., X_d \leq x_d).$$

of a random vector $(X_1, X_2, ..., X_d)$, there exists a copula C such that

$$H(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d)).$$

If H has a density h, then

$$h(x_1, ..., x_d) = c(F_1(x_1), ..., F_d(x_d)) \cdot f_1(x_1) \cdots f_d(x_d),$$

or equivalently,

$$c(x_1, ..., x_d) = \frac{h(F_1(x_1), ..., F_d(x_d))}{f_1(x_1) \cdots f_d(x_d)},$$

where $f_1, f_2, ..., f_d$ are marginal density functions.

Proof. (1) directly from Theorem 15.4.2.(2) Use chain rule. Note that

$$h = \frac{\partial^d H}{\partial x_1 \partial x_2 ... \partial x_d},$$

and

$$\frac{\partial H}{\partial x_1} = \frac{\partial C}{\partial x_1} = \frac{\partial C}{\partial F_1} \frac{\partial F_1}{\partial x_1}.$$

Therefore,

$$h = \frac{\partial^d H}{\partial x_1 \partial x_2 ... \partial x_d} = \frac{\partial^d C}{\partial F_1 \partial F_2 ... \partial F_d} \frac{\partial F_1}{\partial x_1} \frac{\partial F_2}{\partial x_2} ... \frac{\partial F_d}{\partial x_d} = c(F_1, F_2, ..., F_d) f_1 \cdot f_2 \cdots f_d.$$

□

Remark 15.4.3. The marginal distribution and the copulas can be modeled and estimated separately and independently.

Example 15.4.4 (joint density with Gaussian copula). Let $c(u_1, u_2, ..., u_d)$ be a Gaussian copula given by

$$c(u_1, u_2, ..., u_d) = \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} \phi^{-1}(u_1) \\ \vdots \\ \phi^{-1}(u_d) \end{bmatrix}^T \cdot (R^{-1} - I) \begin{bmatrix} \phi^{-1}(u_1) \\ \vdots \\ \phi^{-1}(u_d) \end{bmatrix}),$$

where ϕ is the cdf for a standard normal variable.

- If we have Gaussian margins all given by ϕ, then we have the joint pdf given by

$$\begin{aligned} & h(x_1, x_2, ..., x_d) \\ & = c(\phi(x_1), \phi(x_2), ..., \phi(x_d))\phi(x_1)\phi(x_2)\cdots\phi(x_d) \\ & = \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} \phi^{-1}(\phi(x_1)) \\ \vdots \\ \phi^{-1}(\phi(x_d)) \end{bmatrix}^T \cdot (R^{-1} - I) \begin{bmatrix} \phi^{-1}(\phi(x_1)) \\ \vdots \\ \phi^{-1}(\phi(x_d)) \end{bmatrix}) \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2} \sum_{i=1}^{d} x_i^2) \\ & = \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} x^T R^{-1} x) \end{aligned}$$

- Let $\phi(x)$ be the standard Gaussian cdf, then a Gaussian random variable $N(m, \sigma^2)$ has pdf given by $\phi(\frac{x-m}{\sigma})$.

Then,

$$
\begin{aligned}
& h(x_1, ..., x_d) \\
& = c(F_1(x_1), ..., F_d(x_d)) \cdot f_1(x_1) \cdots f_d(x_d) \\
& = \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} \phi^{-1}(\phi(x_1 - m_1/\sigma_1)) \\ \vdots \\ \phi^{-1}(\phi(x_d - m_d/\sigma_d)) \end{bmatrix}^T \\
& \cdot (R^{-1} - I) \begin{bmatrix} \phi^{-1}(\phi(x_1 - m_1/\sigma_1)) \\ \vdots \\ \phi^{-1}(\phi(x_d - m_d/\sigma_d)) \end{bmatrix}) \\
& \frac{1}{(2\pi)^{d/2}} \exp(\begin{bmatrix} (x_1 - m_1/\sigma_1)) \\ \vdots \\ (x_d - m_d/\sigma_d)) \end{bmatrix}^T \cdot \begin{bmatrix} (x_1 - m_1/\sigma_1)) \\ \vdots \\ (x_d - m_d/\sigma_d)) \end{bmatrix}) \\
& = \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix}^T \cdot (R^{-1}) \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix}) \frac{1}{(2\pi)^{d/2}} \\
& = \frac{1}{\sqrt{det(R)}(2\pi)^{d/2}} \exp(-\frac{1}{2} \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix}^T \cdot (\Sigma^{-1}) \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix})
\end{aligned}
$$

where $\Sigma^{-1} = D^{-1}R^{-1}D^{-1}$, $D = diag(\sigma_1, \sigma_2, ..., \sigma_d)$.

- If we have pdf margins given by $f_1, f_2, ..., f_d$ and cdf margins given by $F_1, F_2, ..., F_d$, then we have the joint pdf given by

$$\begin{aligned} h(x_1, x_2, ..., x_d) &= c(F_1(x_1), F_2(x_2), ..., F_d(x_d)) f_1(x_1) f_2(x_2) \cdots f_d(x_d) \\ &= \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} \phi^{-1}(F_1(x_1)) \\ \vdots \\ \phi^{-1}(F_d(x_d)) \end{bmatrix}^T \\ &\cdot (R^{-1} - I) \begin{bmatrix} \phi^{-1}(F_1(x_1)) \\ \vdots \\ \phi^{-1}(F_d(x_d)) \end{bmatrix}) f_1(x_1) f_2(x_2) \cdots f_d(x_d) \end{aligned}$$

Lemma 15.4.6 (copula density for random vector with independent components). *If a random vector $(X_1, X_2, ..., X_d)$ has independent components, then its copula density is given by*

$$c = 1.$$

Proof. From Theorem 15.4.3, we know that

$$c(x_1, ..., x_d) = \frac{h(F_1(x_1), ..., F_d(x_d))}{f_1(x_1) \cdots f_d(x_d)} = \frac{f_1(x_1) \cdots f_d(x_d)}{f_1(x_1) \cdots f_d(x_d)} = 1.$$

□

Lemma 15.4.7 (copulas invariance under monotone transform). *Let C be the copulas associated with random vector $(X_1, X_2, ..., X_d)$ and its cdf F. It follows that*

- *Let $T_1, T_2, ..., T_d$ be increasing functions. Then C is also the copulas associated with random vector $(T_1(X_1), T_2(X_2), ..., T_n(X_n))$.*
- *Let $T_1, T_2, ..., T_d$ be monotone(either increasing or decreasing) functions. **If** $(X_1, X_2, ..., X_n)$ **have uniform distribution**, then C is also the copulas associated with random vector $(T_1(X_1), T_2(X_2), ..., T_n(X_n))$.*

Proof. (1)Denote $(Y_1, Y_2, ..., Y_n) = (T_1(X_1), T_2(X_2), ..., T_n(X_n))$. Note that $(Y_1, Y_2, ..., Y_n)$ has marginal

$$F_{Y_i}(y_i) = F_{X_i}(T_i^{-1}(y_1)).$$

and cdf

$$F_Y(y_1, ..., y_2) = F_X(T_1^{-1}(y_1), T_2^{-1}(y_2), ..., T_n^{-1}(y_n)).$$

Then

$$\begin{aligned} C_Y(u_1, u_2, ..., u_d) &= F_Y(F_{Y_1}^{-1}(u_1), F_{Y_2}^{-1}(u_2), ..., F_{Y_n}^{-1}(u_n)) \\ &= F_X(T_1^{-1}(F_{Y_1}^{-1}(u_1)), T_2^{-1}(F_{Y_2}^{-1}(u_2)), ..., T_n^{-1}(F_{Y_n}^{-1}(u_n))) \\ &= F_X(F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2), ..., F_{X_n}^{-1}(u_n)) \\ &= C_X(u_1, u_2, ..., u_d) \end{aligned}$$

where we use the relation $F_{Y_i} = F_{X_i} \circ T^{-1}$ such that

$$F_{X_1}^{-1} = (F_{Y_1} \circ T)^{-1} = T^{-1} \circ F_{Y_1}^{-1}.$$

Note that $(F_{Y_1} \circ T)$ is invertible only if T is increasing.

(2) We can replace $F_{X_i} = I$ in the above proof and get the result.

□

Remark 15.4.4 (caution!). This does not hold for decreasing transform.

15.4.2.2 *Survival copula*

Definition 15.4.3 (survival copula).

- *Consider a uniform random vector $(U_1, U_2, ..., U_d)$ with joint cdf/copula C. The **survival copula** associated with C is defined by*

$$C^s(u_1, u_2, ..., u_d) \triangleq Pr(U_1 > u_1, U_2 > u_2, ..., U_d > u_d).$$

- *We have relation*

$$1 - F(U_1 < u_1, U_2 < u_2, ..., U_d < u_d) = C^s(1 - F_1(u_1), 1 - F_2(u_2), ..., 1 - F_d(u_d)).$$

or equivalently

$$S(u_1, u_2, ..., u_d) = C^s(S_1(u_1), S_2(u_2), ..., S_d(u_d)),$$

where $S \triangleq 1 - F, S_i \triangleq 1 - F_i, i = 1, 2, ..., d$.

Lemma 15.4.8 (conversion between copula and survival copula).

- *Let $(U_1, U_2, ..., U_n)$ be uniform random variables. Let C be its joint cdf/copula, let C^s be its survival copula. Then*

$$C(u_1, u_2, ..., u_n) = 1 - C^s(1 - u_1, 1 - u_2, ..., 1 - u_n).$$

Or equivalently,

$$C^s(u_1, u_2, ..., u_n) = 1 - C(1 - u_1, 1 - u_2, ..., 1 - u_n).$$

- *Let $(X_1, X_2, ..., X_n)$ be random variables with margins $F_1, F_2, ..., F_n$ and copula C. Then the joint cdf*

$$\begin{aligned} F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) &= C(F_1(x_1), F_2(x_2), ..., F_n(x_n)) \\ &= 1 - C^s(1 - F_1(x_1), 1 - F_2(x_x), ..., 1 - F_n(x_n)) \end{aligned}$$

Or equivalently,

$$\begin{aligned} F^s_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) &\triangleq Pr(X_1 > x_1, X_2 > x_2, ..., X_n > x_n) \\ &= C(F_1(x_1), F_2(x_2), ..., F_n(x_n)) \\ &= 1 - C^s(1 - F_1(x_1), 1 - F_2(x_x), ..., 1 - F_n(x_n)) \end{aligned}$$

$$C^s(u_1, u_2, ..., u_n) = 1 - C(1 - u_1, 1 - u_2, ..., 1 - u_n).$$

15.4.2.3 *Partial differential and conditional distribution*

Theorem 15.4.4 (partial differential of copula gives conditional distribution of uniform random variables). *[10, p. 16]*

- *(bivariate case)Let $C(u, v)$ be a copula, i.e., the joint cdf associated with uniform random variables (U, V). Assume $C(u, v)$ can be differentiated. Then*

$$\frac{\partial}{\partial v} C(u, v)$$

is the conditional cdf given by

$$\frac{\partial}{\partial u} C(v, v) = Pr(U < u | V = v).$$

- *(multivariate case, multiple conditioning on one)Let $C(u_1, u_2, ..., u_n)$ be a differentiable copula, i.e., the joint cdf associated with uniform random variables $(U_1, U_2, ..., U_n)$. Then*

$$\frac{\partial}{\partial u_k} C(u_1, u_2, ..., u_n)$$

is the conditional cdf given by

$$\frac{\partial}{\partial u_k} C(u_1, u_2, ..., u_n) = Pr(U_i \leq u_i, 1 \leq i \leq n, i \neq k | U_k = u_k).$$

- *(multivariate case, one conditioning on multiple)Let $C(u_1, u_2, ..., u_n)$ be a differentiable copula, i.e., the joint cdf associated with uniform random variables $(U_1, U_2, ..., U_n)$. Then the conditional cdf given by*

$$\frac{\frac{\partial^{n-1}}{\partial u_1 ... \partial u_{k-1} ... \partial u_n} C_{1:n}(u_1, ..., u_n)}{\frac{\partial^{k-1}}{\partial u_1 ... \partial u_{k-1}} C_{1:k-1}(u_1, ..., u_k - 1)} = Pr(U_k \leq u_k | U_i = u_i, 1 \leq i \leq n, i \neq k).$$

Proof. (1)

$$\begin{aligned} Pr(U \leq u | V = v) &= \lim_{h \to 0} Pr(U \leq u | v \leq V \leq v + h) \\ &= \lim_{h \to 0} \frac{C(u, v+h) - C(u, v)}{C_V(v+h) - C_V(v)} \\ &= \lim_{h \to 0} \frac{C(u, v+h) - C(u, v)}{h} \\ &= \lim_{h \to 0} \frac{C(u, v+h) - C(u, v)}{h} \qquad = \frac{\partial}{\partial u} C(u, v) \end{aligned}$$

where C_V is the marginal distribution of V given by [Lemma 15.4.1]

$$C_V(v) = C(1, v) = v.$$

(2)

$$\begin{aligned} &Pr(U_i \leq u_i, 1 \leq i \leq n, i \neq k | U_k) \\ &= \lim_{h \to 0} Pr(U_i \leq x_i, 1 \leq i \leq n, i \neq k | u_k \leq U_k \leq u_k + h) \\ &= \lim_{h \to 0} \frac{C(u_1, ..., u_k + h, ..., u_n) - C(u_1, ..., u_k, ..., u_n)}{C_k(x_k + h) - C_k(x_k)} \\ &= \lim_{h \to 0} \frac{C(u_1, ..., u_k + h, ..., u_n) - C(u_1, ..., u_k, ..., u_n)}{h} \\ &= \frac{\partial}{\partial u_k} C(u_1, ..., u_k, ..., u_n) \end{aligned}$$

where C_k is the marginal distribution of U_k given by Lemma 15.4.1

$$C_k(u_k) = C(1, ...1, , u_k, 1, ..., 1) = u_k.$$

(3) informally, we can think of the upper as

$$Pr(U_k \leq u_k, U_i = u_i, 1 \leq i \leq n, i \neq k);$$

and the lower as

$$Pr(U_i = u_i, 1 \leq i \leq n, i \neq k).$$

□

Lemma 15.4.9 (partial differential of copula gives conditional distribution for general random variables, bivariate distribution). *[10, p. 16] Let $C(u, v)$ be a copula, i.e., the joint cdf associated with uniform random variables (U, V). Assume $C(u, v)$ can be differentiated. Then*

$$\frac{\partial}{\partial u} C(u, v)$$

is the conditional cdf given by

$$\frac{\partial}{\partial u} C(u, v) = Pr(V < v | U = u).$$

Proof.

$$\begin{aligned}
Pr(X \leq x | Y = y) &= \lim_{h \to 0} Pr(X \leq x | y \leq Y \leq y + h) \\
&= \lim_{h \to 0} \frac{F_{XY}(x, y + h) - F_{XY}(x, y)}{F_Y(y + h) - F_Y(y)} \\
&= \lim_{h \to 0} \frac{C(F_X(x), F_Y(y + h)) - C(F_X(x), F_Y(y))}{F_Y(y + h) - F_Y(y)} \\
&= \lim_{h \to 0} \frac{C(F_X(x), F_Y(y) + \Delta(h)) - C(F_X(x), F_Y(y))}{\Delta(h)} \\
&= \frac{\partial}{\partial q} C(p, q)|_{p = F_X(x), q = F_Y(y)}
\end{aligned}$$

where $\Delta(h) = F_Y(y + h) - F_Y(y)$. □

Lemma 15.4.10 (partial differential gives conditional distribution, multiple conditioned on one). *[10, p. 19] Let $X_1, X_2, ..., X_n$ be real-valued random variables with corresponding copula C and continuous marginals $F_1, ..., F_n$. Then for any $k \in \{1, 2, ..., n\}$,*

$$Pr(X_i \leq x_i, 1 \leq i \leq n, i \neq k | X_k) = \frac{\partial}{\partial F_k(X_k)} C(F_1(x_1), ..., F_k(x_k), ..., F_n(x_n)),$$

for all $x_1, x_2, ..., x_n \in \mathbb{R}$.

Proof.

$$\begin{aligned}
&Pr(X_i \leq x_i, 1 \leq i \leq n, i \neq k | X_k) \\
&= \lim_{h \to 0} Pr(X_i \leq x_i, 1 \leq i \leq n, i \neq k | x_k \leq X_k \leq x_k + h) \\
&= \lim_{h \to 0} \frac{F_{X_1:X_n}(x_1, ..., x_k + h, ..., x_n) - F_{X_1:X_n}(x_1, ..., x_k, ..., x_n)}{F_k(x_k + h) - F_k(x_k)} \\
&= \lim_{h \to 0} \frac{C(F_1(x_1), ..., F_k(x_k + h), ..., F_n(x_n)) - C(F_1(x_1), ..., F_k(x_k), ..., F_n(x_n))}{F_k(x_k + h) - F_k(x_k)} \\
&= \lim_{h \to 0} \frac{C(F_1(x_1), ..., F_k(x_k) + \Delta(h), ..., F_n(x_n)) - C(F_1(x_1), ..., F_k(x_k), ..., F_n(x_n))}{\Delta(h)} \\
&= \frac{\partial}{\partial F_k(X_k)} C(F_1(x_1), ..., F_k(x_k), ..., F_n(x_n))
\end{aligned}$$

where $\Delta(h) = F_k(x_k + h) - F_k(x_k)$. □

Lemma 15.4.11 (partial differential gives conditional distribution, one conditioned on multiple). *[10, p. 20] Let $X_1, X_2, ..., X_n$ be real-valued random variables with corresponding copula C and continuous marginals $F_1, ..., F_n$. Then for any $k \in \{1, 2, ..., n\}$,*

$$Pr(X_i \leq x_i, 1 \leq i \leq n, i \neq k | X_k) = \frac{\partial}{\partial F_k(X_k)} C(F_1(x_1), ..., F_k(x_k), ..., F_n(x_n)),$$

for all $x_1, x_2, ..., x_n \in \mathbb{R}$.

Proof.

$$\begin{aligned}
Pr(X \leq x|Y = y) &= \lim_{h\to 0} Pr(X \leq x|y \leq Y \leq y+h) \\
&= \lim_{h\to 0} \frac{F_{XY}(x, y+h) - F_{XY}(x, y)}{F_Y(y+h) - F_Y(y)} \\
&= \lim_{h\to 0} \frac{C(F_X(x), F_Y(y+h)) - C(F_X(x), F_Y(y))}{F_Y(y+h) - F_Y(y)} \\
&= \lim_{h\to 0} \frac{C(F_X(x), F_Y(y) + \Delta(h)) - C(F_X(x), F_Y(y))}{\Delta(h)} \\
&= \frac{\partial}{\partial q} C(p, q)|_{p=F_X(x), q=F_Y(y)}
\end{aligned}$$

where $\Delta(h) = F_Y(y+h) - F_Y(y)$. □

15.4.3 Common copula functions

15.4.3.1 *Gaussian copula*

Definition 15.4.4 (Gaussian copula). *A Gaussian copula characterized by correlation matrix $R \in [-1, 1]^{d\times d}$ is given by*

$$C(u_1, u_2, ..., u_d; R) = \Phi(\phi^{-1}(u_1), \phi^{-1}(u_2), ..., \phi^{-1}(u_d)),$$

where $u_1, u_2, ..., u_d \in [0, 1]$, Φ is the cdf for a multivariate normal distribution with zero mean and covariance matrix R, ϕ is the cdf for a standard normal variable.

The copula density function is given by

$$c(u_1, u_2, ..., u_d) = \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} \phi^{-1}(u_1) \\ \vdots \\ \phi^{-1}(u_d) \end{bmatrix}^T \cdot (R^{-1} - I) \begin{bmatrix} \phi^{-1}(u_1) \\ \vdots \\ \phi^{-1}(u_d) \end{bmatrix}).$$

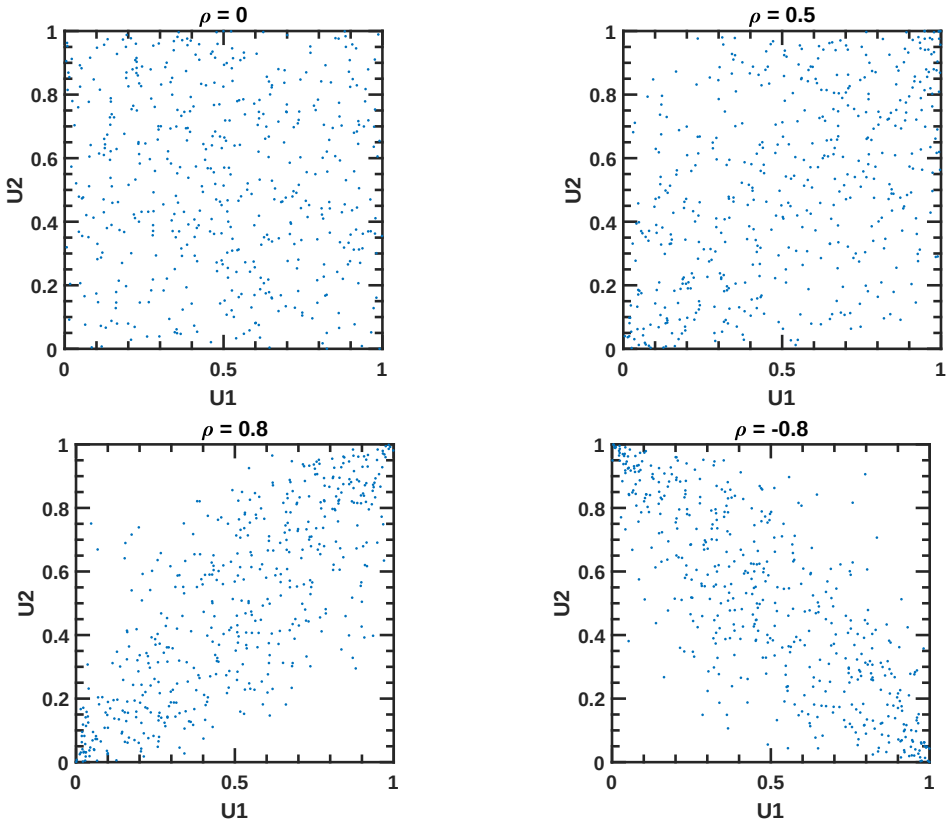


Figure 15.4.1: Gaussian copula with different correlations.

Remark 15.4.5 (derivation of copula density function). From Theorem 15.4.3, we know that

$$c(x_1, ..., x_d) = \frac{h(F_1(x_1), ..., F_d(x_d))}{f_1(x_1) \cdots f_d(x_d)}.$$

Note that

$$h(x) = \frac{1}{(2\pi)^{d/2}|detR|^{1/2}} \exp(-\frac{1}{2}(x)^T R^{-1}(x)), x \in \mathbb{R}^d,$$

and

$$f_1 \cdot f_2 \cdots f_d = \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2}(x)^T(x)).$$

Lemma 15.4.12 (the copula associated with multivariate Gaussian distribution is Gaussian copula).

- *If F is a multivariate normal distribution $MN(0, R)$, where R is correlation matrix and $\Sigma = R$(that is, the margins are standard normal such that covariance matrix is*

equal to correlation matrix); then the copula associated with F is the Gaussian copula with correlation matrix R

- *If F is a multivariate normal distribution $MN(0, \Sigma)$, then the copula associated with F is the Gaussian copula with correlation matrix $R = D^{-1/2}\Sigma D^{-1/2}, D = diag(\Sigma)$.*
- *If F is a multivariate normal distribution $MN(\mu, \Sigma)$, then the copula associated with F is the Gaussian copula with correlation matrix $R = D^{-1/2}\Sigma D^{-1/2}, D = diag(\Sigma)$.*

Proof. (1)Straight forward [Theorem 15.4.2]. (2)(3) Let Φ denote the joint cdf $MN(\mu, \Sigma)$ and ϕ_i the marginal cdf $N(\mu_i, \sigma_i^2)$. Let Φ^S denote the joint cdf $MN(\mu, R)$ and ϕ_i^S the marginal cdf $N(0, 1)$.

Then,

$$\begin{aligned}\Phi(u_1, u_2, ..., u_d) &= \Phi^S(\frac{u_1 - \mu_1}{\sigma_1^2}, \frac{u_2 - \mu_2}{\sigma_2^2}, ..., \frac{u_d - \mu_d}{\sigma_d})\\ \phi_i(u_i) &= \phi_i^S(\frac{u_i - \mu_i}{\sigma_i})\\ \phi_i^{-1}(u_1) &= \sigma_i(\phi_i^S)^{-1} + \mu_i\end{aligned}$$

Therefore,

$$\begin{aligned}&\Phi(\phi_i^{-1}(u_1), \phi_2^{-1}(u_2), ..., \phi_2^{-1}(u_d))\\ =&\Phi^S((\phi_i^{-1}(u_1) - \mu_1)/\sigma_1, (\phi_i^{-1}(u_2) - \mu_2)/\sigma_2, ..., (\phi_i^{-1}(u_d) - \mu_d)/\sigma_d)\\ =&\Phi^S((\phi_1^S)^{-1}, (\phi_2^S)^{-1}, ..., (\phi_d^S)^{-1})\end{aligned}$$

□

Lemma 15.4.13 (construct a multivariate cdf from Gaussian copulas and margins). *Consider d random variables $X_1, X_2, ..., X_d$ with $F_1, F_2, ..., F_d$ being the univariate cdf. Let C be a d dimensional Gaussian copulas given by*

$$C(u_1, u_2, ..., u_d; R) = \Phi(\phi^{-1}(u_1), \phi^{-1}(u_2), ..., \phi^{-1}(u_d)),$$

where $u_1, u_2, ..., u_d \in [0, 1]$, Φ is the cdf for a multivariate normal distribution with zero mean and covariance matrix R, ϕ is the cdf for a standard normal variable.

Then the joint distribution for $X_1, X_2, ..., X_d$ is given by

$$F(x_1, x_2, ..., x_n) = \Phi(\phi^{-1}(F_1(x_1)), \phi^{-1}(F_2(x_2)), ..., \phi^{-1}(F_d(x_d)));$$

and the joint density function is given by

$$f(x) = \frac{1}{(2\pi)^{d/2}|detR|^{1/2}} \exp(-\frac{1}{2}(x)^T R^{-1}(x)) f_1(x_1) f_2(x) \cdots f_d(x_d),$$

where

$$x = (\phi^{-1}(F_1(x_1)), \phi^{-1}(F_2(x_2)), ..., \phi^{-1}(F_d(x_d))).$$

and $f_1, f_2, ..., f_d$ *are the marginal densities of* $X_1, X_2, ..., X_n$.

Proof. (1)See Theorem 15.4.1. (2)See Theorem 15.4.3. □

Example 15.4.5 (Gaussian margin with Gaussian copula will give multivariate Gaussian). Note that from Theorem 15.4.3, we know that the joint cdf can be constructed from margins $f_1, f_2, ..., f_d$ and $c(u_1, u_2, ..., u_d)$ via

$$h(x_1, ..., x_d) = c(F_1(x_1), ..., F_d(x_d)) \cdot f_1(x_1) \cdots f_d(x_d).$$

Let $\phi(x)$ be the standard Gaussian cdf, then a Gaussian random variable $N(m, \sigma^2)$ has pdf given by $\phi(\frac{x-m}{\sigma})$.

Then,

$$
\begin{aligned}
&h(x_1, ..., x_d) \\
&= c(F_1(x_1), ..., F_d(x_d)) \cdot f_1(x_1) \cdots f_d(x_d) \\
&= \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} \phi^{-1}(\phi(x_1 - m_1/\sigma_1)) \\ \vdots \\ \phi^{-1}(\phi(x_d - m_d/\sigma_d)) \end{bmatrix}^T \\
&\cdot (R^{-1} - I) \begin{bmatrix} \phi^{-1}(\phi(x_1 - m_1/\sigma_1)) \\ \vdots \\ \phi^{-1}(\phi(x_d - m_d/\sigma_d)) \end{bmatrix}) \\
&\frac{1}{(2\pi)^{d/2}} \exp(\begin{bmatrix} (x_1 - m_1/\sigma_1)) \\ \vdots \\ (x_d - m_d/\sigma_d)) \end{bmatrix}^T \cdot \begin{bmatrix} (x_1 - m_1/\sigma_1)) \\ \vdots \\ (x_d - m_d/\sigma_d)) \end{bmatrix})) \\
&= \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix}^T \cdot (R^{-1}) \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix}) \frac{1}{(2\pi)^{d/2}} \\
&= \frac{1}{\sqrt{det(R)}(2\pi)^{d/2}} \exp(-\frac{1}{2} \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix}^T \cdot (\Sigma^{-1}) \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix})
\end{aligned}
$$

where $\Sigma^{-1} = D^{-1}R^{-1}D^{-1}$, $D = diag(\sigma_1, \sigma_2, ..., \sigma_d)$.

Example 15.4.6 (bivariate Gaussian copula).

$$C_\rho(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}) dxdy$$

where ρ is the linear correlation coefficient.

15.4.3.2 *t copula*

Definition 15.4.5 (*t* copula). *[11, p. 419] A t-copula characterized by correlation matrix $R \in [-1,1]^{n\times n}$ and degree of freedom parameter v is the copula associated with the multivariate Student's t probability distribution*

$$C(u_1,u_2,...,u_n;\rho,v) = T_n(T_v^{-1}(u_1),T_v^{-1}(u_2),...,T_v^{-1}(u_n);R,v).$$

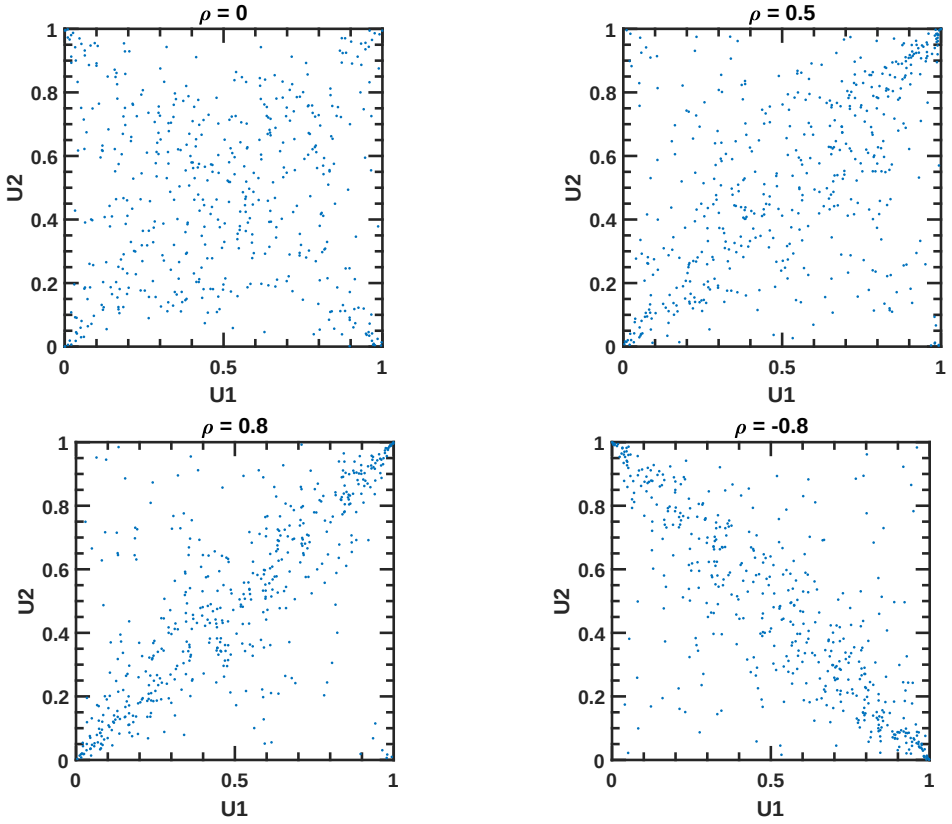


Figure 15.4.2: Student T copula with different correlations.

Example 15.4.7 (bivariate t copula).

$$C(u_1,u_2;\rho,v) = \int_{-\infty}^{T_v^{-1}(u_1)}\int_{-\infty}^{T_v^{-1}(u_2)} \frac{1}{2\pi(1-\rho^2)^{1/2}}(1+\frac{x_1^2+x_2^2-2\rho x_1x_2}{v(1-\rho^2)})dx_1dx_2$$

where ρ is the linear correlation coefficient.

15.4.3.3 *Common copula functions: other copula*

Definition 15.4.6 (product copula, co-monotonicity copula). *[12, p. 187][3, p. 190]*

- ***(product copula, independence copula)****The d-dimensional product copula is given by*
$$C(u_1, u_2, ..., u_d) = u_1 u_2 \cdots u_d, u_i \in [0,1], \forall i.$$
The product copula corresponds to independence and it can be viewed as the cdf of $(U_1, ..., U_d)$, where $U_1, ..., U_d$ are independent uniform random variables.
- ***(co-monotonicity copula)*** *The joint cdf of the random vector $(U_1, U_2, ..., U_d)$, where U is a uniform random variable is called a co-monotonicity copula, which characterizes perfect positive correlation. It is given by*
$$C(u_1, u_2, ..., u_d) = \min(u_1, u_2, ..., u_d).$$
- ***two dimensional counter-monotonicity copula*** *is defined as the joint cdf of $(U, 1-U)$.Therefore,*
$$\begin{aligned} C(u_1, u_2) \triangleq & Pr(U \leq u_1, (1-U) \leq u_2) \\ = & Pr(1-u_2 \leq U \leq u_1) \\ = & \max\{u_1 + u_2 - 1, 0\} \end{aligned}$$

Remark 15.4.6 (interpretation). [12, p. 185]

- The co-monotonicity copula is the joint cdf of $\boldsymbol{U} = (U, U, ..., U)$;that is, $\boldsymbol{U}$ contains d copies of $U(0,1)$. The co-monotonicity copula is the upper bound of all copula functions [Lemma 15.4.4].
- The two dimensional counter-monotonicity copula is the lower bound of all copula functions.

15.4.4 Dependence and copula

15.4.4.1 *Linear correlations*

Example 15.4.8. Consider discrete-valued random variable V_1 and V_2 given by:

- V_1 equally take three different values $-1, 0, +1$.
- If $V_1 = -1$ *or* $+1$, $V_2 = 1$. If $V_1 = 0$, then $V_2 = 0$.

It is clearly that $E[V_1V_2] = 0, E[V_1] = 0 \implies Cov[V_1, V_2] = 0$; however, it is clearly that V_1 and V_2 are uncorrelated but they are dependent since

$$Pr(V_2|V_1 = v) \neq Pr(V_2).$$

where

$$\begin{aligned} Pr(V_2 = 1) &= Pr(V_2 = 1|V_1 = 1)Pr(V_1 = 1) + Pr(V_2 = 1|V_1 = 1)Pr(V_1 = 1) \\ &= 1 \times 1/3 + 1 \times 1/3 = 2/3; \end{aligned}$$

and

$$Pr(V_2 = 0) = Pr(V_2 = 0|V_1 = 0)Pr(V_1 = 0) = 1 \times 1/3 = 1/3.$$

Definition 15.4.7 (linear correlation, Pearson correlation). *[3, p. 202] Given two random variables X and Y. The linear correlation is defined by*

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var[X]Var[Y]}}.$$

Note 15.4.4 (characteristics of linear correlations).

- sensitive to outliers.
- measure the 'average dependence' between X and Y.
- invariant under strictly **increasing linear transformation**.
- May be misleading when multivariate distribution is not elliptical.

Lemma 15.4.14 (invariance of correlation under affine transformation).

- *Let random variables X and Y have correlation $\rho(X, Y)$. Then*

$$\rho(aX + b, cY + d) = \rho(X, Y), \forall a, c > 0, b, d \in \mathbb{R}.$$

- *Let random variables X and Y have correlation $\rho(X, Y)$. Then*

$$\rho(aX + b, cY + d) = \frac{cd}{|cd|}\rho(X, Y), \forall a, b, b, d \in \mathbb{R}.$$

Proof. (1)

$$\rho(aX + b, cY + d) \triangleq \frac{Cov(aX + b, cY + d)}{\sqrt{Var[aX + b]Var[cY + d]}} = \frac{aCov(X, Y)d}{\sqrt{a^2Var[X]Var[Y]c^2}} == \rho(X, Y).$$

(2) straight forward. □

Remark 15.4.7 (generally not invariant under nonlinear transformation). Let $T : \mathbb{R} \to \mathbb{R}$ be a nonlinear strictly increasing function. Then generally

$$\rho(X, Y) \neq \rho(T(X), T(Y)).$$

Lemma 15.4.15 (perfect linear correlation and linear function relationship). *[3, p. 202] Let X and Y be two random variable defined on the same probability space. It follows that*

- *$\rho(X, Y) = 1$ implies $Y = \alpha + \beta X$ almost surely for some $\alpha, \beta \in \mathbb{R}, \beta > 0$.*
- *$\rho(X, Y) = -1$ implies $Y = \alpha + \beta X$ almost surely for some $\alpha, \beta \in \mathbb{R}, \beta < 0$.*
- *Conversely, if $Y = \alpha + \beta X$, then $\rho(X, Y) = sign(\beta)$.*

Proof. (1)(2) Note that the equality holds only when the equality holds in Cauchy inequality [Theorem 12.10.4].Then, X and Y must have linear dependence almost everywhere. (3) Use the affine transformation invariance property of correlation [Lemma 15.4.14], □

15.4.4.2 *Rank correlations*

Definition 15.4.8 (Kendall's tau for observations).

- *Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be a set of joint observations from two random variables X and Y.*
- *A pair of observations (x_i, y_i) and (x_j, y_j) are **concordant** if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$;*
- *They are **discordant**, if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$.*
- *If $x_i = x_j, y_i = y_j$, the pair is neither concordant nor discordant.*
- *The **Kendall τ coefficient** is defined as*

$$\rho_\tau = \frac{num\ concordant\ pairs - num\ disconcordant\ pairs}{n(n-1)/2}.$$

Note that $n(n-1)/2$ is the total number of pairs to compare.

Definition 15.4.9 (Kendall's tau for random variables). *[3, p. 207] Let X_1 and X_2 be two random variables. Then the Kendall's tau is given by*

$$\rho_\tau = E[sign((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2))],$$

where $(\tilde{X}_1, \tilde{X}_2)$ is a independent copy of (X_1, X_2); that is,$(\tilde{X}_1, \tilde{X}_2)$ has the same cdf of (X_1, X_2), but they are statistically independent.

Lemma 15.4.16 (Kendall's tau from copula). *[3, p. 207] The C be the copula associated with the joint cdf of (X, Y), then*

$$\begin{aligned}\rho_\tau(X, Y) &= 4\int_0^1\int_0^1 C(u, v)dC(u, v) - 1\\ &= 4E[C(U, V)] - 1\end{aligned}$$

Proof. From the definition

$$\begin{aligned}\rho_\tau &= E[sign((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2))]\\ &= E[\mathbf{1}((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0)] - E[\mathbf{1}((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) < 0)]\\ &= Pr((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0) - Pr((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) < 0)\\ &= 2Pr((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0) - 1\\ &= 2Pr((X_1 - \tilde{X}_1) > 0, (X_2 - \tilde{X}_2) > 0) + 2Pr((X_1 - \tilde{X}_1) < 0, (X_2 - \tilde{X}_2) < 0) - 1\\ &= 4Pr((X_1 - \tilde{X}_1) < 0, (X_2 - \tilde{X}_2) < 0) - 1\\ &= 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Pr((X_1 < s_1, X_2 < s_2)f_{\tilde{X}_1, \tilde{X}_2}(s_1, s_2)ds_1ds_2 - 1\\ &= 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(s_1, s_2)dF(s_1, s_2) - 1\\ &= 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C(F_1(s_1), F_2(s_2))dC(F_1(s_1), F_2(s_2)) - 1\\ &= 4\int_0^1\int_0^1 C(u_1, u_2)dC(u_1, u_2) - 1\end{aligned}$$

□

Remark 15.4.8 (Kendall's tau is independent of the marginal cdf). Note that Kendall's tau only depends on the correlation structure characterized by the copula; it is independent of the marginal cdf.

Lemma 15.4.17 (Hoffding formula for covariance). *[3, p. 204] If (X_1, X_2) has joint cdf F and marginal cdf F_1 and F_2, then*

- $$Cov(X_1, X_2) = \frac{1}{2}E[(X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2)],$$

where $(\tilde{X}_1, \tilde{X}_2)$ is a independent copy of (X_1, X_2); that is,$(\tilde{X}_1, \tilde{X}_2)$ has the same cdf of (X_1, X_2), but they are statistically independent.

- $$Cov(X_1, X_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(F(x_1, x_2) - F_1(x_1)F_2(x_2))dx_1dx_2.$$

Proof. (1) Directly expand the rhs. Note that $E[X_1\tilde{X}_2] = E[X_1]E[\tilde{X}_2]$ due to independence. (2) A useful identity is for any $a \in \mathbb{R}, b \in R$, we have

$$(a - b) = \int_{-\infty}^{\infty} H(x - b) - H(x - a)dx.$$

We have

$$\begin{aligned}
&E[(X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2)] \\
=&E[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(H(s_1 - X_1) - H(s_1 - \tilde{X}_1))(H(s_2 - X_2) - H(s_2 - \tilde{X}_2))ds_1ds_2] \\
=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}E[(H(s_1 - X_1) - H(s_1 - \tilde{X}_1))(H(s_2 - X_2) - H(s_2 - \tilde{X}_2))]ds_1ds_2 \\
=&2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}Pr(X_1 \le s_1, X_2 \le s_2) - Pr(X_1 \le s_1)Pr(X_2 \le s_2)ds_1ds_2 \\
=&2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(F(s_1, s_2) - F_1(s_1)F_2(s_2))ds_1ds_2
\end{aligned}$$

□

Definition 15.4.10 (Spearman's rho). *[3, p. 207] Let X_1 and X_2 be two random variables with marginal cdf F_1 and F_2. **Spearman's rho** is defined as*

$$\rho_S(X, Y) = \rho(F_1(X_1), F_2(X_2)).$$

In other words, Spearman's rho is the linear correlation of the transform random variables.

Definition 15.4.11 (Spearman's rho for observations).

- *Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be a set of joint observations from two random variables X and Y.*
- *Let $R(x_i)$ denote the rank of x_i among $x_1, x_2, ..., x_n$, where $R(x_i)$ will take value from 1 to n. Similarly we denote $R(y_i)$ as the rank of y_i.*

- *The **Spearman's ρ coefficient** is defined as*

$$\rho_S = \frac{\frac{1}{n}\sum_{i=1}^{n}(R(x_i)-\bar{R}(x))(R(y_i)-\bar{R}(y))}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(R(x_i)-\bar{R}(x))^2\frac{1}{n}\sum_{i=1}^{n}(R(y_i)-\bar{R}(y))^2}},$$

  where $\bar{R}(y) = \frac{1}{n}\sum_{i=1}^{n} R(y_i)$.
- *It can be showed that*

$$\rho_S = 1 - \frac{\sum_{i=1}^{n}(R(x_i)-R(y_i))^2}{n^3-n}.$$

Lemma 15.4.18 (properties of Spearman's rho for observations). *Let $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$ be a set of joint observations from two random variables X and Y. It follows that*

- $$n\bar{R}(y) = \sum_{i=1}^{n} R(y_i) = \frac{1}{2}n(n+1).$$
- $$\sum_{i=1}^{n} R(y_i)^2 = \sum_{i=1}^{n} R(x_i)^2 = \frac{n(n+1)(2n+1)}{6}.$$
- $$\rho_S = 1 - \frac{\sum_{i=1}^{n}(R(x_i)-R(y_i))^2}{n^3-n}.$$

Proof. (1)straight forward.(2) this is the sum of squares from 1 to n. (3) Note that

$$\sum_{i=1}^{n}(R(x_i)-\bar{R}(x))^2 = \sum_{i=1}^{n} R(x_i)^2 - (\sum_{i=1}^{n} R(x_i))^2 = \frac{n(n+1)(2n+1)}{6} - (\frac{1}{2}n(n+1))^2.$$

and

$$\sum(R(x_i)-R(y_i))^2 = \sum R(x_i)^2 + \sum R(y_i)^2 - 2\sum R(x_i)R(y_i)$$

implies

$$2\sum R(x_i)R(y_i) = \sum R(x_i)^2 + \sum R(y_i)^2 - \sum(R(x_i)-R(y_i))^2.$$

□

Lemma 15.4.19 (Spearman's rho from copula). *[3, p. 207]*

$$\rho_S(X,Y) = 12\int_0^1\int_0^1 C(u,v)dudv - 3$$

Proof. From definition and use Lemma 15.4.17, we have

$$\begin{aligned}
\rho_S(X,Y) &= \rho(F_1(X_1), F_2(X_2)) \\
&= Cov(F_1(X_1), F_2(X_2))/(\sqrt{Var[F_1[X_1]]Var[F_2[X_2]]}) \\
&= 12Cov(F_1(X_1), F_2(X_2)) \\
&= 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{F}(F_1(x_1), F_2(x_2)) - \hat{F}_1(F_1(x_1))\hat{F}_2(F_2(x_2))dx_1dx_2 \\
&= 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{F}(F_1(x_1), F_2(x_2)) - x_1x_2dx_1dx_2 \\
&= 12(\int_0^1\int_0^1 C(u,v)dudv - \frac{1}{4})
\end{aligned}$$

where we use the property $F_1(X_1), F_2(X_2)$ are uniform random variable with variance $1/12$, and the joint cdf of $(F_1(X_1), F_2(X_2))$ is the copula associated with joint cdf of (X_1, X_2)(Lemma 15.4.3). □

Lemma 15.4.20 (Spearman's rho for monotonic relation). *Let X be a random variable and let Y be a monotonic function of X, denoted by $Y = f(X)$. It follows that*

- *If f is a monotonically increasing function, the $\rho_S(X,Y) = 1$.*
- *If f is a monotonically decreasing function, the $\rho_S(X,Y) = -1$.*

Proof. (1)Consider the Spearman's rho for observations [Definition 15.4.11]. For each sample (x_i, y_i), each component has the same rank. (2)For each sample (x_i, y_i), each component has ranks satisfying

$$R(x_i) - \bar{R}(x_i) = -(R(y_i) - \bar{R}(y_i)).$$

□

Lemma 15.4.21 (first quadrant probability of bivariate Gaussian distribution). *[3, p. 215] Let $(X_1, X-2)$ be a random vector with joint multivariate Gaussian distribution $MN(0, \Sigma)$. Let $\rho = \rho(X_1, X_2)$. Then*

$$Pr(X_1 > 0, X_2 > 0) = \frac{1}{4} + \frac{\arcsin \rho}{2\pi}.$$

Proof. See reference. □

Lemma 15.4.22 (rank correlation for Gaussian copula). *[3, p. 215] Let (X_1, X_2) be a bivariate random vector with Gaussian copula characterized by correlation coefficient ρ and continuous margins. Then*

- $$\rho_\tau(X_1, X_2) = \frac{2}{\pi} \arcsin \rho$$
- $$\rho_S(X_1, X_2) = \frac{6}{\pi} \arcsin \frac{1}{2}\rho$$

Proof. (1) Note that Kendall's tau only depends on the copula; therefore we can assume (X_1, X_2) has bivariate normal distribution($MN(0, 2\Sigma)$, correlation ρ). From Lemma 15.4.16,

$$\begin{aligned}
\rho_\tau &= 4Pr((X_1 - \tilde{X}_1) > 0, (X_2 - \tilde{X}_2) > 0) - 1 \\
&= 4Pr(Y_1 > 0, Y_2 > 0) - 1 \\
&= 4(\frac{1}{4} + \frac{\arcsin \rho}{2\pi}) - 1 \\
&= \frac{2}{\pi} \arcsin \rho
\end{aligned}$$

where $(\tilde{X}_1, \tilde{X}_2)$ is the independent copy of (X_1, X_2), and $Y_1 = X_1 - \tilde{X}_2 \sim MN(0, 2\Sigma)(Y_1$ has the same correlation ρ), we use Lemma 15.4.21. (2) From Lemma 15.4.19, we have

$$\begin{aligned}
\rho_S(X_1, X_2) &= 12 \int_0^1 \int_0^1 C(u, v) du dv - 3 \\
&= 12 \int_0^1 \int_0^1 \Phi(\phi^{-1}(u), \phi^{-1}(v)) du dv - 3 \\
&= 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi(s_1, s_2) d\phi(s_1) d\phi(s_2) - 3 \\
&= 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi(s_1, s_2) f(s_1) f(s_2) ds_1 ds_2 - 3 \\
&= 12 Pr(X_1 - S_1 < 0, X_2 - S_2 < 0) - 3 \\
&= 12 Pr(Y_1 < 0, Y_2 < 0) - 3 \\
&= 12(\frac{1}{4} + \frac{\arcsin \rho/2}{2\pi}) - 3 \\
&= \frac{6}{\pi} \arcsin \frac{1}{2}\rho
\end{aligned}$$

where $(\tilde{X}_1, \tilde{X}_2)$ is the independent copy of (X_1, X_2), and $Y_1 = X_1 - S_1 \sim MN(0, \Sigma + I_2)(Y_1$ has the correlation $\rho/2$), we use Lemma 15.4.21. □

Remark 15.4.9 (applications in robust correlation estimation for multivariate Gaussian random variables). Note that multivariate Gaussian distribution has Gaussian copula [Lemma 15.4.12]. Therefore, we can estimate ρ_τ, ρ_S first(which is robust) and then convert them to linear correlation coefficients.

15.4.4.3 *Tail dependence*

Definition 15.4.12 (tail dependence). *Let X and Y be random variables with marginal cdf F_X and F_Y.*

- *The **coefficient of upper tail dependence** of X and Y is*

$$\begin{aligned}
\lambda_u(X, Y) \triangleq &= \lim_{\alpha \to 1} Pr(F_Y(Y) > \alpha | F_X(X) > \alpha) \\
&= \lim_{\alpha \to 1} Pr(Y > F_Y^{-1}(\alpha) | X > F_X^{-1}(\alpha)).
\end{aligned}$$

- *The **coefficient of lower tail dependence** of X and Y is*

$$\begin{aligned}\lambda_l(X,Y) \triangleq &= \lim_{\alpha\to 0} Pr(F_Y(Y) \le \alpha | F_X(X) \le \alpha) \\ &= \lim_{\alpha\to 0} Pr(Y > F_Y^{-1}(\alpha) | X > F_X^{-1}(\alpha)).\end{aligned}$$

Tail dependence is the probability of observing a large(small) Y given that X is large(small). If $\lambda_u > 0 (\lambda_l > 0)$, then we say (X,Y) has an upper(lower) tail dependence.

Lemma 15.4.23 (tail dependence from copula).

- $$\lambda_l = \lim_{q\to 0^+} \frac{Pr(F_Y(Y) \le q, F_X(X) \le q)}{Pr(F_X(X) \le q)} = \lim_{q\to 0^+} \frac{C(q,q)}{q}.$$
- $$\lambda_u = \lim_{q\to 1^-} \frac{Pr(F_Y(Y) > q, F_X(X) > q)}{Pr(F_X(X) > q)} = \lim_{q\to 1^-} \frac{C^s(q,q)}{1-q},$$

  where C^s is the survival copula.

Remark 15.4.10 (tail dependence is independent of the marginal cdf).

- Note that tail dependence only depends on the correlation structure characterized by the copula; it is independent of the marginal cdf.
- The existence of tail will depend on margins.

Lemma 15.4.24 (tail independence of Gaussian copula). *Let (X,Y) be a bivariate random vector with Gaussian copula characterized by correlation coefficient ρ and continuous margins. Then*

$$\lambda_u = \lambda_l = 2 \lim_{x\to\infty} \Phi(x\sqrt{1-\rho}/\sqrt{1+\rho}) = 0.$$

Proof. From definition, we have

$$\begin{aligned}
\lambda_l &= \lim_{q\to 0^+} \frac{Pr(F_Y(Y)\leq q, F_X(X)\leq q)}{Pr(F_X(X)\leq q)} \\
&= \lim_{q\to 0^+} \frac{Pr(Y\leq f^{-1}(q), X\leq \phi^{-1}(q))}{Pr(X\leq f^{-1}(q))} \\
&= \lim_{q\to -\infty} \frac{Pr(Y\leq q, X\leq q)}{Pr(X\leq q)} \\
&= \lim_{q\to -\infty} \frac{\int_{-\infty}^{q}\int_{-\infty}^{q} f(x,y)dxdy}{\int_{-\infty}^{q} f(x)dx} \\
&= \lim_{q\to -\infty} \frac{\int_{-\infty}^{q} f(x,q)dx}{f(q)} + \frac{\int_{-\infty}^{q} f(q,y)dy}{f(q)} \\
&= 2\lim_{q\to -\infty} \frac{\int_{-\infty}^{q} f(x,q)dx}{f(q)} \\
&= 2\lim_{q\to -\infty} f(x|y=q)dx
\end{aligned}$$

where $f(x,y)$ is the density of (X,Y), $f(x)$ is the marginal density, and we use L'hospital rule in the derivation. Note that $X|y=q \sim N(\rho q, 1-\rho^2)$ [Theorem 15.1.2]; therefore,

$$\lim_{q\to -\infty} f(x|y=q)dx = \Phi(\frac{q-\rho q}{\sqrt{1-\rho^2}}).$$

□

15.4.5 Estimating copula function

15.4.5.1 *Empirical copula method*

Definition 15.4.13 (empirical copula). *[11, p. 424] Suppose we have observation of n iid d dimensional random vectors,*

$$X^i = (X_1^{(i)}, X_2^{(i)}, ..., X_d^{(i)}), i = 1, 2, ..., n,$$

We can construct its empirical copula associated with the joint distribution of X via the following procedures:

- *(construct empirical marginal cdf)*

$$\hat{F}_k(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_k^{(i)} \leq x), k = 1, 2, ..., d$$

- *(construct transformed uniform sample)*

$$(\hat{U}_1^{(i)}, ..., \hat{U}_d^{(i)}) = (\hat{F}_1(X_1^i), ..., \hat{F}_d(X_d^{(i)})), i = 1, ..., n$$

- *(construct empirical copula)*

$$\hat{C}(u_1, u_2, ..., u_d) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\hat{U}_1^{(i)} \leq u_1, ..., \hat{U}_1^{(i)} \leq u_d)$$

Remark 15.4.11. The nature of copula is the cdf of the transformed uniform random vector $(F_1(X_1), F_2(X_2), ..., F_n(X_n))$ [Lemma 15.4.3].

15.4.5.2 *Maximum likelihood method*

Lemma 15.4.25 (maximum likelihood function and two-stage estimation method). *[11, p. 429] Suppose we have observation of n iid d dimensional random vectors,*

$$X^i = (X_1^{(i)}, X_2^{(i)}, ..., X_d^{(i)}), i = 1, 2, ..., n.$$

Assume the joint cdf for X is given by

$$F(x_1, x_2, ..., x_n) = C(F_1(x_1; \theta_1), ..., F_d(x_d; \theta_d)),$$

such that we have two sets of parameters given by

- *$(\theta_1, ..., \theta_d)$ for univariate distribution function $F_1, F_2, ..., F_d$;*
- *θ_c for the copula function $C(u_1, ..., u_d)$.*

The maximum log likelihood function is given by

$$\begin{aligned} & l(\theta_1, ..., \theta_d, \theta_c) \\ &= \sum_{i=1}^{n} \ln c(F_1(x_1^{(j)}; \hat{\theta}_1), ..., F_1(x_d^{(j)}; \hat{\theta}_d) + \sum_{i=1}^{n} \sum_{j=1}^{d} \end{aligned}$$

The first stage is to estimate univariate parameter $\theta_1, ..., \theta_d$ via

$$\hat{\theta}_i = \arg\max \sum_{j=1}^{N} \ln f_i(x_i^{(j)}; \theta_i).$$

The second stage is to estimate the copula parameters θ_c with the estimated univariate parameters $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_d$ fixed via

$$\hat{\theta}_c = \arg\max \sum_{j=1}^{N} \ln c(F_1(x_1^{(j)}; \hat{\theta}_1), ..., F_1(x_d^{(j)}; \hat{\theta}_d); \theta_c).$$

15.4.6 Applications of copula

15.4.6.1 *Generating correlated uniform random number*

Methodology 15.4.1 (conditional method for generating bivariate uniform random number with arbitrary joint distribution). *[10, p. 96] Suppose that we want to generate a pair of random variables with marginal uniform $U(0,1)$ and joint cdf given by copula $C(u,v)$.*

We use the following procedures:

- *Generate U and T independently from $U(0,1)$;*
- *Set $V = C_u^{-1}(T)$, where $C_u = \frac{\partial}{\partial u} C(u,v)$.*
- *The desired pair is (U,V).*

Proof. Note that [Theorem 15.4.4]

$$C_u = \frac{\partial}{\partial u} C(u,v) = Pr(V < v | U = u),$$

which is the conditional cdf. Then we use inverse transformation method to get V. □

Methodology 15.4.2 (conditional method for generating bivariate uniform random number with arbitrary joint distribution). *[10, p. 96]*

Suppose that we want to generate a set of n random variables with marginal uniform $U(0,1)$ and joint cdf given by copula $C_n(u_1, u_2, ..., u_n)$.

Further denote $C_{1:k}(u_1, ..., u_k)$ be the copula of $(U_1, U_2, ..., U_k), 2 \leq k \leq n$ and set $C_1(u_1) = u_1$.

We use the following procedures:

- *Simulate u_1 from $U(0,1)$.*
- *Simulate u_2 from the conditional distribution function $C_2(u_2|u_1)$.*
- *Simulate u_3 from the conditional distribution $C_3(u_3|u_1, u_2)$.*
- $\cdots$
- *Simulate u_n from the conditional distribution $C_n(u_n|u_1, u_2, ..., u_{n-1})$.*

where

$$\begin{aligned} C_k(u_k|u_1,...,u_{k-1}) &\triangleq Pr(U_k \le u_k | U_1 = u_1,...,U_{k-1} = u_{k-1}) \\ &= \frac{\frac{\partial^{k-1}}{\partial u_1 ... \partial u_{k-1}} C_{1:k}(u_1,...,u_k)}{\frac{\partial^{k-1}}{\partial u_1 ... \partial u_{k-1}} C_{1:k-1}(u_1,...,u_k - 1)} \end{aligned}$$

Proof. See Theorem 15.4.4 for how partial derivative is associated with conditional distribution. □

Example 15.4.9 (generate correlated uniform random variables with Gaussian copula correlation structure). Given a Gaussian copula C characterized by correlation matrix Σ. We can generate random vector $(X_1, X_2, ..., X_n)$ with uniform distribution and Gaussian copula correlation structure using the following procedures:

- First generate $(Y_1, Y_2, ..., Y_n) \sim MN(0, \Sigma)$.
- Then transform $X_1 = \phi(Y_1), X_2 = \phi(Y_2), ..., X_n = \phi(Y_n)$, where ϕ is standard normal cdf.

To understand the mechanism, note that Y_1, Y_2 alone is standard normal variable. Therefore $\phi(Y_1)$ and $\phi(Y_2)$ are uniform random variables [Lemma 15.4.2]. To show C is the copula associated with the cdf F, we have

$$\begin{aligned} F(x_1, x_2) &= Pr(X_1 \le x_1, X_2 \le x_2) \\ &= Pr(\phi(Y_1) \le x_1, \phi(Y_2) \le x_2) \\ &= Pr(\phi(Y_1) \le x_1, \phi(Y_2) \le x_2) \\ &= Pr(Y_1 \le \phi^{-1}(x_1), Y_2 \le \phi^{-1}(x_2)) \\ &= \Phi(\phi^{-1}(x_1), \phi^{-1}(x_2)) \\ &= C(x_1, x_2) \end{aligned}$$

where $C(u_1, u_2) = \Phi(\phi^{-1}(u_1), \phi^{-1}(u_2))$,$\Phi$ is the cdf for a multivariate normal distribution with zero mean and covariance matrix Σ, ϕ is the cdf for a standard normal variable. Note that the copula of a uniform distribution is itself [Lemma 15.4.5].

Example 15.4.10 (generate correlated uniform random variables with t copula correlation structure). [11, p. 420] Suppose we have n marginal distribution $F_i : \mathbb{R} \to [0, 1], i = 1, 2, ..., n$. Given a T copula C characterized by correlation matrix Σ and degree of freedom v. We can generate samples of random vector $(X_1, X_2, ..., X_n)$ characterized by margins $F_1, F_2, ..., F_n$ and copula C using the following procedures:

- Generate $Z_1, Z_2, ..., Z_n$ as iid $N(0, 1)$, and let $Z = (Z_1, Z_2, ..., Z_n)$.

- Generate a random $W \sim \chi^2(n)$ independent of Z.
- Return $X = \sqrt{\frac{v}{W}}CZ$, where C is the Cholesky decomposition of Σ such that $\Sigma = CC^T$.
- Return $U_i = T_v(X_i), i = 1, 2, ..., n$, where T_v is the univariate student t distribution with v degrees of freedom.

15.4.6.2 *Generating general correlated random number*

Methodology 15.4.3 (generate pair correlated random variables with Gaussian copula correlation structure). *Suppose we have two univariate marginal distribution $F_1 : \mathbb{R} \to [0,1]$ and $F_2 : \mathbb{R} \to [0,1]$. Given a Gaussian copula C characterized by correlation matrix Σ. We can generate samples of random vector (X_1, X_2) characterized by margins F_1, F_2 and copula C using the following procedures:*

- *First generate $(Y_1, Y_2) \sim MN(0, \Sigma)$.*
- *Then transform $X_1 = F_1^{-1}(\phi(Y_1)), X_2 = F_2^{-1}(\phi(Y_2))$, where ϕ is standard normal cdf.*
- *The random vector (X_1, X_2) has marginal distribution (F_1, F_2) and cdf*

$$F = C(F_1, F_2).$$

Proof. (1)Note that Y_1, Y_2 alone is standard normal variable. Therefore $\phi(Y_1)$ and $\phi(Y_2)$ are uniform random variables [Lemma 15.4.2]. Therefore X_1 and X_2 have marginal distribution (F_1, F_2). To show C is the copula associated with the cdf F, we have

$$\begin{aligned} F(x_1, x_2) &= Pr(X_1 \leq x_1, X_2 \leq x_2) \\ &= Pr(F_1^{-1}(\phi(Y_1)) \leq x_1, F_2^{-1}(\phi(Y_2)) \leq x_2) \\ &= Pr(\phi(Y_1) \leq F_1(x_1), \phi(Y_2) \leq F_2(x_2)) \\ &= Pr(Y_1 \leq \phi^{-1}(F_1(x_1)), Y_2 \leq \phi^{-1}(F_2(x_2))) \\ &= \Phi(\phi^{-1}(F_1(x_1)), \phi^{-1}(F_2(x_2))) \\ &= C(F_1(x_1), F_2(x_2)) \end{aligned}$$

where $C(u_1, u_2) = \Phi(\phi^{-1}(u_1), \phi^{-1}(u_2))$,$\Phi$ is the cdf for a multivariate normal distribution with zero mean and covariance matrix R, ϕ is the cdf for a standard normal variable. (2) Alternatively, we can use Methodology 15.4.6. □

Methodology 15.4.4 (generate correlated random variables with Gaussian copula correlation structure). *Suppose we have n marginal distribution $F_i : \mathbb{R} \to [0,1], i = 1, 2, ..., n$. Given a Gaussian copula C characterized by correlation matrix Σ. We can generate*

samples of random vector $(X_1, X_2, ..., X_n)$ *characterized by margins* $F_1, F_2, ..., F_n$ *and copula* C *using the following procedures:*

- *First generate* $(Y_1, Y_2, ..., Y_n) \sim MN(0, \Sigma)$.
- *Then transform* $X_1 = F_1^{-1}(\phi(Y_1)), X_2 = F_2^{-1}(\phi(Y_2)), ..., X_n = F_n^{-1}(\phi(Y_n))$,, *where* ϕ *is standard normal cdf.*
- *The random vector* $(X_1, X_2, ..., X_n)$ *has marginal distribution* $(F_1, F_2, ..., F_n)$ *and cdf*

$$F = C(F_1, F_2, ..., F_n).$$

Example 15.4.11 (generation of dependent default time). Let $T_1, T_2, ..., T_n$ denote the random default time for n loan borrowers. Assume the hazard curve for each borrower is given by $h_i(t), t \geq 0$ such that the marginal cdf of default time is given by

$$F_i(t) = Pr(T_i \leq t) = 1 - \exp(-\int_0^t h(s)ds).$$

Further assume the copula associated with the joint cdf is Gaussian copula with correlation matrix R.

Consider the following random number generating process

- Simulate $Y_1, Y_2, ..., Y_n$ from $MN(0, R)$.
- Obtain $T_1, T_2, ..., T_n$ using $T_i = F_i^{-1}(\phi(Y_i))$, where ϕ is the cdf of a standard normal.

Then samples of $T_1, T_2, ..., T_n$ will follow the joint cdf F.

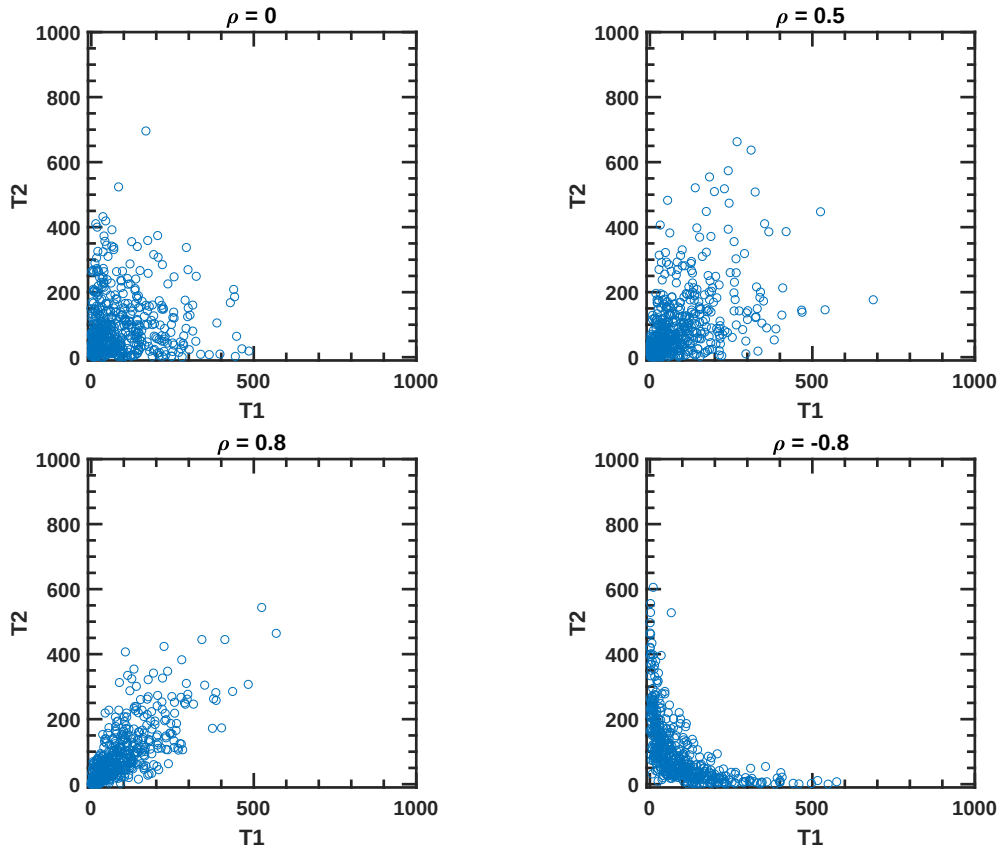


Figure 15.4.3: Generated correlated default time via Gaussian copula with different correlations. The hazard rate for both parties is $h(t) = 0.01$.

Methodology 15.4.5 (generating correlated random number with t copula). *Suppose we have n marginal distribution $F_i : \mathbb{R} \to [0,1], i = 1,2,...,n$. Given a T copula C characterized by correlation matrix Σ and degree of freedom v. We can generate samples of random vector $(X_1, X_2, ..., X_n)$ characterized by margins $F_1, F_2, ..., F_n$ and copula C using the following procedures:*

- *Generate $Z_1, Z_2, ..., Z_n$ as iid $N(0,1)$, and let $Z = (Z_1, Z_2, ..., Z_n)$.*
- *Generate a random $W \sim \chi^2(n)$ independent of Z.*
- *Return $X = \sqrt{\frac{v}{W}} CZ$, where C is the Cholesky decomposition of Σ such that $\Sigma = CC^T$.*
- *Set $U_i = T_v(X_i), i = 1,2,...,n$, where T_v is the univariate student t distribution with v degrees of freedom.*
- *Return $Y_i = F_i^{-1}(U_i), i = 1,2,...,n$.*

Example 15.4.12 (copula method for pricing basket option). Consider a financial basket option with payoff at future time T given by

$$V_T = max(\sum_{i=1}^{n} \frac{1}{n} S_T^{(i)} - K, 0),$$

where $S_T^{(i)}, i = 1, ..., n$ is the stochastic stock price i at time T, K is a constant.

To evaluate V_T via Monte Carlo method, we can generate samples of $S_T^{(i)}, i = 1, ..., n$ and then evaluate the expectation.

We make the following assumptions:

- We can construct the implied cdf F_i for each $S_T^{(i)}, i = 1, 2, ..., n$ by estimating historical data.
- The joint distribution of $S_T^{(1)}, S_T^{(2)}, ..., S_T^{(n)}$ has Gaussian copula with correlation matrix Σ. Note that the correlation matrix can be estimated from historical data.

Then we can use the following simulation method to generate one sample

- First generate $(Y_1, Y_2, ..., Y_n) \sim MN(0, \Sigma)$.
- Return $S_T^{(1)} = F_1^{-1}(\phi(Y_1)), S_T^{(2)} = F_2^{-1}(\phi(Y_2)), ..., S_T^{(n)} = F_n^{-1}(\phi(Y_n))$,, where ϕ is standard normal cdf.

We can similarly generate samples with t copula.

Methodology 15.4.6 (transform correlated uniform distribution to arbitrary distribution with same copula structure). *Let $(U_1, U_2, ..., U_n)$ be a uniform random vector with cdf F(or equivalently copula$C$ [Lemma 15.4.5]. Let $F_1, F_2, ..., F_n$ be the marginal cdf of a target cdf F^{target}. It follows that the random vector $(F_1^{-1}(U_1), F_2^{-1}(U_2), ..., F_n^{-1}(U_2))$ has marginal cdf $F_1, F_2, ..., F_n$ and copula structure C; that is, the cdf of $(F_1^{-1}(U_1), F_2^{-1}(U_2), ..., F_n^{-1}(U_2))$ is F^{target}.*

Proof. Let $U_i \sim U(0, 1)$, then

$$Pr(F_i^{-1}(u_i) < x_i) = Pr(u_i < F_i(x_i)) = F_i(x_i),$$

which implies that $F_i^{-1}(U_i)$ has marginal F_i.

The show the cdf of $(F_1^{-1}(U_1), F_2^{-1}(U_2), ..., F_n^{-1}(U_2))$ has the same copula as $(U_1, U_2, ..., U_n)$, we use the monotone transform invariance property of copula [Lemma 15.4.7]. □

15.4.6.3 *Multivariate distribution approximation with Gaussian copula*

Theorem 15.4.5. *Given random variable $X_1, X_2, ..., X_n$ with margins $F_{X_1}, F_{X_2}, ..., F_{X_n}$ and pair-correlation matrix R. We can construct a multivariate distribution such that recovers the margins and correlation via*

$$F(x_1, x_2, ..., x_n) = \Phi(\phi^{-1}(F_1(x_1)), \phi^{-1}(F_2(x_2)), ..., \phi^{-1}(F_n(x_n))),$$

Φ is the cdf for a multivariate normal distribution with zero mean and covariance matrix R, ϕ is the cdf for a standard normal variable.

Proof. From the theorem of constructing multivariate distribution from margins [Theorem 15.4.1], we have

$$F(x_1, x_2, ..., x_n) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d)).$$

Note that the Gaussian copula with correlation matrix R [Definition 15.4.4] is given by

$$C(u_1, u_2, ..., u_d; R) = \Phi(\phi^{-1}(u_1), \phi^{-1}(u_2), ..., \phi^{-1}(u_d)),$$

where $u_1, u_2, ..., u_d \in [0, 1]$, Φ is the cdf for a multivariate normal distribution with zero mean and covariance matrix R, ϕ is the cdf for a standard normal variable. □

Remark 15.4.12 (implications).

- To fully determine the multivariate distribution, we usually need all the margins and all cross-term moments.
- With only margins and correlations given, we can construct a multivariate Gaussian distribution as an approximation.

15.5 Covariance structure and factor analysis

15.5.1 The orthogonal factor model

15.5.1.1 *Motivation and factor models*

In some real world applications, we often face the challenges of modeling the dependence for a large set of random variables. Directly modeling of the joint distribution is usually intractable. In section 15.4, we introduce Copula method as a parametric method to model dependence. In the section, we introduce factor model, which aims model the correlation structure by connecting them to a few random variables, called latent factors[2, p. 482].

Consider a p dimensional observable random vector X and a $m, m \ll p$ dimensional random vector F, called common/latent factors. Let $E[X] = \mu, \mu \in \mathbb{R}^p, Cov(X) = \Sigma, \Sigma \in \mathbb{R}^{p \times p}$. In the **orthogonal factor model**, **we assume** X **is related to** F via

$$X - \mu = LF + \epsilon,$$

where

- $L \in \mathbb{R}^{p \times m}$ is called matrix of **factor loadings**;
- ϵ is a p dimensional random vector such that $E[\epsilon] = 0_{p \times 1}, Cov(\epsilon) = \psi, Cov(\epsilon)_{ij} = \psi_i, i = j, Cov(\epsilon)_{ij} = 0, i \neq j$.
- $E[F] = 0_{m \times 1}, Cov(F) = I_{m \times m}$.
- F and ϵ are independent such that $Cov(\epsilon, F) = 0$.

orthogonal factor model

Remark 15.5.1 (parameter summary in factor models). The original covariance structure $\Sigma \in \mathbb{R}^{p \times p}$ has $p(p+1)/2$ parameters. In the new covariance structure resulted from factor model, we have in total $p(m+1)$ parameters:

- mp factor loadings in the matrix L.
- p specific variance in the matrix ψ.

15.5.1.2 *Covariance structure implied by factor model*

An important application of factor models is to serve as a low dimensional approximation to high-dimensional covariance matrix of the random vector $X_1, ..., X_p$. In the factor model, we use a L matrix of pm elements and ψ matrix of p elements to approximate/reproduce the original covariance matrix Σ of X, containing $p(p+1)/2$ elements.

Lemma 15.5.1 (covariance structure implied by factor model). *[2, p. 483] In the factor model, the covariance of X is given by*

$$E[(X-\mu)(X-\mu)^T] = LL^T + \psi.$$

In addition, the covariance matrix between X and F is given by

$$Cov(X,F) \triangleq E[(X-\mu)F^T] = L.$$

Proof. (1)

$$\begin{aligned} E[(X-\mu)(X-\mu)^T] &= E[(LF+\epsilon)(LF+\epsilon)^T] \\ &= E[LFF^TL^T + \epsilon(LF)^T + LF\epsilon^T + \epsilon\epsilon^T] \\ &= LIL^T + 0 + 0 + \psi. \end{aligned}$$

(2)

$$\begin{aligned} E[(X-\mu)F^T] &= E[(LF+\epsilon)F^T] \\ &= E[LFF^T + \epsilon F^T] \\ &= LI + 0 = L. \end{aligned}$$

□

Remark 15.5.2 (non-uniqueness of factor model). **[anderson2009introduction]**[2, pp. 488, 504] Let Σ, L, F, ψ be the results of factorization such that

$$X - \mu = LF + \epsilon,$$

and

$$\Sigma = LL^T + \psi.$$

Let $U \in \mathbb{R}^{m\times m}$ be a orthonormal matrix such that $UU^T = I$. Then the factors and the associated factor loadings can also take form $\hat{L} = LU, \hat{F} = U^TF$ also satisfy

$$X - \mu = \hat{L}\hat{F} + \epsilon,$$

and

$$\Sigma = \hat{L}\hat{L}^T + \psi.$$

To understand this, note that

$$\hat{L}\hat{F} = LUU^TF = LUU^TL^T = LL^T$$

and

$$\hat{L}\hat{L}^T = LU(LU)^T = LUU^TL^T = LL^T.$$

Example 15.5.1. Consider a one factor model given by

$$V_i = a_i Y + \sqrt{1 - a_i^2} Z_i, i = 1, 2, ..., n,$$

where $(F, Z_1, Z_2, ..., Z_n)$ are independent standard normal variables.

Then the covariance structure implied by the factor model is given by

$$\begin{aligned} Cov &= \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix} [a_1, a_2, ..., a_n] + \begin{bmatrix} 1 - a_1^2 & & & \\ & 1 - a_2^2 & & \\ & & \ddots & \\ & & & 1 - a_n^2 \end{bmatrix} \\ &= \begin{bmatrix} a_1^2 & a_1 a_2 & \cdots & a_1 a_n \\ a_2 a_1 & a_2 a_2 & \cdots & a_2 a_n \\ \vdots & \cdots & \ddots & \vdots \\ a_n a_1 & a_n a_2 & \cdots & a_n a_n \end{bmatrix} \begin{bmatrix} 1 - a_1^2 & & & \\ & 1 - a_2^2 & & \\ & & \ddots & \\ & & & 1 - a_n^2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & a_1 a_2 & \cdots & a_1 a_n \\ a_2 a_1 & 1 & \cdots & a_2 a_n \\ \vdots & \cdots & \ddots & \vdots \\ a_n a_1 & a_n a_2 & \cdots & 1 \end{bmatrix} \end{aligned}$$

which is also a valid correlation matrix.

15.5.2 Parameter estimation

15.5.2.1 *Data collection and preparation*

In the factor model, we need to estimate L and Ψ. We will go through two methods PCA and maximum likelihood estimation. The first step is to prepare data.

- Suppose each data vector x_i has p components. Then we can form a $n \times p$ data matrix X.

- We can transform the data matrix to sample covariance matrix [subsection 15.1.1] via
$$S = \frac{1}{n-1} X^T (I - \frac{1}{n} J) X.$$
- The sample covariance matrix and sample correlation matrix will then be used to build factor models.

15.5.2.2 *PCA method*

Lemma 15.5.2 (PCA method for estimation of factor loadings). *[2, p. 488] Let $S \in \mathbb{R}^{p \times p}$ be the sample covariance matrix of data matrix X. Let S adopt eigendecomposition*

$$S = \sum_{i=1}^{p} \lambda_i u_i u_i^T,$$

where $\lambda_1 \leq \lambda_2 \cdots \leq \lambda_p$.

It follows that

- *Then an estimation of column i of factor loading matrix L is given by*
$$\hat{L}_i = \sqrt{\lambda_i} u_i$$
such that
$$\hat{L}\hat{L}^T = \sum_{i=1}^{m} \lambda_i u_i u_i^T.$$
- *An estimation of ψ is given by*
$$\hat{\psi} = diag(S - \hat{L}\hat{L}^T).$$
- *The factors $F_1, F_2, ..., F_m$ are assumed to be independent standard normal random variables estiamted to have samplings given by*

Proof. Note that
$$\hat{L}_i \hat{L}_j^T = 0, \forall i \neq j.$$
Then
$$LL^T = \sum_{i=1}^{m} \hat{L}_i \hat{L}_i^T = \sum_{i=1}^{m} \lambda_i u_i u_i^T.$$
□

Remark 15.5.3 (how to determine the number of factors)**.** We can empirically determine the number of factors based on the following considerations:

- m is chosen to explain most of the variance.
- m is chosen such that the number of parameters $p(m+1)$ is smaller than the number of parameters $p(p+1)/2$ in the full covariance matrix.

15.5.2.3 *Maximum likelihood method*

Using the Maximum Likelihood Estimation Method we must assume that the data are independently sampled from a multivariate normal distribution with mean vector μ and variance-covariance structure take the form of

$$\Sigma = LL^T + \psi,$$

where $L \in \mathbb{R}^{p\times m}$ is the factor loadings and ψ is the diagonal matrix of specific variances.

Methodology 15.5.1. *[2, p. 496] Suppose we have data vectors from n observations $x_1, x_2, ..., x_n \in \mathbb{R}^p$. The maximum likelihood estimation involves estimating the mean μ, the matrix of factor loadings, and the specific variance matrix ψ.*

The likelihood function is given by

$$L(\mu,\Sigma) = (2\pi)^{-\frac{np}{2}}|\Sigma|^{-\frac{n}{2}}\exp(-\frac{1}{2}Tr(\Sigma^{-1}(\sum_{j=1}^{n}(x_j-\mu)(x_j-\mu)^T)))$$

where $\Sigma = LL^T + \psi$. and the log likelihood function is given by

$$l(\mu,\Sigma) \triangleq \ln L = \frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}Tr(\Sigma^{-1}(\sum_{j=1}^{n}(x_j-\mu)(x_j-\mu)^T)).$$

15.5.3 Factor score estimation

Methodology 15.5.2 (weighted least square method to find factor score). *Suppose we have data vectors from n observations $x_1, x_2, ..., x_n \in \mathbb{R}^p$. Further suppose that we have already had the factor model given by*

$$X = \mu + Lf + \epsilon,$$

with $\mu \in \mathbb{R}^p, L \in \mathbb{R}^{p\times m}$ being given. Then the factor scores $f_1, f_2, ..., f_n \in \mathbb{R}^m$ associated with data vector can be estimated by minimizing the following optimization problem

$$\min_{f_i \in \mathbb{R}^m} (x_i - \mu - Lf_i)^T \psi^{-1} (x_i - \mu - Lf_i).$$

Furthermore, the minimizer is given by [see generalized least square Theorem 16.1.12]

$$\hat{f}_i = (L^T \psi^{-1} L)^{-1} L^T \psi^{-1} (x_i - \mu).$$

15.5.4 Application I: Joint default modeling

15.5.4.1 *Single factor model*

We consider a financial application of factor modeling. Consider n parties that will default before the next period t with unconditional probability p_i (i.e., a Bernoulli random variable). Because default behavior of multiple parties can be quite correlated, particularly during periods of financial crisis, we also like to capture their **joint default behavior** for risk management purpose.

Define a new proxy **standard random variable** X_i by

$$X_i = a_i F + \sqrt{1 - a_i^2} Z_i, i = 1, 2, ..., n,$$

where F is a common factor (such as GDP) affecting defaults for all parties and Z_i is a noise term affecting only party i. F and Z_i are independent standard normal variables.

It follows that

- $$Pr(X_i \ will \ default) \triangleq p_i = Pr(X_i < \phi^{-1}(p_i)),$$
  where ϕ is the cdf of the standard normal variable.
- The **conditional default probability** of firm i before t conditioning on the observation of F is given by

$$Pr(X_i \ will \ default | F = f) \triangleq Pr(T_i < t | F = f) = \phi(\frac{\phi^{-1}(p_i) - a_i f}{\sqrt{1 - a_i^2}}),$$

To see this, note that $X_i \sim N(0, 1)$. Therefore,

$$Pr(X_i < \phi^{-1}(p_i)) = \phi(\phi^{-1}(p_i)) = p_i.$$

For the second point,

$$\begin{aligned} Pr(X_i < \phi^{-1}(p_i)|F = f) &= Pr(a_i f + \sqrt{1 - a_i^2} Z_i < \phi^{-1}(p_i)) \\ &= Pr(Z_i < \frac{\phi^{-1}(p_i) - a_i f}{\sqrt{1 - a_i^2}}) \end{aligned}$$

We can see that when a_i is large, all parties tend to default together when an event $f < 0$ occurs. On the other hand, when $a_i = 0$, all parties will default independently.

The single factor approach can be extended to capture the **joint default time modeling**. Let $T_1, T_2, ..., T_n$ denote the random default time for n parties. Assume marginal cdf of default time is given by $Q_i(t)$. Define a new proxy random variable $X_i = \phi^{-1}(Q_i(T_i)), i = 1, 2, ..., n$, and **assume** X_i can be modeled by

$$X_i = a_i F + \sqrt{1 - a_i^2} Z_i, i = 1, 2, ..., n,$$

where F is a common factor affecting defaults for all firms and Z_i is a factor affecting only firm i. F and Z_i are independent standard normal variables. It follows that the **conditional default probability** of firm i before t conditioning on the observation of F is given by

$$Q_i(t|F = f) \triangleq Pr(T_i < t|F = f) = \phi(\frac{\phi^{-1}(Q_i(t)) - a_i f}{\sqrt{1 - a_i^2}}),$$

where ϕ is the standard normal cdf.

To see this, we have

$$\begin{aligned} Pr(T_i < t) &= Pr(T_i < t) \\ &= Pr(Q_i(T_i) < Q_i(t)) \\ &= Pr(\phi^{-1}(Q_i(T_i)) < \phi^{-1}(Q_i(t))) \\ &= Pr(X_i < \phi^{-1}(Q_i(t))) \\ &= Pr(a_i F + \sqrt{1 - a_i^2} Z_i < \phi^{-1}(Q_i(t))) \\ &= Pr(Z_i < \frac{\phi^{-1}(Q_i(t)) - a_i F}{\sqrt{1 - a_i^2}}) \\ \implies Q_i(T|F = f) \triangleq Pr(T_i < t|F) &= Pr(Z_i < \frac{\phi^{-1}(Q_i(T)) - a_i F}{\sqrt{1 - a_i^2}}|F = f) \\ &= \phi(\frac{\phi^{-1}(Q_i(t)) - a_i f}{\sqrt{1 - a_i^2}}) \end{aligned}$$

Remark 15.5.4 (the correlation structure for the Gaussian copula implied by the factor model)**.** Consider a one factor model given by

$$V_i = a_i Y + \sqrt{1 - a_i^2} Z_i, i = 1, 2, ..., n,$$

where $(F, Z_1, Z_2, ..., Z_n)$ are independent standard normal variables.

Then the correlation structure implied by the factor model is given by [Figure 15.5.1]

$$\begin{aligned} Cov &= \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix} [a_1, a_2, ..., a_n] + \begin{bmatrix} 1 - a_1^2 & & & \\ & 1 - a_2^2 & & \\ & & \ddots & \\ & & & 1 - a_n^2 \end{bmatrix} \\ &= \begin{bmatrix} a_1^2 & a_1 a_2 & \cdots & a_1 a_n \\ a_2 a_1 & a_2 a_2 & \cdots & a_2 a_n \\ \vdots & \cdots & \ddots & \vdots \\ a_n a_1 & a_n a_2 & \cdots & a_n a_n \end{bmatrix} \begin{bmatrix} 1 - a_1^2 & & & \\ & 1 - a_2^2 & & \\ & & \ddots & \\ & & & 1 - a_n^2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & a_1 a_2 & \cdots & a_1 a_n \\ a_2 a_1 & 1 & \cdots & a_2 a_n \\ \vdots & \cdots & \ddots & \vdots \\ a_n a_1 & a_n a_2 & \cdots & 1 \end{bmatrix} \end{aligned}$$

(a) Calculation of first default probability(thick solid black) from individual default probability of 10 reference names using Gaussian one factor model with $\rho = 0.5$.

(b) First default probability as a funciton of correlation of the underlying names.

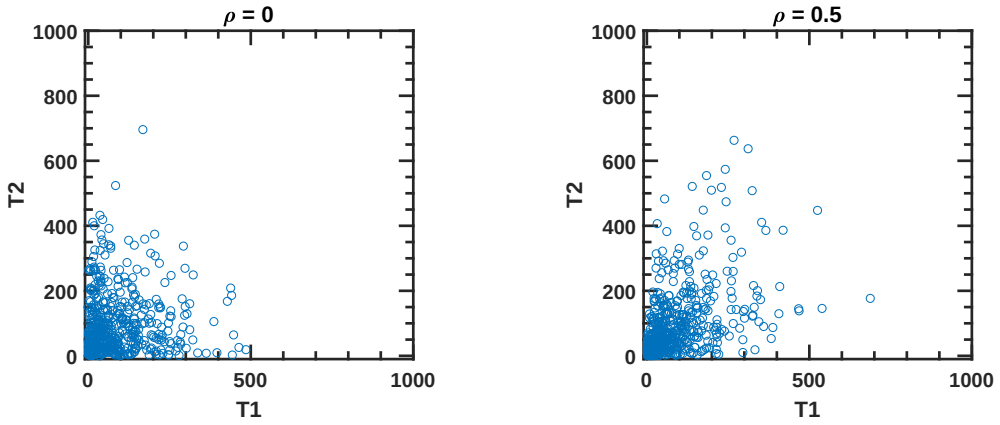


Figure 15.5.1: Correlation structure for the Gaussian copula implied by the factor model.

15.5.4.2 *Multiple factor model*

Finally, we can introduce multiple factors to give a richer representation of the correlation structure, as we make the following modification. Let $T_1, T_2, ..., T_n$ denote the random default time for n parties. Assume marginal cdf of default time is given by $Q_i(t)$. Define a new proxy random variable $X_i = \phi^{-1}(Q_i(T_i)), i = 1, 2, ..., n$, and **assume** X_i can be modeled by

$$X_i = \sum_{j=1}^{m} a_{ij} F_j + \sqrt{1 - \sum_{j=1}^{m} a_{ij}^2} Z_i, i = 1, 2, ..., n,$$

where $F_1, F_2, ..., F_m$ are common factors affecting defaults for all firms and Z_i is a factor affecting only firm i. $F_1, F_2, ..., F_m$ and Z_i are mutually independent standard normal variables.

The **conditional default probability** of firm i before t conditioning on the observation of $F_1, F_2, ..., F_m$ is given by

$$Q_i(t|F_1 = f_1, ..., F_m = f_m) \triangleq Pr(T_i < t|F_1 = f_1, ..., F_m = f_m) = \phi\left(\frac{\phi^{-1}(Q_i(t)) - \sum_{j=1}^{m} a_{ij} f_j}{\sqrt{1 - \sum_{j=1}^{m} a_{ij}^2}}\right),$$

where ϕ is the standard normal cdf.

15.5.5 Application II: factor models for stock return

15.5.5.1 *Overview*

In portfolio management and financial risk analytics, it is often desirable to identify a small set of risk factors that underlying the stochastic dynamics of hundreds and thousands of stocks.

Mathematically, let $r_1, ..., r_n$ be the stochastic return of n assets. Seeking risk factors can be simplified to a linear risk model given by

$$r_i(t) = a_i(t) + \sum_{j=1}^{m} b_{i,j}(t) f_j(t) + e_i(t), \forall i = 1, 2, ..., n$$

where $f_1, ..., f_m, m \ll n$ are stochastic risk factors, e_i is noise term specific to asset i satisfying $E[e_i] = 0, E[e_i e_j] = 0, \forall i \neq j$.

The applications of this linear factor model include

- Capture the correlation structure via fewer parameters.
- Managing risks from $r_1, ..., r_n$ is reduced to managing risks from a smaller set of risks $f_1, ..., f_n$.
- Understand the risk dynamics by studying these common factors.

These factors $f_1, ..., f_M$ can be external, such that GDP, inflation, etc. they can also be internal, such as linear combination of X_1, X_2, X_3 obtained via PCA or other statistical methods like data mining.

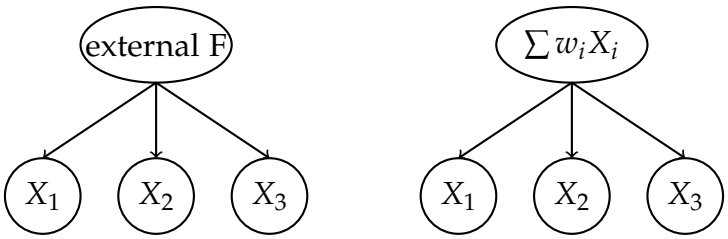


Figure 15.5.2: Factor model using (a) external factors or (b) internal factors.

Example 15.5.2 (a macroeconomic factor model). [13, p. 635] Consider a factor model in which the returns of stocks are correlated with surprises in interest rates and surprises in GDP growth. For stock $i = 1, 2, ..., N$, the return is modeled by

$$R_i = a_i + b_{i1} F_{INT} + b_{i2} F_{GDP} + \epsilon_i,$$

where

- R_i is the return of stock i
- a_i is the expected return to stock i
- b_{i1} is the sensitivity of the return to stock i to interest rate surprise
- b_{i2} is the sensitivity of the return to stock i to GDP growth surprise
- F_{INT} is the surprise in interest rate
- F_{GDP} is the surprise in GDP
- ϵ_i an error term with a zero mean that represents the portion of the return to stock i not explained by the factor model

Note that we define **surprise** in general as the actual value minus the predicted(or exptected) value. For example,

$$actual\ inflation = \ predicted\ inflation\ +\ inflation\ surprise.$$

Example 15.5.3 (factor model for portfolio optimization). In the classical portfolio optimization framework, we require the mean and covariance matrix for the n assets. They can be related to the single-factor model parameters as:

$$E[r_i] = a_i + \sum_{j=1}^{m} b_{ij} E[f_j]$$

$$\sigma_i^2 = Var[r_i] = \sum_{j=1}^{m} b_{ij}^2 \sigma_{f_j}^2 + \sum_{k<j}^{m} 2 b_{ik} b_{ij} cov(f_k, f_j) + \sigma_{e_i}^2$$

$$\sigma_{ij} = Cov(r_i, r_j) = \sum_{k=1}^{m} \sum_{p=1}^{m} b_{ik} b_{jp} cov(f_k, f_p), i \neq j$$

Remark 15.5.5 (cross-panel parameter fitting). Consider a model of return

$$r = Xb + u,$$

where r is an N vector of excess returns, X is an N by K matrix of factor exposures, b is a K dimension vector of factor returns, and u is an N dimensional vector of specific returns, and we further assume the factor loading matrix X is given.

Then we can obtain factor vector b from the weighted least square minimization given by

$$\min_b (Xb - r)^T \Delta^{-1} (Xb - r).$$

The final result is from [Theorem 16.1.12]:

$$b = (X^T \Delta^{-1} X)^{-1} X^T \Delta^{-1} r.$$

15.5.5.2 *The Fama-French 3 factor model*

The section we study the arguably most famous factor model, known as the Fama-French 3 factor model[14, 15], for stock market returns. In the Fama-French 3 factor model, the asset i return is given by

$$r_i = E[r_i] + \beta_{i1}(R_m - E[R_m]) + \beta_{i2}(SMB - E[SMB]) + \beta_{i3}(HML - E[HML]) + \epsilon_i$$

where

- $R_m - r_f$ is the return on a market value-weighted index in excess of the one-month T-bill rate.
- SMB ='small [market capitalization] minus big' factor. Mathematically, SMB is the average **raw**[1] return on three small-cap portfolios minus the average return on three large-cap portfolio. When small stocks do well relative to large stocks this will be positive, and when they do worse than large stocks, this will be negative. SMB also refers to as **size premium**. Note that $E[SMB] \neq 0$.
- HML = 'high [book/price] minus low' factor. Mathematically, HML is the average **raw** return on two high book-to-market portfolios minus the average return on two low book-to-market portfolios. HML also refers to a **value premium**. $E[HML] \neq 0$.

Remark 15.5.6 (factor portfolios as proxy risk drivers). [16, p. 70]

- The fundamental reasons determines a stock's return is the company's management system, technology, the ability to adaptive, efficiency etc. However, these reasons/factors are hard to quantify and observe. Therefore, we use the performance of different companies as the proxy to these factors.
- The Fama-French model views the size and value factors as representing ('proxying for') a set of underlying risk factors. For example, small market-cap companies may be subject to risk factors such as less ready access to private and public credit markets and competitive disadvantages. High book-to-market may represent shares with depressed prices because of exposure to financial distress. The model views the return premiums to small size and value as compensation for bearing types of systematic risk.
- Fama and French create a portfolio designed to have returns that mimic the returns associated with the size effect and propose using the returns of this portfolio as a risk factor.

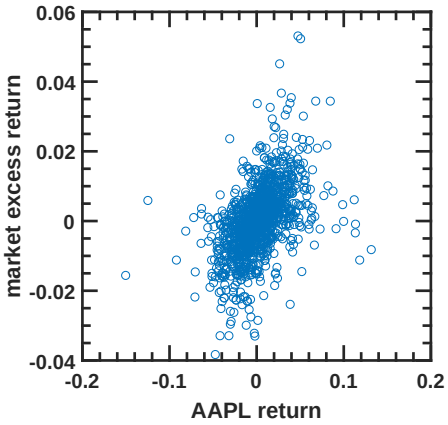
We first analyze the statical properties of the risk factors. Table 15.5.1 shows the monthly excess return statistics of Fama-french 3 factor portfolios. We can see that SMB and HML portfolios have positive premium, which agree with the size and value premium we discussed.

1 when say here raw return to emphasize that it is not excess return

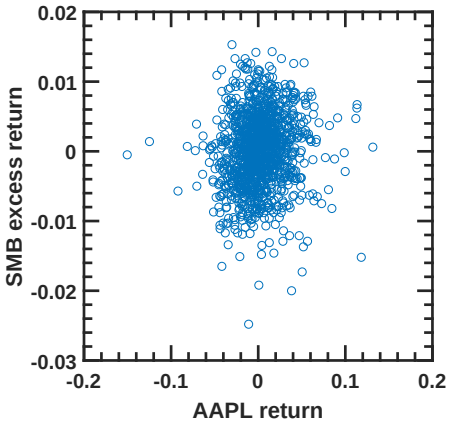
Table 15.5.1: statistics on Fama-French 3 factors from July 1963 to Dec. 1991.

factor name	**mean**	**std**	**correlation**		
			$r_M - r_f$	$SMB - r_f$	$HML - r_f$
$r_M - r_f$	0.43	4.54	1		
$SMB - r_f$	0.27	2.89	-0.38	1	
$HML - r_f$	0.40	2.54	0.34	-0.08	1

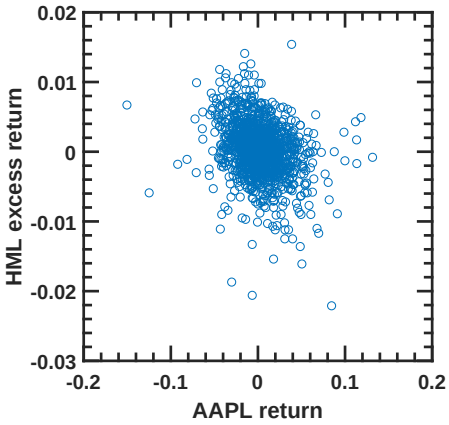
For each asset i, say AAPL stock, we can perform a linear regression based on observed returns and factor data to get coefficients $\beta_1, \beta_2, \beta_3$. Figure 15.5.3 shows the AAPL return data vs. factor realization in a historical period. Clearly, AAPL return is quite positively correlated with the factor R_M, and much less with SMB and HML. More detailed results can be found in Table 15.5.2.



(a) Scatter plot for AAPL daily return vs. market excess daily return from 2001-Oct to 2006-Oct.



(b) Scatter plot for AAPL daily return vs. SMB factor excess daily return from 2001-Oct to 2006-Oct.



(c) Scatter plot for AAPL daily return vs. HML factor excess daily return from 2001-Oct to 2006-Oct.

Figure 15.5.3: Scatter plot of AAPL return vs. market excess return, SMB excess return and HML excess return.

Table 15.5.2: AAPL stock return modeled by the Fama-French 3 factor model.

	estimate	**SE**	**tStat**	**pValue**
α	0.00186	0.00065	2.877	0.0041
β_{MKT}	1.2058	0.07012	17.196	1.4e-59
β_{SMB}	0.3753	0.12183	3.0806	0.00211
β_{HML}	-0.6583	0.17566	-3.7475	0.000187
R^2	0.264	adjusted R^2	0.263	

15.6 Graphical models

15.6.1 Fundamentals

Graphical models are an intuitive way of representing and visualizing the relationships between many random variables. A graph can help extract the conditional independence relationships among random variables. Thus we can answer questions like "Is A independent from B given that we know the value of C?"

Graphical models can be roughly divided into directed graphical models and undirected graphical models, Our focus is directed graphical models, known as Bayesian graphical model. More formally, a graph $G = (\mathcal{V}, \mathcal{E})$ consists of a set of nodes or vertices $\mathcal{V} = \{1, 2, ..., V\}$, and a set of edges $\mathcal{E} = \{(s, t) : s, t \in \mathcal{V}\}$. Nodes represent random variables, and edges representing assumed causal relationship and conditional independence between nodes or random variables. The connectivity between nodes can be represented by a matrix G, where $G(s, t) = 1$ if $(s, t) \in \mathcal{E}$. **Directed acyclic graph(DAG)** is a directed graph with no directed cycles. Related random variables of a random variable are classified as **parents**, **children**, **ancestors**, and **descendants**, which are defined by

- Parent of a node: $Pa(s) = \{t | G(t, s) = 1\}$
- Child of a node: $ch(s) = \{t | G(s, t) = 1\}$
- ancestors of a node: $anc(t)$ is the set of nodes s that has a directed path from s to t.
- descendants of a node: $anc(t)$ is the set of nodes s that has a directed path from t to s.

A directed graphical model $G = (\mathcal{V}, \mathcal{E})$ is a type of graphical model whose graph is a directed acyclic graph(DAG). In the graph, each node s is conditional independent of all its ancestor nodes except the parent nodes when conditioned on the parent nodes of s. Directed graphical model is also called **Bayes network**(the parameter as a random variable can be represented by a node), **brief network**, and **causal network**(because the directed arrows can be interpreted as causal relations.)

The first important application of graphical model is to allow decomposition and factorization of a complex joint distribution into simpler one.

We consider a naive decomposition of joint distributions.

Lemma 15.6.1 (joint distribution decomposition,chain rule). *Let P be the joint distribution on random variables $X_{1:N} \triangleq (X_1, X_2, ..., X_N)$, then we can decompose the joint distribution as*

$$P(X_{1:V}) = P(X_1)P(X_2|X_1)P(X_3|X_{1:2}...P(X_N|X_{1:N-1}) = \prod_{i=1}^{N} P(X_i|X_{1:i-1}).$$

This decomposition still holds if we permute the index of the X_i.

Proof. Apply the definition of conditional distribution $P(X, Y) = P(X)P(Y|X)$ from front to back. □

For any joint distribution, we can always represent it by a fully connected graphical model, as showed in Figure 15.6.1.

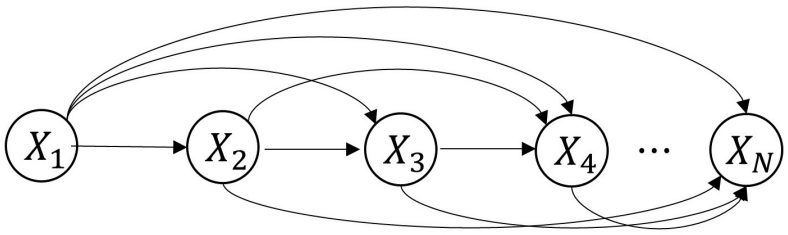


Figure 15.6.1: A fully connected graphical model representing the joint distribution $P(X_1, ..., X_N)$.

Graphical models can be used to represent the conditional independence among a set of random variable. Formally, we have

Definition 15.6.1 (set of conditional independence). *[17, p. 60][18, p. 324]Let P be a distribution over a set of random variables $\mathcal{X}$. Random variables X and Y are conditional independent given Z, denoted as $X \perp Y|Z$, if*

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

the set of all such conditional independence is denoted as $I(P)$.

A graph model G is an independence map for a joint distribution P if

$$I(G) \subseteq I(P)$$

where $I(G)$ the set of all conditional independence assumptions encoded by G.

Therefore, a fully connected graph G [Figure 15.6.1] is the I-map for all distributions P defined over the same random variables. Note that for a fully connected graph G, $I(G) = \emptyset$.

The goal of graphical model is to capture all the conditional independence relationship, but not omit any, for a joint distribution P. In other words, we are seeking a graph model G that the **minimal I-map** of P.

A graphical model can lead to simplified factorization of a joint distribution.

Theorem 15.6.1 (factorization theorem). *If graph G is an I-map of P, then*

$$P(X_1...X_n) = \prod_i^n P(X_i|Pa(X_i))$$

Proof. by the chain rule lemma, we have

$$P(X_1...X_n) = \prod_i^n P(X_i|X_1...X_i))$$

□

Example 15.6.1. Consider the following graph models in Figure 15.6.2.

- For graphical model (A), we can have the following factorization

$$P(A, B, C, D, E) = P(A)P(B)P(C|A, B)P(D|B, C)P(E|C, D).$$

- Consider the graphical model (B). A data set of N points are generated iid from a Gaussian distribution with parameters μ and σ. The joint probability is given by

$$P(X_1, X_2, ..., X_N, \mu, \sigma) = P(\mu)P(\sigma)\prod_{n=1}^N P(X_n|\mu, \sigma).$$

- Consider the graphical model (C) representing a Markov chain. The joint probability can be decomposed by

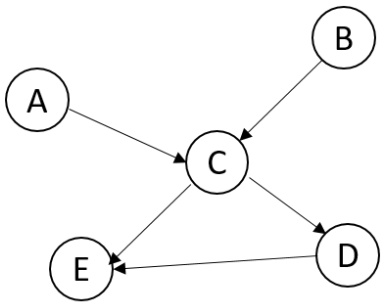
$$P(X_1, X_2, ..., X_N) = P(X_N|X_{N-1})P(X_{N-1}|X_{N-2})\cdots P(X_1).$$

- Consider the graphical model (D) representing a second-order Markov chain. The joint probability can be decomposed by

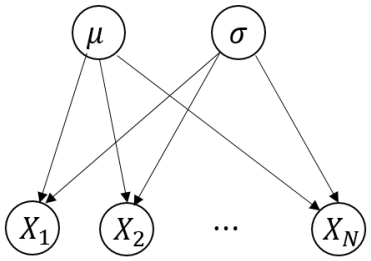
$$P(X_1, X_2, ..., X_N) = P(X_N|X_{N-1}, X_{N-2})P(X_{N-1}|X_{N-2}, X_{N-3}) \cdots P(X_3|X_1, X_2)P(X_1)P(X_2).$$

- Consider the graphical model (E) representing a hidden Markov chain. The joint probability can be decomposed by

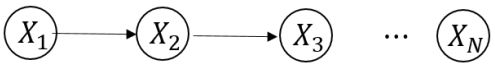
$$P(X_1, ..., X_N, Z_1, ..., Z_N) = P(Z_N|Z_{N-1})P(Z_{N-1}|Z_{N-2}) \cdots P(Z_1) \prod_{n=1}^{N} P(X_i|Z_i).$$



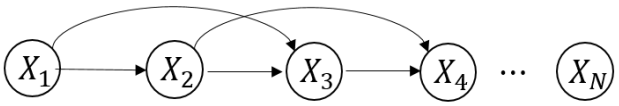
(a) Graphicial model (a).



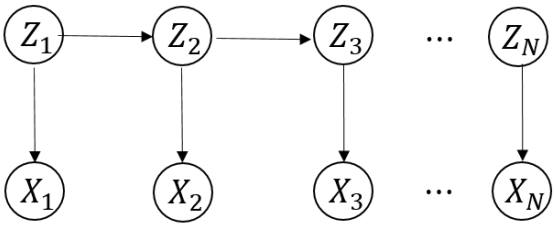
(b) Graphicial model (b). μ, σ are model parameters.



(c) Graphical model (c) for a Markov chain model.



(d) Graphical model (d) for a second-order Markov chain model.



(e) Graphical model (e) for a hidden Markov chain model.

Figure 15.6.2: Graphical model examples.

Example 15.6.2 (efficient inference via factorization). Consider a joint distribution on random variables A, B, C, D, E has the following factorization given by

$$P(A, B, C, D, E) = P(A)P(B)P(C|A, B)P(D|B, C)P(E|C, D).$$

To calculate

$$P(A|C = c) = \frac{P(A, C = c)}{P(C = c)},$$

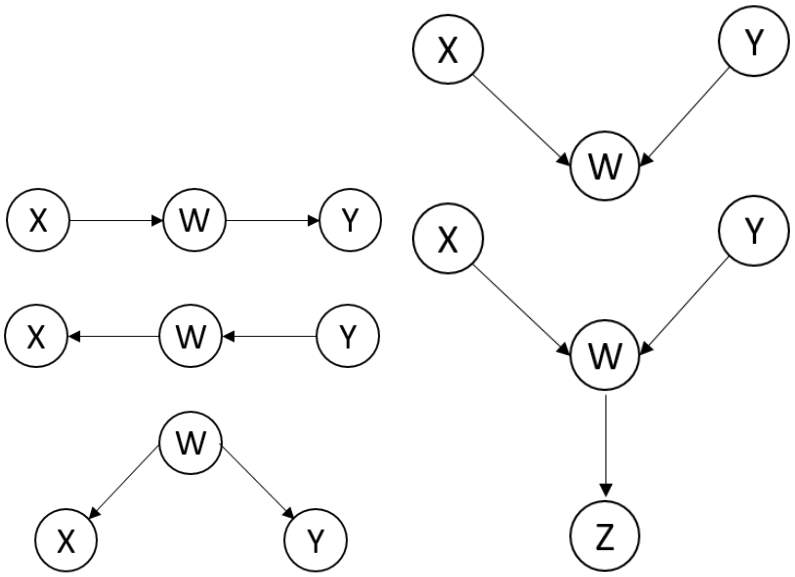
we have

$$\begin{aligned} P(A, C=c) &= \sum_{B,D,E} P(A)P(B)P(C=c|A,B)P(D|B,C=c)P(E|C=c,D) \\ &= \sum_{B} P(A)P(B)P(C=c|A,B) \sum_{D} P(D|B,C=c) \sum_{E} P(E|C=c,D) \\ &= \sum_{B} P(A)P(B)P(C=c|A,B) \end{aligned}$$

We now study how a graph model encodes conditional independence relationship. An important concept, d-separation [Figure 15.6.3], is introduced as follows.

Definition 15.6.2 (d-separation). *[18, p. 324] In a directed acyclic graphic model, a set of nodes $\mathcal{V}$ d-separates X from Y if **every undirected path** between X and Y is blocked by $\mathcal{V}$. A path is blocked by $\mathcal{V}$ if there is a node W on the path such that either*

- *W has converging arrows along the path $(\rightarrow W \leftarrow)$ and neither W nor its descendants are in $\mathcal{V}$.*
- *W does not have converging arrows along the path $(\rightarrow W \rightarrow)$ or $\leftarrow W \rightarrow$ and $W \in \mathcal{V}$.*



(a) X and Y are d-separated by W

(b) X and Y are not d-separated by W and Z.

Figure 15.6.3: d-separation examples.

Theorem 15.6.2 (d-sepration and conditional independence). *In a directed acyclic graphic model, X is conditional independent from Y given $\mathcal{V}$ if $\mathcal{V}$ d-separates X from Y.*

Remark 15.6.1 (correlation vs. causation)**.** The graphical modeling framework also help us distinguish correlation and causation

- causation will imply correlation.
- correlation will **not** imply causation.

For any two correlated events, A and B, the following relationships are possible:

- A causes B; (direct causation)
- B causes A; (reverse causation)
- A and B are **consequences of a common cause of C**, but do not cause each other; graphically, we have $A \leftarrow C \rightarrow B$;
- A causes B and B causes A (bidirectional or cyclic causation);
- A causes C which causes B (indirect causation); graphically, we have $A \rightarrow C \rightarrow B$;

15.7 Notes on Bibliography

For linear regression models, see [19][20]. For, linear models with *R* resources, see [21].

For multivariate statistical analysis, see [2][**anderson2009introduction**].

For copula, see [9][22][3][23].

BIBLIOGRAPHY

1. Hogg, R. V., McKean, J. & Craig, A. T. Introduction to Mathematical Statistics, 7 ed (2012).
2. Johnson, R. & Wichern, D. *Applied Multivariate Statistical Analysis* ISBN: 9780131877153 (Pearson Prentice Hall, 2007).
3. McNeil, A. J., Frey, R. & Embrechts, P. *Quantitative risk management: Concepts, techniques and tools* (Princeton university press, 2015).
4. Ma, Y. & Vidal, R. Generalized principal component analysis. *Unpublished Notes* (2002).
5. Tipping, M. E. & Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61,** 611–622 (1999).
6. Witten, D. M. & Tibshirani, R. J. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology* **8,** 1–27 (2009).
7. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10,** 515–534 (2009).
8. Suo, X., Minden, V., Nelson, B., Tibshirani, R. & Saunders, M. Sparse canonical correlation analysis. *arXiv preprint arXiv:1705.10865* (2017).
9. Rüschendorf, L. *Mathematical Risk Analysis: Dependence, Risk Bounds, Optimal Allocations and Portfolios* ISBN: 9783642335907 (Springer Berlin Heidelberg, 2013).
10. Schmitz, V. *Copulas and stochastic processes* (Bibliothek der RWTH Aachen, 2003).
11. Roncalli, T. *Lecture Notes on Risk Management & Financial Regulation* (2016).
12. Ruppert, D. *Statistics and data analysis for financial engineering, 2ed* (Springer, 2015).
13. DeFusco, R. A., McLeavey, D. W., Pinto, J. E., Anson, M. J. & Runkle, D. E. *Quantitative investment analysis* (John Wiley & Sons, 2015).
14. Fama, E. F. & French, K. R. The cross-section of expected stock returns. *the Journal of Finance* **47,** 427–465 (1992).
15. Fama, E. F. & French, K. R. Common risk factors in the returns on stocks and bonds. *Journal of financial economics* **33,** 3–56 (1993).

16. Henry, E., Robinson, T. R., Stowe, J. D., *et al. Equity asset valuation* (John Wiley & Sons, 2010).

17. Koller, D. & Friedman, N. *Probabilistic graphical models: principles and techniques* (MIT press, 2009).

18. Murphy, K. P. *Machine learning: a probabilistic perspective* (MIT press, 2012).

19. Kutner, M., Nachtsheim, C. & Neter, J. *Applied Linear Regression Models* ISBN: 9780072955675 (McGraw-Hill Higher Education, 2003).

20. Seber, G. A. & Lee, A. J. *Linear regression analysis* (John Wiley & Sons, 2012).

21. Faraway, J. J. *Linear models with R* (CRC press, 2014).

22. Lindskog, F. *et al. Modelling dependence with copulas and applications to risk management* ().

23. Cherubini, U., Luciano, E. & Vecchiato, W. *Copula methods in finance* (John Wiley & Sons, 2004).