... du hasard il n'est point de science:
S'il en était, on aurait tort
De l'appeler hasard...

La Fontaine

Contents

Foreword

These lecture notes are designed to accompany the author's course in Probability Theory and Stochastic Processes at the Technical University of Cluj-Napoca.

The course itself has a particular character. The interest in Probability Theory is traditional at our University. In the spirit of this tradition, some years ago Sergiu Nedevschi offered me the opportunity to give a course to undergraduate students, and to his collaborators and Ph.D. students. As an active participant, he suggested some topics. Other topics were suggested in different years by several enthusiastic Ph.D. students interested in specific applications. A few topics are the result of my whim. This explains the eclectic content and the application-oriented approach.

Needless to say, I enjoyed the experience. In time I prepared a handwritten version of the lectures. Voica Baraian talked me into systematizing the material, and afterwards she converted the handwritten version into a computer-edited one.

In preparing this version of the notes I removed some errors, but there is a high probability that other errors remain. Read with care, and - of course - use also other books.

Ioan Rasa
Ioan.Rasa@math.utcluj.ro

CHAPTER 1

Probability, Entropy, Information

1.1 Probability spaces

Let Ω be a set, $\mathscr{P}(\Omega)$ the family of all subsets of Ω, and $\mathscr{F} \subset \mathscr{P}(\Omega)$. For $A \in \mathscr{P}(\Omega)$ we denote by $\overline{A}$ the complement of A with respect to Ω.

Definition 1.1.1 *$\mathscr{F}$ is called a σ-**field** if*

a) $\Omega \in \mathscr{F}$;

b) $\overline{A} \in \mathscr{F}$ *for all* $A \in \mathscr{F}$;

c) If I is a finite or a countable set and $A_i \in \mathscr{F}, i \in I$, then $\bigcup\limits_{i \in I} A_i \in \mathscr{F}$.

It is easy to verify that if $\mathscr{F}$ is a σ-field, then

1) $\emptyset \in \mathscr{F}$;

2) $\bigcap\limits_{i \in I} A_i \in \mathscr{F}$ for all finite or countable I and all $A_i \in \mathscr{F}$, $i \in I$;

3) $A - B = A \cap \overline{B} \in \mathscr{F}$ for all $A, B \in \mathscr{F}$.

To prove 2) it suffices to remark that $\bigcap\limits_{i \in I} A_i = \overline{\bigcup\limits_{i \in I} \overline{A_i}}$.

Definition 1.1.2 *Let $\mathscr{F} \subset \mathscr{P}(\Omega)$ be a σ-field and $P : \mathscr{F} \longrightarrow \mathbb{R}$ such that:*

a) $P(A) \geq 0$, *for all* $A \in \mathscr{F}$;

b) $P(\Omega) = 1$;

c) If I is finite or countable, and $A_i \in \mathscr{F}, i \in I$, are pairwise disjoint, then

$$P\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} P(A_i).$$

Then P is called a probability *and $(\Omega, \mathscr{F}, P)$ a* probability space.

Remark 1.1.3 If $(\Omega, \mathscr{F}, P)$ is a probability space, then

1) $P(\emptyset) = 0$
2) $P(\overline{A}) = 1 - P(A),\ A \in \mathscr{F}$
3) $P(A - B) = P(A) - P(A \cap B),\ A, B \in \mathscr{F}$
4) $P(A) \le P(B)$ for all $A, B \in \mathscr{F}$ with $A \subset B$
5) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
6) $P\left(\bigcup\limits_{i\in I} A_i\right) \le \sum_{i\in I} P(A_i)$ for all finite or countable I and all $A_i \in \mathscr{F},\ i \in I$.

Let $(\Omega, \mathscr{F}, P)$ be a probability space. The elements of $\mathscr{F}$ are called *events.* $\emptyset$ is the *impossible event* and Ω is the *certain event.* If A is an event, then $\overline{A}$ is called the *contrary event.* $P(A)$ is the *probability of the event* $A \in \mathscr{F}$.

☞ **Example 1.1.4** A die is rolled. Then $\Omega = \{1, \dots, 6\}$, $\mathscr{F} = \mathscr{P}(\Omega)$ and $P(A) = \dfrac{1}{6}\, card(A),\ A \in \mathscr{F}$.

☞ **Example 1.1.5** Consider an experiment with possible outcomes forming the set $\Omega = \{r_1, \dots, r_n\}$. Let p_i be the probability of getting the outcome $r_i,\ i = 1, \dots, n$. Then $p_i \ge 0,\ i = 1, \dots, n$, and $p_1 + \cdots + p_n = 1$. Let $A = \{r_{k_1}, \dots, r_{k_j}\} \in \mathscr{F} = \mathscr{P}(\Omega)$. Then $P(A) = p_{k_1} + \cdots + p_{k_j}$.

☞ **Example 1.1.6** Let the outcomes of a random experiment form the set $\Omega = \{0, 1, 2, \dots\}$ and $\mathscr{F} = \mathscr{P}(\Omega)$. Let p_i be the probability of the event $\{i\}$; then $p_i \ge 0,\ i = 0, 1, \dots$, and $\sum\limits_{i=0}^{\infty} p_i = 1$. Moreover, $P(A) = \sum\limits_{i\in A} p_i,\ A \in \mathscr{F}$.

☞ **Example 1.1.7** Let $\Omega = [0, 1] \times [0, 1]$ and take $\mathscr{F}$ as the family of all the subsets of Ω having area. Then $\mathscr{F}$ is strictly contained in $\mathscr{P}(\Omega)$. Define $P(A) = \text{area}(A),\ A \in \mathscr{F}$. Thus $(\Omega, \mathscr{F}, P)$ is a probability space.

Remark 1.1.8 With notation from Example 1.1.5, consider the function

$$H(p_1, \dots, p_n) = -\sum_{i=1}^{n} p_i \log p_i.$$

The function H is called the *entropy* of the probability space. It can be considered as a measure of the uncertainty associated with the random experiment before performing the experiment; consequently, it can be also considered as measuring the quantity of *information* obtained after performing the experiment.

These considerations are justified by the following properties of H:

a) $H \geq 0$;

b) $H(1, 0, \dots, 0) = 0$;

c) $H(p_1, \dots, p_n, 0) = H(p_1, \dots, p_n)$;

d) $H(p_1, \dots, p_n) \leq H(\frac{1}{n}, \dots, \frac{1}{n})$.

The first three properties are obvious. To prove the fourth, remark that the function $f(x) = -x \log x$ is concave ($f''(x) = -\frac{1}{x} < 0$); according to Jensen's inequality,

$$\frac{f(p_1) + \cdots + f(p_n)}{n} \leq f\left(\frac{p_1 + \cdots + p_n}{n}\right) = f\left(\frac{1}{n}\right),$$

which leads to

$$H(p_1, \dots, p_n) = f(p_1) + \cdots + f(p_n) \leq nf\left(\frac{1}{n}\right) = -\log\frac{1}{n} = H\left(\frac{1}{n}, \dots, \frac{1}{n}\right).$$

1.2 Conditional probability. Independence

Consider a random experiment with n outcomes, each of them having probability $\frac{1}{n}$. Let A and B be events associated to the experiment, with

$$P(A) = \frac{m}{n},\ P(B) = \frac{p}{n},\ P(A \cap B) = \frac{q}{n}.$$

We are interested in knowing the probability of B subject to the condition that A is certain to occur; this is called the *conditional probability* of B, given that A has occurred, and is denoted by $P_A(B)$ or $P(B|A)$.

If A is certain to occur, we have to consider only m equally likely outcomes; the occurrence of B is guaranteed by q of them. It follows that

$$P(B|A) = \frac{q}{m} = \frac{q}{n} / \frac{m}{n} = P(A \cap B)/P(A).$$

So the conditional probability is given by

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

provided that $P(A) \neq 0$. (If $P(A) = 0$, the conditional probability is undefined).

Suppose now that a die is rolled and consider the events $A = \{1, 2\}$, $B = \{2, 3, 4\}$. It is easy to verify that

$$P(A) = P(A|B) = P(A|\overline{B}) = \frac{1}{3},$$
$$P(B) = P(B|A) = P(B|\overline{A}) = \frac{1}{2}.$$

That is, the occurrence of B or of $\overline{B}$ does not change the probability of A, and the occurrence of A or of $\overline{A}$ does not change the probability of B. In such a situation it is natural to say that A and B are independent, and we are lead to the following definition;

Definition 1.2.1 *The events A and B are said to be* stochastically independent *(or just* independent*) if*

1. $P(B|A) = P(B)$,
2. $P(B|\overline{A}) = P(B)$,
3. $P(A|B) = P(A)$,
4. $P(A|\overline{B}) = P(A)$.
   In fact, these four relations are equivalent, and each of them is equivalent to
5. $P(A \cap B) = P(A) \cdot P(B)$.

Indeed,

$$P(B|A) = P(B) \Longleftrightarrow \frac{P(A \cap B)}{P(A)} = P(B) \Longleftrightarrow P(A \cap B) = P(A) \cdot P(B),$$

which means that (1) $\Longleftrightarrow$ (5).
Similarly,

$P(B|\overline{A}) = P(B) \Longleftrightarrow \dfrac{P(\overline{A} \cap B)}{P(\overline{A})} = P(B) \Longleftrightarrow P(\overline{A} \cap B) = P(\overline{A}) \cdot P(B) \Longleftrightarrow$
$P(B - A) = (1 - P(A)) \cdot P(B) \Longleftrightarrow P(B) - P(A \cap B) = P(B) - P(A) \cdot P(B)$,
i.e., (2) $\Longleftrightarrow$ (5).
One proves analogously that (3) $\Longleftrightarrow$ (5) and (4) $\Longleftrightarrow$ (5).

We conclude that *A and B are independent events if and only if* $P(A \cap B) = P(A) \cdot P(B)$.

Similarly, A, B, C are independent if and only if $P(A \cap B) = P(A)P(B)$, $P(A \cap C) = P(A)P(C)$, $P(B \cap C) = P(B)P(C)$, $P(A \cap B \cap C) =$
$= P(A)P(B)P(C)$. This extends to an arbitrary number of events.

From a practical point of view, some experiments E_1 and E_2 are independent in an intuitive sense (e.g., "spin a coin" and "throw a dice").
An event associated with E_1 and one associated with E_2 will be independent events.

1.3 The Poisson model

Let us record the moments of time at which incoming calls reach a telephone switchboard. In a given interval of time $[0, t]$ it is possible to have $0, 1, 2, \ldots$ calls; we are interested in knowing the probability of having a certain number n of calls. More generally, consider events E (e.g., the above calls) which

- occur *singly* along the axis $[0, +\infty)$, i.e., the probability of two events occurring simultaneously is zero;

- occur *uniformly*, i.e., the number of occurrences in an interval of length t does not depend on the position of the interval;

- occur *independently.*

Moreover, we suppose that

(a) P (just one occurrence in the interval $(t, t + \delta t)) = \lambda\delta t + o(\delta t)$, for all $t \geq 0$; here $\lambda > 0$ is a constant and $o(\delta t)$ is Landau's symbol characterized by
$\lim\limits_{\delta t \to 0} \dfrac{o(\delta t)}{\delta t} = 0.$

(b) P (two or more occurrences in $(t, t + \delta t)) = o(\delta t)$, for all $t \geq 0$.

For given $t > 0$ and $n = 0, 1, 2, \ldots$ denote by $P_n(t)$ the probability of having exactly n occurrences of E in the interval $(0, t)$ (and hence in each time interval of length t).

For each $r = 0, 1, 2\ldots$ consider the events
A_r : r occurrences of E in $(t, t+\delta t)$
B_r : r occurrences of E in $(0, t)$
C_r : r occurrences of E in $(0, t+\delta t)$.
Then we have

$$P_n(t+\delta t) = P(C_n) = P[(A_0 \cap B_n) \cup (A_1 \cap B_{n-1}) \cup \cdots \cup (A_n \cap B_0)] =$$

$$P(A_0 \cap B_n) + P(A_1 \cap B_{n-1}) + \sum_{r=2}^{n} P(A_r \cap B_{n-r}) =$$

$$P(A_0) \cdot P(B_n) + P(A_1) \cdot P(B_{n-1}) + \sum_{r=2}^{n} P(A_r) \cdot P(B_{n-r}).$$

On the other hand,

$$P(A_1) = \lambda \delta t + o(\delta t) \text{ and } P(A_0) + P(A_1) +$$

$$+P[\text{ two or more occurrences in}(t, t+\delta t)] = 1,$$

which entails, according to (b),

$$P(A_0) = 1 - \lambda \delta t - o(\delta t).$$

Now we obtain

$$P_n(t+\delta t) = (1-\lambda \delta t) P_n(t) + (\lambda \delta t + o(\delta t)) P_{n-1}(t) + o(\delta t).$$

This implies

$$\frac{P_n(t+\delta t) - P_n(t)}{\delta t} = -\lambda P_n(t) + \lambda P_{n-1}(t) + \frac{o(\delta t)}{\delta t}\ .$$

Letting $\delta t \longrightarrow 0$ we get, for $n \geq 1$,

$$P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t). \tag{1.3.1}$$

Similarly, for $n = 0$ one obtains

$$P_0'(t) = -\lambda P_0(t). \tag{1.3.2}$$

Let us remark that $P_0(0) = 1$ and $P_n(0) = 0,\ n \geq 1$.
The Cauchy problem

$$\begin{cases} P_0'(t) = -\lambda P_0(t) \\ P_0(0) = 1 \end{cases}$$

has the solution $P_0(t) = e^{-\lambda t}$. Now (1.3.1) yields

$$\begin{cases} P_1'(t) = -\lambda P_1(t) + \lambda e^{-\lambda t} \\ P_1(0) = 0 \end{cases}$$

with solution $P_1(t) = \lambda t e^{-\lambda t}$.
Furthermore

$$\begin{cases} P_2'(t) = -\lambda P_2(t) + \lambda^2 t e^{-\lambda t} \\ P_2(0) = 0 \end{cases}$$

and consequently $P_2(t) = \frac{(\lambda t)^2}{2!} e^{-\lambda t}$. By induction we find

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad n \geq 0.$$

This is the Poisson model; as illustrated at the beginning of this section, it has many practical applications.

☞ **Example 1.3.1** After a long period of records it is known that approximatively 40 calls reach a certain switchboard in 2000 seconds. This means that $\lambda = \dfrac{40}{2000} = 0.02$. Suppose we are interested in the number of calls corresponding to an interval of 40 seconds. Then

$$P_n(40) = \frac{(0.8)^n}{n!} e^{-0.8}, \quad n \geq 0.$$

So the probabilities to have $0; 1; 2; 3; 4$ calls are, respectively, $0.45; 0.36; 0.14; 0.04; 0.01$.

1.4 Bayes' formula

Consider a probability space $(\Omega, \mathscr{F}, P)$ and n mutually exclusive events $A_1, \dots, A_n \in \mathscr{F}$ such that $A_1 \cup \dots \cup A_n = \Omega$. Let $X \in \mathscr{F}$. We have

$$X = X \cap \Omega = X \cap (A_1 \cup \dots \cup A_n) = (X \cap A_1) \cup \dots \cup (X \cap A_n)$$

and, since $X \cap A_1, \dots, X \cap A_n$ are mutually exclusive,

$$P(X) = P(X \cap A_1) + \dots + P(X \cap A_n).$$

On the other hand,

$$P(X \cap A_j) = P(A_j) P(X|A_j), \quad j = 1, \dots, n.$$

It follows that

$$P(X) = P(A_1)P(X|A_1) + \cdots + P(A_n)P(X|A_n).$$

Now let us remark that

$$P(X \cap A_j) = P(A_j)P(X|A_j) = P(X)P(A_j|X).$$

This implies

$$P(A_j|X) = \frac{P(A_j)P(X|A_j)}{P(X)}, \quad j = 1, \ldots, n.$$

Finally we obtain Bayes' formula:

$$P(A_j|X) = \frac{P(A_j)P(X|A_j)}{P(A_1)P(X|A_1) + \cdots + P(A_n)P(X|A_n)}, \ j = 1, \ldots, n.$$

If $P(A_1) = \cdots = P(A_n)$, it becomes

$$P(A_j|X) = \frac{P(X|A_j)}{P(X|A_1) + \cdots + P(X|A_n)}, \ j = 1, \ldots, n.$$

1.5 Problems

1. Show that the probability that at least one of the events A, B, C will occur is given by

   $$P(A)+P(B)+P(C)-P(A\cap B)-P(B\cap C)-P(C\cap A)+P(A\cap B\cap C).$$

   Extend this result to an arbitrary number of events.

2. Three players, A, B, C, toss a coin in this order. The first one to throw a head wins. Find their respective chances of winning.

3. A box contains two ordinary coins and one coin that has two tails. A coin is selected at random from the box and tossed twice. Find the probability of getting two tails. If two tails are obtained, find the probability that the selected coin was the dishonest one.

4. Two dice are rolled. Find the probability of obtaining a sum greater than 9, given that at least a face is a 6.

5. A lottery sells n tickets and gives 4 prizes. You have 3 tickets. What is the probability that you will win at least one prize?

CHAPTER 2

Random variables

2.1 Definitions and examples

Let $(\Omega, \mathscr{F}, P)$ be a probability space.

Definition 2.1.1 *A function $X : \Omega \longrightarrow \mathbb{R}$ is called a **random variable** if*

$$\{\omega \in \Omega \ : \ X(\omega) < a\} \in \mathscr{F}, \quad a \in \mathbb{R}. \tag{2.1.1}$$

For the sake of simplicity, the set $\{\omega \in \Omega \ : \ X(\omega) < a\}$ is denoted by $\{X < a\}$.

It can be proved that condition (2.1.1) is equivalent to each of the following conditions:

$$\{X \leq a\} \in \mathscr{F}, \ a \in \mathbb{R}; \tag{2.1.2}$$

$$\{X > a\} \in \mathscr{F}, \ a \in \mathbb{R}; \tag{2.1.3}$$

$$\{X \geq a\} \in \mathscr{F}, \ a \in \mathbb{R}. \tag{2.1.4}$$

Clearly, if $\mathscr{F} = \mathscr{P}(\Omega)$ any function $X : \Omega \longrightarrow \mathbb{R}$ is a random variable.

Let $A \in \mathscr{F}$ and $X_A : \Omega \longrightarrow \mathbb{R}$, $X_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \in \overline{A}. \end{cases}$

X_A, the characteristic function of A, is a random variable; indeed,

$$\{X_A < a\} = \begin{cases} \emptyset & , a \leq 0; \\ \overline{A} & , a \in (0, 1]; \\ \Omega & , a > 1. \end{cases}$$

It can be proved that if $X, Y : \Omega \longrightarrow \mathbb{R}$ are random variables, then $X + Y$, XY, X/Y (provided that $Y \neq 0), |X|$, and cX $(c \in \mathbb{R})$ are random variables.
Concerning a random variable X, interest usually centers on determining the probability that X will assume specified values. For example, if X represents the sum of the points obtained in rolling two dice, then it could be of interest to calculate the probability that X will assume the value 8 (this probability is, of course, $\frac{5}{36}$).

If X represents the distance a dart will land from the center of a target, it may be of interest to know the probability that X will assume some value less than 10.

So, it is useful to introduce the following concept.

Definition 2.1.2 *Let $X : \Omega \longrightarrow \mathbb{R}$ be a random variable. The function $F : \mathbb{R} \longrightarrow \mathbb{R}$, $F(x) = P(X < x)$, is called the* distribution function *of X.*

It is easy to verify that F is non-decreasing, $0 \leq F(x) \leq 1,\ x \in \mathbb{R}$, and $\lim\limits_{x \to -\infty} F(x) = 0,\ \lim\limits_{x \to \infty} F(x) = 1.$
Moreover, $P(a \leq X < b) = F(b) - F(a),\ a, b \in \mathbb{R}, a < b.$

2.2 Discrete random variables

A random variable X is called *discrete* if it can assume only a finite or a countable number of values. Let $\{x_i : i \in I\}$ be these values, where I is a finite or a countable set of indices.
Denote $p_i = P(X = x_i),\ i \in I.$ Then $p_i \geq 0,\ i \in I$, and $\sum\limits_{i \in I} p_i = 1.$ X can be described by its *repartition* (or *distribution*):

$$X = \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}.$$

☞ **Example 2.2.1** Consider a random experiment in which only the occurrence or nonoccurrence of a given event E is recorded. Let p be the probability that E occurs when the experiment is performed, and $q = 1 - p$. Let n independent trials be made and denote by X the number of occurrences of E.
Then X is a random variable:

$$X \begin{pmatrix} i \\ \binom{n}{i} p^i q^{n-i} \end{pmatrix}_{i=0,1,\ldots,n}.$$

This is the *binomial distribution* with parameters n and p.

☞ **Example 2.2.2** Consider the Poisson model described in Section 1.3. Let X be the number of occurrences of E in the interval $(0,1)$. Then X is a random variable:

$$X\begin{pmatrix} n \\ \frac{\lambda^n}{n!}e^{-\lambda} \end{pmatrix}_{n=0,1,2,\ldots}.$$

This is the *Poisson distribution* with parameter $\lambda > 0$.

It is easy to see that the distribution function of a discrete random variable $X\binom{x_i}{p_i}_{i\in I}$ is a step-function with steps of height p_i at $x_i,\ i \in I$.

2.3 Continuous random variables

Definition 2.3.1 *A function* $f : \mathbb{R} \longrightarrow \mathbb{R}$ *is called a* probability density function *if*

a) $f(t) \geq 0,\ t \in \mathbb{R}$;

b) f *is integrable on* $\mathbb{R}$*;*

c) $\int\limits_{-\infty}^{\infty} f(t)dt = 1.$

Definition 2.3.2 *Let X be a random variable and F its distribution function. X is called a* ***continuous random variable*** *if a probability density function exists such that*

$$F(x) = \int\limits_{-\infty}^{x} f(t)dt,\ x \in \mathbb{R}.$$

If X is continuous with density f, then for $a, b \in \mathbb{R},\ a < b$ we have

$$P(a \leq X < b) = F(b) - F(a) = \int_{-\infty}^{b} f(t)dt - \int_{-\infty}^{a} f(t)dt = \int_{a}^{b} f(t)dt.$$

Moreover, $P(X = c) = 0,\ c \in \mathbb{R}$, and so if I is one of the intervals $[a,b]$, $[a,b)$, $(a,b]$ or (a,b), then

$$P(X \in I) = \int_{a}^{b} f(t)dt.$$

The term "probability density" is justified by

$$P(x \leq X \leq x + \Delta x) = \int_{x}^{x+\Delta x} f(t)dt \approx f(x)\Delta x$$

which can be interpreted as

$$f(x) \approx \frac{P(x \leq X \leq x + \Delta x)}{\Delta x}.$$

If f is continuous at $x \in \mathbb{R}$, then $F'(x) = f(x)$.

☞ **Example 2.3.3** Let $a, b \in \mathbb{R},\ a < b$. The function

$$f(x) = \begin{cases} \dfrac{1}{b-a} & , \ x \in [a, b] \\ 0 & , \ \text{elsewhere} \end{cases}$$

is a probability density function. We say that a random variable X with density f has a *uniform distribution*. In this case the function F is

$$F(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0 & , \ x < a \\ \dfrac{x-a}{b-a} & , \ a \leq x \leq b \\ 1 & , \ x > b. \end{cases}$$

☞ **Example 2.3.4** The most useful of all distributions for continuous random variables is a distribution called the *normal* or *Gaussian* distribution. It serves as a model distribution for many real-life continuous random variables and arises naturally in many theoretical researches. The density of a normal variable is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \quad x \in \mathbb{R},$$

where $m \in \mathbb{R}$ and $\sigma > 0$ are real parameters. If X has the above density f, we say that X is a $N(m, \sigma^2)$ variable.

Suppose that X is $N(0, 1)$, thus having the density

$$f(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right), \quad t \in \mathbb{R}.$$

Then

$$F(x) = \int_{-\infty}^{x} f(t)dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}dt =$$
$$\frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{0} e^{-t^2/2}dt + \int_{0}^{x} e^{-t^2/2}dt\right).$$

We have $\int_{-\infty}^{0} e^{-t^2/2}dt = \sqrt{\frac{\pi}{2}}$. Denote

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2}dt, \quad x \in \mathbb{R}.$$

So we obtain

$$F(x) = \frac{1}{2} + \Phi(x), \quad x \in \mathbb{R}.$$

Φ is called *Laplace's integral function* and its values can be found in tables. It is not difficult to prove that $\Phi(-x) = -\Phi(x)$, so that it suffices to know the values of Φ for $x > 0$.

Here are some particular (approximative) values of Φ:

x	1	2	3	4
$\Phi(x)$	0.3413	0.4772	0.4987	0.4999

For a given $x > 0$ we have

$$P(-x \le X \le x) = \int\limits_{-x}^{x} f(t)dt = F(x) - F(-x) = \frac{1}{2} + \Phi(x) - \frac{1}{2} - \Phi(-x) =$$
$$= 2\Phi(x).$$

In particular, $P(-1 \le X \le 1) = 0.6826$;
$P(-2 \le X \le 2) = 0.9544$;
$P(-3 \le X \le 3) = 0.9974$.
So, if X is $N(0,1)$, then with probability 0.9974 X assumes values between -3 and 3.

Now suppose that X is a $N(m, \sigma^2)$ random variable. The distribution function will be

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} \exp(-\frac{(t-m)^2}{2\sigma^2})dt = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{0} e^{-y^2/2}dy +$$
$$\frac{1}{\sqrt{2\pi}} \int\limits_{0}^{\frac{x-m}{\sigma}} e^{-y^2/2}dy = \frac{1}{2} + \Phi\left(\frac{x-m}{\sigma}\right).$$

In this case we have
$P(m-3\sigma \leq X \leq m+3\sigma) = F(m+3\sigma) - F(m-3\sigma) =$
$= \Phi(3) - \Phi(-3) = 0.9974.$

So, a $N(m,\sigma^2)$ variable X assumes values between $m-3\sigma$ and $m+3\sigma$ with a probability very close to 1.

☞ **Example 2.3.5** The *Cauchy distribution* is given by the density

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2} \quad x \in \mathbb{R}.$$

☞ **Example 2.3.6** Let $\alpha > 0$, $\beta > 0$. *The Gamma distribution* is given by the probability density function

$$f(x) = \begin{cases} \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)} & , \ x > 0 \\ 0 & , \ x \leq 0 \end{cases}$$

Here $\Gamma(\alpha) = \int\limits_0^{\infty} x^{\alpha-1}e^{-x}dx$ is the Gamma function for which $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$; in particular, if n is a positive integer, $\Gamma(n+1) = n!$.

☞ **Example 2.3.7** The special case of the Gamma distribution that occurs when $\alpha = 1$ is called the *exponential distribution.* It is used sufficiently often to justify listing it separately. The density is in this case:

$$f(x) = \begin{cases} \frac{1}{\beta}e^{-x/\beta} & , \ x > 0 \\ 0 & , \ x \leq 0. \end{cases}$$

This distribution arises in studying the lifetime of a radioactive material. From Physics it is known that the rate at which a mass y of radioactive material is decaying is proportional to the amount of material remaining at any time t. Thus y satisfies the differential equation

$$\frac{dy}{dt} = -\lambda y$$

where $\lambda > 0$ is a constant depending upon the kind of material being studied.

Let y_0 denote the amount of radioactive material at time $t = 0$. The solution of the above differential equation is

$$y(t) = y_0 e^{-\lambda t}.$$

Since $\dfrac{y_0 - y(t)}{y_0}$ represents the proportion of the original material that has decayed in t units of time, this quantity may be taken as the probability that an atom selected at random from this material will decay in t units of time. Thus, if X is the length of life of such an atom, then

$$F(t) = P(X < t) = \frac{y_0 - y(t)}{y_0} = 1 - e^{-\lambda t}, \quad t > 0.$$

This means that the distribution function of X is

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & , \ x > 0 \\ 0 & , \ x \leq 0. \end{cases}$$

Consequently, X has the density

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & , \ x > 0 \\ 0 & , \ x \leq 0, \end{cases}$$

which is an exponential density with parameter $\beta = 1/\lambda$.

☞ **Example 2.3.8** *The Chi-Square* (χ^2) *distribution* is characterized by the density

$$f(x) = \begin{cases} \dfrac{x^{\nu/2-1} e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)} & , \ x > 0 \\ 0 & , \ x \leq 0. \end{cases}$$

This is a particular case of the Gamma distribution, obtained for $\alpha = \nu/2$ and $\beta = 2$.

☞ **Example 2.3.9** The *Student distribution* is given by the probability density function

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} , \quad x \in \mathbb{R}.$$

The particular case when $n = 1$ is the Cauchy distribution.

2.4 Multidimensional random variables

Let X and Y be random variables defined on Ω. Then the pair (X, Y) can be viewed as a *bidimensional random variable* $(X, Y) : \Omega \longrightarrow \mathbb{R}^2$. The distribution function of (X, Y) is $F : \mathbb{R}^2 \longrightarrow \mathbb{R}$,

$$F(x, y) = P(X < x, Y < y), \ (x, y) \in \mathbb{R}^2;$$

here $(X < x, Y < y)$ means $(X < x$ and $Y < y)$.

Let $F_1, F_2 : \mathbb{R} \longrightarrow \mathbb{R}$ be the distribution functions of X, respectively Y; it can be proved that

$$\lim_{y\to\infty} F(x,y) = F_1(x) \text{ and } \lim_{x\to\infty} F(x,y) = F_2(y).$$

A function $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ is called a *probability density function* on $\mathbb{R}^2$ if

a) $f(x,y) \geq 0,\ (x,y) \in \mathbb{R}^2$;

b) f is integrable on $\mathbb{R}^2$;

c) $\iint_{\mathbb{R}^2} f(x,y)dxdy = 1$.

We say that (X,Y) is a continuous random variable if there exists a density f on $\mathbb{R}^2$ such that

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(s,t)dsdt,\ (x,y) \in \mathbb{R}^2.$$

In this case

$$P((X,Y) \in D) = \iint_D f(s,t)dsdt$$

for each domain $D \subset \mathbb{R}^2$.

Let $f_1(x) = \int_{\mathbb{R}} f(x,y)dy$ and $f_2(y) = \int_{\mathbb{R}} f(x,y)dx$; f_1 and f_2 are called *marginal density functions.* It can be proved that f_1 is the density of X and f_2 that of Y.

Let x be fixed, with $f_1(x) > 0$. The function

$$y \longrightarrow f(y|\ x) := \frac{f(x,y)}{f_1(x)}$$

is called the *conditional probability density* of Y when $X = x$ is fixed. Obviously we have

$$\int_{\mathbb{R}} f(y|\ x)dy = \int_{\mathbb{R}} \frac{f(x,y)}{f_1(x)}dy = \frac{1}{f_1(x)}\int_{\mathbb{R}} f(x,y)dy = 1.$$

Moreover,

$$P(Y < b|\ X = x) = \int_{-\infty}^{b} f(y|\ x)dy.$$

The conditional density function of X for Y fixed is defined in a similar manner.

All these considerations apply to multidimensional random variables $(X_1, X_2, \ldots, X_n) : \Omega \longrightarrow \mathbb{R}^n$.

2.5 Independent random variables

Let X_1, X_2 be random variables on Ω. Let F_1, F_2 be their distribution functions and F the distribution function of the bidimensional random variable (X_1, X_2). Thus we have

$$F_1(x_1) = P(X_1 < x_1),\ F_2(x_2) = P(X_2 < x_2),$$
$$F(x_1, x_2) = P(X_1 < x_1,\ X_2 < x_2).$$

Definition 2.5.1 *We say that X_1 and X_2 are independent random variables if the events $\{X_1 < x_1\}$ and $\{X_2 < x_2\}$ are independent for all $x_1, x_2 \in \mathbb{R}$.*

In this case we have $F(x_1, x_2) = F(x_1)F(x_2),\ x_1, x_2 \in \mathbb{R}$.

Moreover, if $X_1, X_2, (X_1, X_2)$ are continuous with densities f_1, f_2, f, then

$$f(x_1, x_2) = f_1(x_1)f_2(x_2), \quad x_1, x_2 \in \mathbb{R}.$$

Similar definitions and results can be formulated for n random variables $X_1, X_2, \ldots, X_n$.

2.6 Problems

1. Let X be a random variable with density

$$f(x) = \begin{cases} cx^2e^{-x} & , \ x > 0 \\ 0 & , \ x \le 0. \end{cases}$$

Find c; $P(X < 1)$; $P(1 < X < 2)$; $P(X > 2)$.

2. Assume that (X, Y) possesses a uniform distribution over the square with vertices at $(1, 1),\ (-1, 1),\ (1, -1),\ (-1, -1)$. Determine the probabilities of the following events:

   (i) $X + Y \ge -1$.

   (ii) $X^2 + Y^2 < \dfrac{1}{4}$.

   (iii) $|X - Y| < 1$.

3. Let (X, Y) have the density

$$f(x, y) = \begin{cases} cye^{-y(x+1)} & , \ x, y > 0 \\ 0 & , \ \text{otherwise}. \end{cases}$$

Find c, the marginal density functions, and the conditional density functions.

4. Assume that (X, Y) has the density

$$f(x,y) = \begin{cases} e^{-(x+y)} & , \ x > 0, y > 0 \\ 0 & , \ \text{otherwise.} \end{cases}$$

Find $P(X < 1),\ P(X < 1 \mid Y = 1),\ P(X > Y \mid Y < 1)$.

CHAPTER 3

Operations with random variables

3.1 Operations with discrete variables

Consider the discrete random variables

$$X\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I} \quad , \quad Y\begin{pmatrix} y_j \\ q_j \end{pmatrix}_{j\in J}.$$

Denote $p_{ij} = P(X = x_i,\ Y = y_j)$. The sum of X and Y is the random variable

$$X+Y\begin{pmatrix} x_i + y_j \\ p_{ij} \end{pmatrix}_{(i,j)\in I\times J}.$$

If X and Y are independent, then

$$p_{ij} = P(X = x_i,\ Y = y_j) = P(X = x_i)P(Y = y_j) = p_i q_j,$$

hence

$$X+Y\begin{pmatrix} x_i + y_j \\ p_i q_j \end{pmatrix}_{(i,j)\in I\times J}.$$

If X and Y are independent, then

$$XY\begin{pmatrix} x_i y_j \\ p_i q_j \end{pmatrix}_{(i,j)\in I\times J}.$$

☞ **Example 3.1.1** Let X and Y be Poisson random variables with parameters λ, respectively μ. Suppose that X and Y are independent.

We have $P(X = i) = \frac{\lambda^i}{i!}e^{-\lambda}$, $P(Y = j) = \frac{\mu^j}{j!}e^{-\mu}$, $i, j = 0, 1, 2, \ldots$. For $k = 0, 1, 2, \ldots$ we have

$$\{X+Y = k\} = \{X = 0,\ Y = k\} \cup \{X = 1,\ Y = k-1\} \cup \cdots \cup \{X = k,\ Y = 0\}.$$

Consequently,

$$P(X + Y = k) = \sum_{i=0}^{k} P(X = i, Y = k - i) = \sum_{i=0}^{k} P(X = i)P(Y = k - i) =$$
$$= \sum_{i=0}^{k} \frac{\lambda^i}{i!}e^{-\lambda}\frac{\mu^{k-i}}{(k-i)!}e^{-\mu} = \frac{e^{-(\lambda+\mu)}}{k!}\sum_{i=0}^{k}\frac{k!}{i!(k-i)!}\lambda^i\mu^{k-i} =$$
$$= \frac{(\lambda+\mu)^k}{k!}e^{-(\lambda+\mu)}.$$

We conclude that $X + Y$ is a Poisson variable with parameter $\lambda + \mu$.

3.2 Operations with continuous variables

Let X and Y be continuous random variables with densities $f_1(x)$, respectively $f_2(y)$. Let $f(x, y)$ be the density of (X, Y). Denote $Z = X + Y$ and let $F(x)$ be the distribution function of Z. For a given $z \in \mathbb{R}$ we have

$$F(z) = P(Z < z) = P(X + Y < z) = P((X, Y) \in D),$$

where $D = \{(x, y) \in \mathbb{R}^2 : x + y < z\}$. Thus

$$F(z) = \iint_D f(x, y)dxdy.$$

Now let $x = u$, $x + y = v$; then $\dfrac{D(x, y)}{D(u, v)} = 1$ and D is transformed into $\Delta = \{(u, v) \in \mathbb{R}^2 : v < z\}$.

We have

$$F(z) = \iint_\Delta f(u, v - u)dudv = \int_{-\infty}^{z} dv \int_{-\infty}^{\infty} f(u, v - u)du.$$

The probability density of Z is

$$g(z) = F'(z) = \int_{-\infty}^{\infty} f(u, z - u)du.$$

If X and Y are independent, then $f(x,y)=f_1(x)f_2(y)$, and we have

$$g(z)=\int_{-\infty}^{\infty} f_1(u)f_2(z-u)du.$$

In the general case it can be proved that the densities of XY and $\dfrac{X}{Y}$ are, respectively,

$$v(z)=\int_{-\infty}^{\infty} f\left(u,\frac{z}{u}\right)\frac{1}{|u|}du,$$
$$w(z)=\int_{-\infty}^{\infty} f(uz,u)|u|du.$$

☞ **Example 3.2.1** Let X and Y be independent, uniformly distributed in $[a,b]$ random variables (see Ex. 2.3.3). This means that they have densities

$$f_1(x)=\begin{cases}\dfrac{1}{b-a} & , \; x\in[a,b],\\ 0 & , \; \text{otherwise;}\end{cases}$$

$$f_2(y)=\begin{cases}\dfrac{1}{b-a} & , \; y\in[a,b],\\ 0 & , \; \text{otherwise.}\end{cases}$$

The probability density of $X+Y$ is

$$g(z)=\int_{-\infty}^{\infty} f_1(u)f_2(z-u)du=\int_a^b \frac{1}{b-a}f_2(z-u)du.$$

Setting $z-u=t$ we have

$$g(z)=\frac{1}{b-a}\int_{z-b}^{z-a} f_2(t)dt.$$

If $z-a<a$ or $z-b>b$, then $g(z)=0$; we have to study the case when $2a\le z\le 2b$.

If $2a\le z\le a+b$, then $z-b\le a$ and $a\le z-a\le b$, so that

$$g(z)=\frac{1}{b-a}\left(\int_{z-b}^{a} f_2(t)dt+\int_a^{z-b} f_2(t)dt\right)=\frac{z-2a}{(b-a)^2}.$$

If $a+b\le z\le 2b$, we get similarly

$$g(z)=\frac{2b-z}{(b-a)^2}.$$

Consequently,

$$g(z)=\begin{cases}\dfrac{z-2a}{(b-a)^2} & ,\ 2a\le z\le a+b\\ \dfrac{2b-z}{(b-a)^2} & ,\ a+b\le z\le 2b\\ 0 & ,\ \text{otherwise.}\end{cases}$$

☞ **Example 3.2.2** Let $f(x,y)$ be the density of (X,Y). Consider a bijective mapping $(U,V)\longrightarrow(X,Y),\ X=\varphi(U,V),\ Y=\psi(U,V)$ and let $g(u,v)$ be the density of (U,V).
Let $D\subset\mathbb{R}^2$ be an arbitrary domain and

$$\Delta=\{(x,y)\in\mathbb{R}^2\ :\ x=\varphi(u,v),\ y=\psi(u,v),\ (u,v)\in D\}.$$

Then $P((U,V)\in D)=P((X,Y)\in\Delta)$, i.e.,

$$\iint_D g(u,v)dudv=\iint_\Delta f(x,y)dxdy=\iint_D f(\varphi(u,v),\psi(u,v)|J|dudv,$$

where

$$J=\frac{D(\varphi,\psi)}{D(u,v)}.$$

Since D is arbitrary, we conclude that

$$g(u,v)=f(\varphi(u,v),\psi(u,v)|J|.$$

3.3 Problems

1. The probability density of (X,Y) is

$$f(x,y)=\begin{cases}\frac{1}{a^2} & ,\ 0\le x\le a,\ 0\le y\le a\\ 0 & ,\ \text{otherwise.}\end{cases}$$

   Prove that $|X-Y|$ and $\min(X,Y)$ have the same distribution function.

2. Find the density of $1-X^3$, if X is a Cauchy random variable.

3. The density of (X,Y) is

$$f(x,y)=\begin{cases}e^{-(x+y)} & ,\ x,y\ge 0\\ 0 & ,\ \text{otherwise.}\end{cases}$$

   Let $U=X+Y,\ V=\frac{X}{Y}$. Find the densities of (U,V) and U.

4. Find the density of $\frac{X}{Y}$ if X and Y are independent, uniformly distributed in $[0,1]$ random variables.

5. X is uniformly distributed in $[-1,1]$. Find the densities of e^X, $2X+1$, $2X^2+1$.

6. Let X and Y be binomial variables with parameters (n,p), respectively (m,p). If X and Y are independent, find the distribution of $X+Y$.

7. X and Y are independent, $N(0,\sigma^2)$ random variables. Find the density of $\max\{|X|,\ |Y|\}$.

8. (X,Y) has the density $f(x,y)$. Find the densities of $\min(X,Y)$ and $\max(X,Y)$.

9. X is $N(0,1)$, Y is Chi-Square with parameter ν, and X,Y are independent. Prove that $\dfrac{X}{\sqrt{Y/\nu}}$ is a Student variable.

CHAPTER 4

Expectation and Variance

4.1 The Expectation

Let X be a discrete random variable. The *expectation* (or *mean value*, or *expected value*) of

$$X\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}$$

is defined by

$$E(X) = \sum_{i\in I} p_i x_i$$

provided the sum exists.

If X is continuous with probability density f, then its *expectation* is

$$E(X) = \int_{\mathbb{R}} x f(x) dx.$$

Let $h : \mathbb{R} \longrightarrow \mathbb{R}$ be continuous. Then

$$E(h(X)) = \sum_{i\in I} p_i h(x_i),$$

respectively

$$E(h(X)) = \int_{\mathbb{R}} h(x) f(x) dx.$$

Consider two discrete random variables:

$$X\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}, Y\begin{pmatrix} y_i \\ q_j \end{pmatrix}_{j\in J}.$$

Then $X+Y$ has the distribution

$$X+Y\begin{pmatrix} x_i+y_j \\ p_{ij} \end{pmatrix}_{(i,j)\in I\times J},$$

where

$$p_{ij}=P(X=x_i,\ Y=y_j).$$

This yields

$$E(X+Y)=\sum_{i\in I,j\in J}(x_i+y_j)p_{ij}=\sum_{i\in I}\sum_{j\in J}x_ip_{ij}+\sum_{j\in J}\sum_{i\in I}y_jp_{ij}=$$
$$\sum_{i\in I}x_i\sum_{j\in J}p_{ij}+\sum_{j\in J}y_j\sum_{i\in I}p_{ij}=\sum_{i\in I}x_ip_i+\sum_{j\in J}y_jq_j=$$
$$E(X)+E(Y).$$

Thus we have

$$E(X+Y)=E(X)+E(Y).$$

This is true also when X and Y are continuous.
Indeed, let $f_1(x), f_2(y), f(x,y)$ be the probability densities of X, Y, respectively (X,Y). The density of $X+Y$ is

$$g(z)=\int_{-\infty}^{+\infty}f(u,z-u)du.$$

Therefore,

$$E(X+Y)=\int_{-\infty}^{+\infty}zg(z)dz=\int_{-\infty}^{+\infty}zdz\int_{-\infty}^{+\infty}f(u,z-u)du=$$
$$\iint_{\mathbb{R}^2}zf(u,z-u)dudz.$$

Taking as new variables u and $v=z-u$, we get $\dfrac{D(u,z)}{D(u,v)}=1$, hence

$$\begin{aligned}E(X+Y)&=\iint_{\mathbb{R}^2}(u+v)f(u,v)dudv=\iint_{\mathbb{R}^2}uf(u,v)dudv+\iint_{\mathbb{R}^2}vf(u,v)dudv\\&=\int_{\mathbb{R}}udu\int_{\mathbb{R}}f(u,v)dv+\int_{\mathbb{R}}vdv\int_{\mathbb{R}}f(u,v)du=\\&=\int_{\mathbb{R}}uf_1(u)du+\int_{\mathbb{R}}vf_2(v)dv=E(X)+E(Y).\end{aligned}$$

Now suppose that X and Y are *independent* discrete random variables,

$$X\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}, Y\begin{pmatrix} y_j \\ q_j \end{pmatrix}_{j\in J}.$$

Then XY has the distribution

$$XY\begin{pmatrix} x_iy_j \\ p_iq_j \end{pmatrix}_{(i,j)\in I\times J},$$

which leads to

$$\begin{aligned} E(XY) &= \sum_{i\in I}\sum_{j\in J} x_iy_jp_iq_j = \sum_{i\in I} p_ix_i \sum_{j\in J} q_jy_j = \\ &= E(X)\cdot E(Y). \end{aligned}$$

Thus, for independent discrete variables X and Y we have

$$E(XY) = E(X)\cdot E(Y).$$

This is true also for independent continuous random variables X and Y. Indeed, let $f_1(x)$ and $f_2(y)$ be the densities of X and respectively Y.

Then the density of XY is

$$g(z) = \int_{-\infty}^{+\infty} f_1(u)f_2\left(\frac{z}{u}\right)\frac{1}{|u|}du,$$

and, consequently,

$$\begin{aligned} E(XY) &= \int_{\mathbb{R}} zg(z)dz = \int_{\mathbb{R}} zdz \int_{\mathbb{R}} f_1(u)f_2\left(\frac{z}{u}\right)\frac{1}{|u|}du = \\ &= \iint_{\mathbb{R}} zf_1(u)f_2\left(\frac{z}{u}\right)\frac{1}{|u|}dudz. \end{aligned}$$

Taking as new variables u and $v = \frac{z}{u}$ we get $\dfrac{D(u,z)}{D(u,v)} = u$, so that

$$\begin{aligned} E(XY) &= \iint_{\mathbb{R}^2} uvf_1(u)f_2(v)\frac{1}{|u|}|u|dudv = \\ &= \int_{\mathbb{R}} uf_1(u)du \int_{\mathbb{R}} vf_2(v)dv = E(X)\cdot E(Y). \end{aligned}$$

Remark 4.1.1 If X is a random variable and $a \in \mathbb{R}$, then $E(aX) = aE(X)$.

Remark 4.1.2 If $X \geq 0$ and $E(X) = 0$, then $P(X = 0) = 1$; in this case we say that $X = 0$ *almost surely.*

4.2 The Variance

Let X be a random variable with finite expectation $E(X)$. The variance of X is defined by

$$Var(X) = E((X - EX)^2).$$

Let us remark that

$$Var(X) = E(X^2 - 2(EX)X + (EX)^2) = E(X^2) - 2(EX)^2 + (EX)^2,$$

so that

$$Var(X) = E(X^2) - (EX)^2.$$

If $Var(X) = 0$, then X=const. almost surely.

Indeed, $E((X - EX)^2) = 0$ implies $(X - EX)^2 = 0$ almost surely, i.e., $X = EX$ almost surely.

The variance of X may serve as a measure of the degree to which X is concentrated about the expectation EX.

4.3 Schwarz' inequality

Theorem 4.3.1 *Let X and Y be random variables. Then*

$$(E(XY))^2 \leq E(X^2)E(Y^2).$$

The equality holds if and only if there exists $c \in \mathbb{R}$ such that $Y = cX$ almost surely.

Proof. From $(tX - Y)^2 \geq 0$ for all $t \in \mathbb{R}$ we deduce $t^2X^2 - 2tXY + Y^2 \geq 0$, $t \in \mathbb{R}$, and so

$$t^2E(X^2) - 2tE(XY) + E(Y^2) \geq 0, \quad t \in \mathbb{R}. \tag{4.3.1}$$

If $E(X^2) = 0$, then $2tE(XY) \leq E(Y^2), \quad t \in \mathbb{R}$, which implies $E(XY) = 0$ and Schwarz' inequality is verified. If $E(X^2) > 0$, then from (4.3.1) we infer

$$(E(XY))^2 - E(X^2)E(Y^2) \leq 0$$

which is exactly Schwarz' inequality.

Finally, if $(E(XY))^2 = E(X^2)E(Y^2)$, then the discriminant in (4.3.1) is zero, which means that there exists $c \in \mathbb{R}$ such that

$$c^2E(X^2) - 2cE(XY) + E(Y^2) = 0.$$

This entails $E((cX - Y)^2) = 0$, and so $Y = cX$ almost surely.

4.4 Problems

Find $E(X)$ and $Var(X)$ if X is

1. Binomial with parameters n and p.
2. Poisson with parameter λ.
3. Uniformly distributed in $[a, b]$.
4. Normal with parameters m and σ.
5. Gamma with parameters α and β.
6. Chi-Square with parameter ν.
7. Let X and Y be independent. Prove that
$$Var(XY) = (VarX)(VarY) + E^2(X)VarY + E^2(Y)VarX.$$
8. Let X be $N(m, \sigma^2)$. Find $E(|X - m|)$.

CHAPTER 5

Least-squares problems

5.1 The orthogonal projection

Let $(V, <, >)$ be a real inner-product space and $v_1, \ldots, v_n \in V$. The matrix

$$G(v_1, \ldots, v_n) = \begin{pmatrix} < v_1, v_1 > & < v_1, v_2 > & \ldots & < v_1, v_n > \\ < v_2, v_1 > & < v_2, v_2 > & \ldots & < v_2, v_n > \\ \ldots & \ldots & \ldots & \ldots \\ < v_n, v_1 > & < v_n, v_2 > & \ldots & < v_n, v_n > \end{pmatrix}$$

is called the *Gram matrix* of the vectors $v_1, \ldots, v_n$; $detG(v_1, \ldots, v_n)$ is called the *Gram determinant* of $v_1, \ldots, v_n$. Remark that the matrix G is symmetric .

The following result is well-known in Linear Algebra:

Theorem 5.1.1 *a) The matrix $G(v_1, \ldots, v_n)$ is positive semidefinite.*

b) The vectors $v_1, \ldots, v_n$ are linearly independent if and only if the matrix $G(v_1, \ldots, v_n)$ is nonsingular (i.e., iff $detG(v_1, \ldots, v_n) \neq 0$).

In the sequel we shall suppose that $v_1, \ldots, v_n$ are linearly independent; let $W_n = Span(v_1, \ldots, v_n)$ be the linear subspace of V generated by $v_1, \ldots, v_n$.

Let $v \in V$ be a given vector. Consider the linear system of equations

$$(s) \begin{cases} < v_1, v_1 > x_1 + \cdots + < v_n, v_1 > x_n =< v, v_1 > \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ < v_1, v_n > x_1 + \cdots + < v_n, v_n > x_n =< v, v_n > . \end{cases}$$

According to Theorem 5.1.1 it has a unique solution $(x_1, \ldots, x_n)$. Let $w = x_1 v_1 + \cdots + x_n v_n$.
From the above system we infer:

$$< w, v_1 >=< v, v_1 >, \ldots, < w, v_n >=< v, v_n > .$$

This means that $< v - w, v_j >= 0, j = 1, \ldots, n$, which entails

$$< v - w, u >= 0, \quad u \in W_n. \tag{5.1.1}$$

So we have a vector $w \in W_n$ satisfying (5.1.1).

It is easy to see that $w \in W_n$ is uniquely determined by the condition (5.1.1). Indeed if $w' \in W_n$ satisfies the similar condition

$$< v - w', u >= 0, \ u \in W_n,$$

then we have $< w', u >=< v, u >=< w, u >, \ u \in W_n$, i.e.,

$$< w' - w, u >= 0, \ u \in W_n;$$

since $w' - w \in W_n$, we get $< w' - w, w' - w >= 0$ which yields $w' = w$.

Definition 5.1.2 *The vector $w \in W_n$ uniquely determined by condition (5.1.1) is called the orthogonal projection of v onto W_n.*

As usual, the norm associated to the inner-product is defined by

$$\|v\| = \sqrt{< v, v >}, \quad v \in V,$$

and the distance between $u, v \in V$ by

$$d(u, v) = \|u - v\|.$$

Moreover, the distance from the vector $v \in V$ to the subspace W_n is defined by

$$dist(v, W_n) = \inf_{u \in W_n} d(v, u).$$

Let $v \in V$ be given and $w \in W_n$ the orthogonal projection of v onto W_n. For an arbitrary $u \in W_n$ we have, according to (5.1.1),

$$< v - w, w - u >= 0.$$

Now by the Pythagorean theorem,

$$\|v - u\|^2 = \|(v - w) + (w - u)\|^2 = \|v - w\|^2 + \|w - u\|^2,$$

so that

$$\|v-u\|^2 \geq \|v-w\|^2$$

with strict inequality unless $u = w$.
(This can be expressed by saying that $\|v-w\|^2$ is the *least-square*).

Consequently,

$$dist(v, W_n) = d(v, w).$$

We shall apply these general results in the case when the vectors are the random variables on a given probability space and the inner-product of the random variables f and g is defined by

$$< f, g >= E(fg).$$

In this case the norm of f is

$$\|f\| = \sqrt{< f, f >} = \sqrt{E(f^2)}$$

and the distance between f and g:

$$d(f, g) = \|f-g\| = \sqrt{E((f-g)^2)}.$$

5.2 The Gram-Schmidt orthogonalization process

The description of the orthogonal projection w via the system (s) is particularly simple when the basis $\{v_1, \ldots, v_n\}$ of W_n is *orthonormal*, i.e., when $< v_i, v_j >= \delta_{ij},\ i, j = 1, \ldots, n$. Indeed, in this case the solution of (s) is given by

$$x_i =< v, v_i >, \quad i = 1, \ldots, n$$

and consequently

$$w = \sum_{i=1}^{n} < v, v_i > v_i.$$

Given an arbitrary basis $\{b_1, \ldots, b_n\}$ of W_n, it is possible to construct a related orthonormal basis $\{v_1, \ldots, v_n\}$ by using the *Gram-Schmidt orthogonalization process*; let us recall it from Linear Algebra.

1. Normalize the first vector:

$$v_1 = \frac{b_1}{\|b_1\|}.$$

2. The orthogonal projection of b_2 onto $Span(v_1)$ is the vector $< b_2, v_1 > v_1$; consider the difference $e_2 = b_2 - < b_2, v_1 > v_1$ and normalize e_2:
$$v_2 = \frac{e_2}{\|e_2\|}.$$

3. The orthogonal projection of b_3 onto $Span(v_1, v_2)$ is
$$< b_3, v_1 > v_1 + < b_3, v_2 > v_2;$$
consider the difference
$$e_3 = b_3 - < b_3, v_1 > v_1 - < b_3, v_2 > v_2$$
and normalize e_3:
$$v_3 = \frac{e_3}{\|e_3\|}.$$

4. Now proceed inductively: to form the next orthogonal vector using b_k, determine the projection $\sum\limits_{i=1}^{k-1} < b_k,\ v_i > v_i$ of b_k onto $Span(v_1, \ldots, v_{k-1})$, consider the difference
$$e_k = b_k - \sum_{i=1}^{k-1} < b_k, v_i > v_i$$
and normalize e_k:
$$v_k = \frac{e_k}{\|e_k\|}.$$
The orthonormal basis $\{v_1, \ldots, v_n\}$ has also the property that
$$Span(v_1, \ldots, v_j) = Span(b_1, \ldots, b_j), \quad j = 1, \ldots, n.$$

☞ **Example 5.2.1** Let $X, W_1, W_2, \ldots$ be independent random variables such that
$$E(X) = E(W_j) = 0,\ E(X^2) = a^2,\ E(W_j^2) = m^2,\ j = 1, 2, \ldots$$
Let $H_j = X + W_j$. We want to find the best linear estimate of X based on $H_1, \ldots, H_k$, i.e., we need $\widehat{X}_k \in Span(H_1, \ldots, H_k)$ such that
$$d(X, \widehat{X}_k) \leq d(X, H) \text{ for all } H \in Span(H_1, \ldots, H_k).$$

(Here $d(X,Y) = \|X-Y\| = \sqrt{< X-Y, X-Y >} = \sqrt{E((X-Y)^2)}$). We already know that $\widehat{X}_k$ is the orthogonal projection of X onto $Span(H_1, \dots, H_k)$. To describe it explicitly, denote

$$Y_k = \frac{1}{k}(H_1 + \cdots + H_k),\ a_k = \frac{a^2}{a^2 + m^2/k},\ U_k = a_k Y_k.$$

Then $U_k \in Span(H_1, \dots, H_k)$. For $1 \le i \le k$ we have

$$\begin{aligned} &E((X-U_k)H_i) = E(XH_i) - a_k E(Y_k H_i) = \\ &= E(X(X+W_i)) - a_k \frac{1}{k} \sum_{j=1}^{k} E(H_j H_i) = \\ &E(X^2) + E(X)E(W_i) - \frac{1}{k} a_k \sum_{j=1}^{k} E((X+W_j)(X+W_i)) = \\ &= a^2 - \frac{1}{k} a_k (ka^2 + m^2) = 0. \end{aligned}$$

This means that

$$\langle X - U_k,\ H_i \rangle = 0,\ i = 1, \dots, k,$$

and so $U_k \in Span(H_1, \dots, H_k)$ is the orthogonal projection of X onto $Span(H_1, \dots, H_k)$.

We conclude that

$$\widehat{X}_k = \frac{a^2}{ka^2 + m^2}(H_1 + \cdots + H_k).$$

5.3 Problems

1. Prove Theorem 5.1.1.
2. Let X_i be independent random variables, $i \in \{-2, -1, 0, 1, 2, 3\}$. Suppose that X_i is uniformly distributed in $[i-1, i+1]$.

   Find the orthogonal projection of X_3 onto the space generated by the other variables.

CHAPTER 6

Correlation and regression

6.1 Covariance

Let X and Y be random variables. The number

$$C(X,Y) = E\left((X - EX)(Y - EY)\right)$$

is called the *covariance* of X and Y.

Let us remark that

$$\begin{aligned} C(X,Y) =& E(XY - (EX)Y - (EY)X + (EX)(EY)) = \\ & E(XY) - (EX)(EY). \end{aligned}$$

Thus we have also

$$C(X,Y) = E(XY) - (EX)(EY).$$

Obviously if X and Y are independent then $C(X,Y) = 0$. On the other hand:

(a) $C(X,X) = E(X^2) - (EX)^2 = Var(X)$

(b) $Var(\sum\limits_{i=1}^{n} X_i) = \sum\limits_{i=1}^{n} Var(X_i) + 2\sum\limits_{i<j} C(X_i, X_j)$.

Indeed,

$$\begin{aligned} Var\left(\sum_{i=1}^{n} X_i\right) &= E\left(\sum_{i=1}^{n}(X_i - EX_i)\right)^2 = E\left(\sum_{i=1}^{n}(X_i - EX_i)^2\right) + \\ + 2E&\left(\sum_{i<j}(X_i - EX_i)(X_j - EX_j)\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} C(X_i, X_j). \end{aligned}$$

As a consequence of the above results, if $X_1, \dots, X_n$ are pairwise independent, then

$$Var(X_1 + \cdots + X_n) = Var(X_1) + \cdots + Var(X_n).$$

6.2 Coefficient of correlation

The number

$$r(X,Y) = \frac{C(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

is called the *coefficient of linear correlation*, or simply, the *correlation coefficient* of X and Y.

Theorem 6.2.1 *(i)* $-1 \le r(X,Y) \le 1$.

(ii) $r(X,Y) = 1$ *if and only if there exist* $a, b \in \mathbb{R},\ a > 0$ *such that* $P(Y = aX + b) = 1$.

(iii) $r(X,Y) = -1$ *if and only if there exist* $a, b \in \mathbb{R},\ a < 0$ *such that* $P(Y = aX + b) = 1$.

Proof. (i) Let us apply Schwarz' inequality to $X - EX$ and $Y - EY$; we obtain

$$|E\left((X - EX)(Y - EY)\right)| \le \sqrt{E(X - EX)^2}\sqrt{E(Y - EY)^2}.$$

This is equivalent to

$$|C(X,Y)| \le \sqrt{Var(X)}\sqrt{Var(Y)}$$

and therefore to $|r(X,Y)| \le 1$. This proves (i).
Suppose now that there exist $a, b \in \mathbb{R}$ such that $Y = aX + b$ almost surely, i.e., $P(Y = aX + b) = 1$. Then $C(X,Y) = C(X, aX + b) = E((X - EX)(aX + b - aEX - b)) = aE(X - EX)^2 = aVar(X)$.
Moreover, $Var(Y) = Var(aX + b) = E(aX + b - aEX - b)^2 = a^2E(X - EX)^2 = a^2Var(X)$.
This yields

$$r(X,Y) = \frac{aVar(X)}{|a|Var(X)} = \frac{a}{|a|},$$

so that $|r(X,Y)| = 1$. Of course, $a > 0$ implies $r(X,Y) = 1$, and $a < 0$ implies $r(X,Y) = -1$.

Conversely, suppose that $|r(X,Y)| = 1$. This entails

$$(C(X,Y))^2 = Var(X)Var(Y),$$

i.e.,

$$(E((X-EX)(Y-EX)))^2 = E(X-EX)^2 \cdot E(Y-EY)^2.$$

So we have equality in Schwarz' inequality, which means that there exists $t_0 \in \mathbb{R}$ such that $Y - EY = t_0(X - EX)$, i.e.,

$$Y = t_0 X + EY - t_0 EX \quad \text{almost surely.}$$

Thus $Y = aX + b$ with $a = t_0$ and $b = EY - t_0 EX$. As before, $r(X,Y) = 1$ is equivalent to $a > 0$ and $r(X,Y) = -1$ to $a < 0$. So (ii) and (iii) are proved.

6.3 Regression lines. The least squares method

Suppose we have some experimental values $(x_i, y_i), i = 1, \ldots, n$, of the pair (X,Y). Each experimental value provides one point in the plane, and the collection of these points is the *scatter diagram* for the experiment.

If all the pairs (x_i, y_i) satisfy some equation of the form $y = f(x)$ (with a given function f), we say that the random variables X, Y are *perfectly correlated* (or *functionally related*).

Otherwise, some lesser degree of correlation may be present, depending on how closely the points on the scatter diagram tend to cluster around a certain curve. If the points are irregularly scattered in the plane, with no apparent clustering, we say that the variables are *uncorrelated.*

Suppose that the points (x_i, y_i) on a scatter diagram appear to cluster closely round some curve; the problem is to find the curve (i.e., its equation $y = f(x)$) which approximates the "cluster-curve". It will be called the *curve of regression.* Such a curve may be used to predict the value of Y by using the known value of X. The general problem of fitting mathematical curves to statistical data for prediction purposes is termed *regression analysis.*

We shall discuss here only the case of linear correlation and the corresponding *regression lines.*.

Let us estimate the numbers $C(X,Y), Var(X), Var(Y)$ as follows:

$$C(X,Y) = E(XY) - (EX)(EY) = \frac{x_1 y_1 + \cdots + x_n y_n}{n} -$$
$$- \frac{x_1 + \cdots + x_n}{n} \frac{y_1 + \cdots + y_n}{n},$$
$$Var(X) = E(X^2) - (EX)^2 = \frac{x_1^2 + \cdots + x_n^2}{n} - \left(\frac{x_1 + \cdots + x_n}{n}\right)^2,$$
$$Var(Y) = E(Y^2) - (EY)^2 = \frac{y_1^2 + \cdots + y_n^2}{n} - \left(\frac{y_1 + \cdots + y_n}{n}\right)^2.$$

Thus we have an estimation of $r(X,Y) = \frac{C(X,Y)}{\sqrt{Var(X)Var(Y)}}$.

(a) If $r = 1$ or $r = -1$, the sample points lie on a straight line $y = ax + b$.

(b) If r is near to 1 or to -1, there is a strong linear correlation between the variables; the sample points cluster around a certain straight line $y = ax + b$. There is a *positive linear correlation* when r is near 1 (i.e., $a > 0$), and a *negative linear correlation* when r is near -1 (i.e., $a < 0$).

(c) If r is close to zero, there is little correlation between the variables, unless the scatter diagram indicates the existence of some non-linear relationship.

Now suppose that r is near to 1 or to -1; the problem is to "fit" a line $y = ax+b$ through the points (x_i, y_i), $i = 1, \ldots, n$ of the scatter diagram. How can we judge which line is best? In what sense is it "best"? The most important method is known as the *method of least squares.* It consists in finding the line $y = ax + b$ for which the "sum of squares"

$$S(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

has the smallest possible value.

The values of a and b which minimize S are the solution of the equations

$$\frac{\partial S}{\partial a} = 0, \quad \frac{\partial S}{\partial b} = 0.$$

which reduce to

$$\begin{cases} \sum (y_i - ax_i - b)x_i = 0 \\ \sum (y_i - ax_i - b) = 0. \end{cases}$$

This system becomes

$$\begin{cases} a\sum x_i + bn = \sum y_i \\ a\sum x_i^2 + b\sum x_i = \sum x_i y_i. \end{cases}$$

We get immediately

$$a = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{(\sum y_i)(\sum x_i^2) - (\sum x_i)(\sum x_i y_i)}{n\sum x_i^2 - (\sum x_i)^2}.$$

This leads to

$$\begin{aligned}
a &= \frac{E(XY) - (EX)(EY)}{E(X^2) - (EX)^2} = \frac{C(X,Y)}{Var(X)} \\
b &= \frac{(EY)E(X^2) - (EX)E(XY)}{E(X^2) - (EX)^2} = \\
&= \frac{(EY)(E(X^2) - (EX)^2) + (EY)(EX)^2 - (EX)E(XY)}{Var(X)} = \\
&= EY - E(X)\frac{C(X,Y)}{Var(X)}.
\end{aligned}$$

So the regression line has the equation

$$y = \frac{C(X,Y)}{Var(X)}x + EY - (EX)\frac{C(X,Y)}{Var(X)}$$

which can be written as

$$y - EY = \frac{C(X,Y)}{Var(X)}(x - EX).$$

This is called the "Y on X" regression line.
A version of the same method of least squares consists in finding the line $y = px + q$ for which

$$S(p,q) = \sum_{i=1}^{n}(x_i - (py_i + q))^2$$

has the smallest possible value.

It leads to the "X on Y" regression line:

$$x - EX = \frac{C(X,Y)}{Var(Y)}(y - EY).$$

Let us remark that both regression lines pass through the point (EX, EY). In general, the closer the regression lines are to one another the stronger is the linear relationship between the variables.

6.4 The covariance matrix

Let $X : \Omega \longrightarrow \mathbb{R}^n$ be an n-dimensional random variable. Then

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_n \end{pmatrix}, \quad EX = \begin{pmatrix} EX_1 \\ EX_2 \\ \dots \\ EX_n \end{pmatrix},$$

where $X_1, X_2, \dots, X_n : \Omega \longrightarrow \mathbb{R}$ are 1-dimensional random variables.

We define the *covariance matrix* of X by

$$C_X = E((X - EX)(X - EX)^t).$$

Considering the usual inner-product $< f, g >= E(fg)$, we have

$$C_X = E \begin{pmatrix} X_1 - EX_1 \\ \dots \\ X_n - EX_n \end{pmatrix} \cdot (X_1 - EX_1, \dots, X_n - EX_n) =$$

$$= \begin{pmatrix} < X_1 - EX_1, X_1 - EX_1 > & \dots & < X_1 - EX_1, X_n - EX_n > \\ \dots & \dots & \dots \\ < X_n - EX_n, X_1 - EX_1 > & \dots & < X_n - EX_n, X_n - EX_n > \end{pmatrix} =$$

$$= G(X_1 - EX_1, \dots, X_n - EX_n),$$

where G is the Gram matrix.

It follows that

(a) C_X is positive semidefinite.

(b) C_X is nonsingular if and only if the vectors $X_1 - EX_1, \dots, X_n - EX_n$ are linearly independent.

On the other hand we have also

$$C_X = \begin{pmatrix} C(X_1, X_1) & \dots & C(X_1, X_n) \\ \dots & \dots & \dots \\ C(X_n, X_1) & \dots & C(X_n, X_n) \end{pmatrix}.$$

6.5 Problems

1. Prove that

$$C(X + Y, X + Y) = C(X, X) + C(Y, Y) + 2C(X, Y).$$

2. Let $a, b, c, d \in \mathbb{R}$, $bd > 0$.
Let X, Y be random variables, $U = a + bX$, $V = c + dY$. Prove that $r(U, V) = r(X, Y)$.

3. Let $X_1, \dots, X_n$ be independent random variables such that $VarX_1 = \dots = VarX_n > 0$. Prove that

$$C(a_1X_1 + \dots + a_nX_n,\ b_1X_1 + \dots + b_nX_n) = 0$$

if and only if

$$a_1b_1 + \dots + a_nb_n = 0.$$

4. Let X and Y be independent random variables, each of them normally distributed $N(m, \sigma^2)$. Find $r(aX + bY,\ aX - bY)$.

5. Find the angle between the regression lines of X and Y.

CHAPTER 7

The discrete Kalman filter

7.1 States and observations

Consider a system whose state is described by the unidimensional random variable $x_t,\ t = 0, 1, 2, \ldots$.

The process is governed by the equation

$$x_{t+1} = A_t x_t + w_t, \quad t = 0, 1, 2, \ldots \tag{7.1.1}$$

The sequence (A_t) relates the state at time t to the state at time $t+1$, in the presence of the noise w_t which is a zero-mean random variable such that

$$E w_t w_\tau = Q_t \delta_{t,\tau}. \tag{7.1.2}$$

We perform noisy observations on the system; the observation y_t at time t is related to the state x_t by

$$y_t = C_t x_t + v_t \tag{7.1.3}$$

where the noise v_t is a zero-mean random variable such that

$$E v_t v_\tau = R_t \delta_{t,\tau}. \tag{7.1.4}$$

Assume that

(a) The state noise w_t and the observation noise v_t have correlation

$$E w_t v_\tau = M_t \delta_{t,\tau}. \tag{7.1.5}$$

(b) w_t is uncorrelated with $x_0, \ldots, x_t$ and with $y_0, \ldots, y_{t-1}$.

(c) v_t is uncorrelated with $y_0, \dots, y_{t-1}$ and with $x_0, \dots, x_t$.

(d) x_0 has mean 0 and variance π_0.

By using (7.1.1) and assumption (d) we deduce that

$$Ex_t = 0, \quad t = 0, 1, 2, \dots \tag{7.1.6}$$

and from (7.1.3):

$$Ey_t = 0, \quad t = 0, 1, 2, \dots \tag{7.1.7}$$

The discrete Kalman filter addresses the problem of estimating the state x_t by a linear combination of the vectors from the set $Y_t = \{y_0, y_1, \dots, y_t\}$. The optimality criterion is the minimum averaged mean-squared error; this means that the best estimate is the orthogonal projection of x_t onto *Span* Y_t. (The inner-product of the random variables f, g is -of course-$E(fg)$).

7.2 The innovation process

In order to find the projection of x_t onto *Span* Y_t it is useful to construct an orthogonal basis of this subspace. We shall apply the Gram-Schmidt procedure to the given basis $\{y_0, y_1, \dots, y_t\}$:

$$\varepsilon_0 = y_0,\ \varepsilon_\tau = y_\tau - \sum_{j=0}^{\tau-1} < y_\tau,\ \varepsilon_j > \frac{\varepsilon_j}{\|\varepsilon_j\|^2},\ \tau = 1, \dots, t. \tag{7.2.1}$$

The resulting orthogonal basis $\{\varepsilon_0, \varepsilon_1, \dots, \varepsilon_t\}$ of *Span* y_t is called the *innovation process.*

We shall denote the projection of x onto Span Y_t by x^t. Then

$$y_\tau^{\tau-1} = \sum_{j=0}^{\tau-1} < y_\tau,\ \varepsilon_j > \frac{\varepsilon_j}{\|\varepsilon_j\|^2},$$

which leads to

$$\varepsilon_\tau = y_\tau - y_\tau^{\tau-1}, \quad \tau = 1, \dots, t. \tag{7.2.2}$$

Due to assumption (c) we infer that

$$v_t^{t-1} = 0$$

and from (7.1.3):

$$y_t^{t-1} = C_t x_t^{t-1}. \tag{7.2.3}$$

7.3 The discrete Kalman filter

From (7.1.1) we get

$$x_{t+1}^t = A_t x_t^t + w_t^t. \tag{7.3.1}$$

On the other hand,

$$w_t^t = \sum_{j=0}^{t} < w_t,\ \varepsilon_j > \frac{\varepsilon_j}{\|\varepsilon_j\|^2}$$

where, by virtue of (7.2.2),(7.2.3) and (7.1.3),

$$\begin{aligned}&< w_t, \varepsilon_j >= E[w_t(y_j - y_j^{j-1})] = E[w_t(C_j x_j + v_j - C_j x_j^{j-1})] = \\ &= C_j E(w_t x_j) + E(w_t v_j) - C_j E(w_t x_j^{j-1}).\end{aligned}$$

Using (7.1.5) and assumption (b) we get

$$< w_t, \varepsilon_j >= M_t \delta_{t,j}, \quad j = 0, 1, \ldots, t,$$

which means that

$$w_t^t = \sum_{j=1}^{t} M_t \delta_{t,j} \frac{\varepsilon_j}{\|\varepsilon_j\|^2} = M_t \frac{\varepsilon_t}{\|\varepsilon_t\|^2}.$$

Finally we get

$$x_{t+1}^t = A_t x_t^t + \frac{M_t}{\|\varepsilon_j\|^2}\varepsilon_t. \tag{7.3.2}$$

Let us remark also that

$$x_{t+1}^{t+1} = \sum_{i=0}^{t+1} < x_{t+1},\ \varepsilon_i > \frac{\varepsilon_i}{\|\varepsilon_i\|^2},$$

that is,

$$x_{t+1}^{t+1} = x_{t+1}^t + E(x_{t+1}\varepsilon_{t+1}) \frac{\varepsilon_{t+1}}{\|\varepsilon_{t+1}\|^2}. \tag{7.3.3}$$

Equations (7.3.2) and (7.3.3) constitute the heart of the Kalman filter. They are recursive formulas for determining the best estimate of the state x_{t+1} based on the observations $y_0, y_1, \ldots, y_{t+1}$.

What remains is to find explicit, recursive representations of the expectations appearing in (7.3.2) and (7.3.3): $E(\varepsilon_{t+1}\varepsilon_{t+1}) = \|\varepsilon_{t+1}\|^2$ and $E(x_{t+1}\varepsilon_{t+1})$.

7.4 Estimation of the error

Let $P_t^{t-1} = \|x_t - x_t^{t-1}\|^2$ be the estimation of the error. We have (see (7.2.2) and (7.2.3):

$$\varepsilon_t = y_t - y_t^{t-1} = y_t - C_t x_t^{t-1} = C_t(x_t - x_t^{t-1}) + v_t,$$

so that

$$E(\varepsilon_t)^2 = C_t^2 P_t^{t-1} + 2C_t E(x_t v_t) - 2C_t E(x_t^{t-1} v_t) + E(v_t)^2.$$

Now (7.1.4) and assumption (c) yield

$$\|\varepsilon_t\|^2 = C_t^2 P_t^{t-1} + R_t. \tag{7.4.1}$$

Moreover,

$$\begin{aligned}&E(x_t\varepsilon_t) = E[x_t(C_t(x_t - x_t^{t-1}) + v_t)] = \\ &= C_t E[x_t(x_t - x_t^{t-1})] + E(x_t v_t) = \\ &= C_t E(x_t - x_t^{t-1})^2 + C_t E[x_t^{t-1}(x_t - x_t^{t-1})].\end{aligned}$$

Since $x_t - x_t^{t-1}$ is orthogonal to *Span* Y_{t-1}, we get

$$E(x_t\varepsilon_t) = C_t P_t^{t-1}. \tag{7.4.2}$$

7.5 The Riccati equation

We shall establish a recursive expression for the estimation of the error P_t^{t-1}.

Denote

$$\Pi_t = E(x_t)^2, \quad \textstyle\sum_t^{t-1} = E(x_t^{t-1})^2.$$

We have the orthogonal decomposition

$$x_t = x_t^{t-1} + (x_t - x_t^{t-1})$$

and consequently

$$\Pi_t = E(x_t)^2 = E(x_t^{t-1})^2 + E(x_t - x_t^{t-1})^2 = \textstyle\sum_t^{t-1} + P_t^{t-1}.$$

Thus

$$P_t^{t-1} = \Pi_t - \textstyle\sum_t^{t-1}. \tag{7.5.1}$$

Now, using (7.1.2),

$$\Pi_{t+1} = E(x_{t+1})^2 = E(A_t x_t + w_t)^2 = A_t^2 \Pi_t + Q_t.$$

So we have the recursive representation

$$\Pi_{t+1} = A_t^2 \Pi_t + Q_t, \quad t = 0, 1, 2, \dots \tag{7.5.2}$$

and $\Pi_0 = E(x_0)^2$ is given.

On the other hand, by virtue of (7.3.2),

$$\begin{aligned} \textstyle\sum_{t+1}^t &= E(x_{t+1}^t)^2 = E(A_t x_t^t + \tfrac{M_t}{\|\varepsilon_t\|^2}\varepsilon_t)^2 = \\ &= A_t^2 E(x_t^t))^2 + 2A_t M_t \tfrac{1}{\|\varepsilon_t\|^2} E(x_t^t \varepsilon_t) + \tfrac{M_t^2}{\|\varepsilon_t\|^2}. \end{aligned}$$

Now (7.3.3) gives

$$x_t^t = x_t^{t-1} + E(x_t \varepsilon_t)\frac{\varepsilon_t}{\|\varepsilon_t\|^2} \tag{7.5.3}$$

and since x_t^{t-1} is orthogonal to ε_t, we find:

$$E(x_t^t)^2 = \textstyle\sum_t^{t-1} + (E(x_t \varepsilon_t))^2 \tfrac{1}{\|\varepsilon_t\|^2} \tag{7.5.4}$$

and, moreover,

$$E(x_t^t \varepsilon_t) = E(x_t \varepsilon_t). \tag{7.5.5}$$

Using (7.5.4) and (7.5.5) we deduce

$$\begin{aligned} \textstyle\sum_{t+1}^t &= A_t^2 \textstyle\sum_t^{t-1} + A_t^2 (E(x_t\varepsilon_t))^2 \tfrac{1}{\|\varepsilon_t\|^2} + 2A_t M_t \tfrac{1}{\|\varepsilon_t\|^2} E(x_t^t \varepsilon_t) + \\ + \tfrac{M_t^2}{\|\varepsilon_t\|^2} &= A_t^2 \textstyle\sum_t^{t-1} + \tfrac{1}{\|\varepsilon_t\|^2}\left(A_t^2 (E(x_t\varepsilon_t))^2 + 2A_t M_t E(x_t \varepsilon_t) + M_t^2\right) \end{aligned}$$

which entails

$$\textstyle\sum_{t+1}^t = A_t^2 \textstyle\sum_t^{t-1} + \tfrac{1}{\|\varepsilon_t\|^2}(A_t E(x_t \varepsilon_t) + M_t)^2. \tag{7.5.6}$$

From (7.5.1), (7.5.2), (7.5.6) and (7.4.2) we get

$$\begin{aligned} P_{t+1}^t = \Pi_{t+1} - \textstyle\sum_{t+1}^t &= A_t^2 \Pi_t + Q_t - A_t^2 \textstyle\sum_t^{t-1} - \\ &- \tfrac{1}{\|\varepsilon_t\|^2}(A_t C_t P_t^{t-1} + M_t)^2 = \\ = A_t^2 \left(\Pi_t - \textstyle\sum_t^{t-1}\right) &+ Q_t - \tfrac{1}{\|\varepsilon_t\|^2}(A_t C_t P_t^{t-1} + M_t)^2. \end{aligned}$$

Now using (7.5.1) and (7.4.1) we deduce

$$P_{t+1}^t = A_t^2 P_t^{t-1} + Q_t - \frac{(A_t C_t P_t^{t-1} + M_t)^2}{C_t^2 P_t^{t-1} + R_t}\ ,\ t = 0, 1, \dots \tag{7.5.7}$$

This recursive formula for P_{t+1}^t is known as the *Riccati equation.* Let us remark that taking $x_0^{-1} = 0$ we get $P_0^{-1} = \|x_o\|^2 = \Pi_0$.

7.6 Kalman filter at work

With the development of the Riccati equation (7.5.7) we have completed the steps needed for the celebrated Kalman filter. It works as follows:

First of all, P_t^{t-1} is obtained via (7.5.7) for $t = 1, 2, \dots$, starting from the known $P_0^{-1} = \Pi_0$.

From (7.2.2) and (7.2.3) we derive

$$\varepsilon_t = y_t - C_t x_t^{t-1}.$$

Now (7.3.2) and (7.4.1) yield the *time-update equation:*

$$x_{t+1}^t = A_t x_t^t + \frac{M_t}{C_t^2 P_t^{t-1} + R_t}(y_t - C_t x_t^{t-1}). \tag{7.6.1}$$

From (7.3.3), (7.4.1) and (7.4.2) we deduce the *measurement-update equation:*

$$x_{t+1}^{t+1} = x_{t+1}^t + C_{t+1} P_{t+1}^t \frac{1}{C_{t+1}^2 P_{t+1}^t + R_{t+1}}(y_{t+1} - C_{t+1} x_{t+1}^t). \tag{7.6.2}$$

Take $x_0^{-1} = 0$. With $t = -1$ in (7.6.2) we get x_0^0.
Putting $t = 0$ in (7.6.1) yields x_1^0; $t = 0$ in (7.6.2) yields x_1^1. We continue with $t = 1$ in (7.6.1) and then in (7.6.2), getting x_2^1 and x_2^2; and so on!

7.7 Uncorrelated noises

The presentation of the Kalman filter can be simplified under the hypothesis that $M_t = 0$, i.e., the state-update noise and the measurement noise are uncorrelated. Let us see the results in this case.

Define the *Kalman gain*

$$K_t = \frac{C_t}{\|\varepsilon_t\|^2} P_t^{t-1} = \frac{C_t}{C_t^2 P_t^{t-1} + R_t} P_t^{t-1}. \tag{7.7.1}$$

Then (7.6.2) becomes

$$x_{t+1}^{t+1} = x_{t+1}^t + K_{t+1}(y_{t+1} - C_{t+1} x_{t+1}^t) = x_{t+1}^t + K_{t+1}\varepsilon_{t+1}. \tag{7.7.2}$$

Denote

$$P_t^t = E(x_t - x_t^t)^2.$$

Then

$$P_t^t = E(x_t - x_t^{t-1} - K_t\varepsilon_t)^2 = P_t^{t-1} + K_t^2\|\varepsilon_t\|^2 - 2K_t E((x_t - x_t^{t-1})\varepsilon_t).$$

Since, by (7.2.2) and (7.2.3), $\varepsilon_t = y_t - C_t x_t^{t-1} = C_t(x_t - x_t^{t-1}) + v_t$, we have (see (c)): $E((x_t - x_t^{t-1})\varepsilon_t) = C_t P_t^{t-1}$, so that (see (7.7.1))

$$\begin{aligned} P_t^t &= P_t^{t-1} + K_t^2(C_t^2 P_t^{t-1} + R_t) - 2K_t C_t P_t^{t-1} = \\ &= P_t^{t-1} + K_t C_t P_t^{t-1} - 2K_t C_t P_t^{t-1} = P_t^{t-1}(1 - K_t C_t). \end{aligned}$$

Thus we have

$$P_t^t = (1 - K_t C_t) P_t^{t-1}. \tag{7.7.3}$$

From (7.5.7) with $M_t = 0$ we get

$$\begin{aligned} P_{t+1}^t &= A_t^2 P_t^{t-1} + Q_t - \frac{A_t^2 C_t^2 (P_t^{t-1})^2}{C_t^2 P_t^{t-1} + R_t} = \\ &= A_t^2 P_t^{t-1} \frac{R_t}{C_t^2 P_t^{t-1} + R_t} + Q_t. \end{aligned}$$

But (7.7.3) and (7.7.1) yield

$$P_t^t = \left(1 - \frac{C_t^2 P_t^{t-1}}{C_t^2 P_t^{t-1} + R_t}\right) P_t^{t-1} = \frac{R_t P_t^{t-1}}{C_t^2 P_t^{t-1} + R_t},$$

so that

$$P_{t+1}^t = A_t^2 P_t^t + Q_t. \tag{7.7.4}$$

Summarizing from (7.6.1), (7.7.4), (7.7.1), (7.7.2) and (7.7.3), we have the following formulation of the Kalman filter in the case of uncorrelated noises (i.e., $M_t = 0$):

Time update:

$$x_{t+1}^t = A_t x_t^t \tag{7.7.5}$$

$$P_{t+1}^t = A_t^2 P_t^t + Q_t. \tag{7.7.6}$$

Kalman gain:

$$K_{t+1} = \frac{C_{t+1} P_{t+1}^t}{C_{t+1}^2 P_{t+1}^t + R_{t+1}}. \tag{7.7.7}$$

Measurement update:

$$x_{t+1}^{t+1} = x_{t+1}^t + K_{t+1}(y_{t+1} - C_{t+1} x_{t+1}^t) \tag{7.7.8}$$

$$P_{t+1}^{t+1} = (1 - K_{t+1} C_{t+1}) P_{t+1}^t. \tag{7.7.9}$$

7.8 Linearization of nonlinear systems

Consider now a nonlinear discrete-time system

$$x_{t+1} = A(x_t, t) + w_t \tag{7.8.1}$$

$$y_t = C(x_t, t) + v_t \tag{7.8.2}$$

where $(w_t)_{t=0,1,\ldots}$ and $(v_t)_{t=0,1,\ldots}$ are uncorrelated , zero-mean noise sequences.

Suppose that we know a *nominal* (or *reference*) trajectory $\overline{x_t}$ satisfying

$$\overline{x}_{t+1} = A(\overline{x}_t, t) \quad , \ t = 0, 1, \ldots \tag{7.8.3}$$

and the deviations δ_{x_t} defined by

$$\delta_{x_t} = x_t - \overline{x}_t \tag{7.8.4}$$

are small. Then the equation (7.8.1) may be linearized:

$$\overline{x}_{t+1} + \delta x_{t+1} = A(\overline{x}_t, t) + A'_x(\overline{x}_t, t)\delta x_t + w_t.$$

Together with (7.8.3), this leads to

$$\delta x_{t+1} = A'_x(\overline{x}_t, t)\delta x_t + w_t. \tag{7.8.5}$$

Define the reference observations $\overline{y}_t$ by

$$\overline{y}_t = C(\overline{x}_t, t) \quad , t = 0, 1, \ldots \tag{7.8.6}$$

and suppose that the deviations

$$\delta y_t = y_t - \overline{y}_t \tag{7.8.7}$$

are small. Then (7.8.2) may be linearized:

$$\overline{y}_t + \delta y_t = C(\overline{x}_t, t) + C'_x(\overline{x}_t, t)\delta x_t + v_t.$$

Thus we get

$$\delta y_t = C'_x(\overline{x}_t, t)\delta x_t + v_t. \tag{7.8.8}$$

Now define $A_t := A'_x(\overline{x}_t, t)$ and $C_t := C'_x(\overline{x}_t, t)$.
Then (7.8.5) and (7.8.8) become

$$\delta x_{t+1} = A_t \delta x_t + w_t \tag{7.8.9}$$

$$\delta y_t = C_t \delta x_t + v_t. \tag{7.8.10}$$

Applying the Kalman filter to this *linear* model we obtain the *deviation estimates* $(\delta x_t)^t$ and then the *trajectory estimates*

$$x_t^t = \overline{x}_t + (\delta x_t)^t.$$

This approach is called *global linearization.* The development of a nominal trajectory may be problematic. In some cases it may be possible to generate such a trajectory via computer simulations; experience and intuition may be useful. Sometimes one may simply rely on guesses.

7.9 The extend Kalman filter

Global linearization using a predetermined nominal trajectory is not the only way to approach the linearization problem. Another approach is to determine at each step a local nominal trajectory and update it as information becomes available.

We need initial conditions in the form of x_0^{-1} and P_0^{-1}, the *a priori* state estimate and covariance. In the role of x_0^{-1} we take the best information we have concerning the value x_0, and use it as the first point in the nominal trajectory:

$$\overline{x}_0 = x_0^{-1}. \tag{7.9.1}$$

Consequently, the value of C_0 will be

$$C_0 = C'_x(x_0^{-1}, 0) \tag{7.9.2}$$

and, moreover (see (7.8.6), (7.8.7)):

$$\delta y_0 = y_0 - \overline{y}_0 = y_0 - C(x_0^{-1}, 0). \tag{7.9.3}$$

Using (7.7.1) we have

$$K_0 = \frac{C_0}{C_0^2 P_0^{-1} + R_0} P_0^{-1} \tag{7.9.4}$$

and from (7.7.8), (7.7.9):

$$(\delta x_0)^0 = (\delta x_0)^{-1} + K_0(\delta y_0 - C_0(\delta x_0)^{-1}) \tag{7.9.5}$$

$$P_0^0 = (1 - K_0 C_0) P_0^{-1}. \tag{7.9.6}$$

On the other hand, from $\delta x_0 = x_0 - \overline{x}_0$ we get $(\delta x_0)^{-1} = x_0^{-1} - \overline{x}_0 = 0$ and $(\delta x_0)^0 = x_0 - \overline{x}_0 = x_0^0 - x_0^{-1}$; thus (using also (7.9.3)), (7.9.5) and (7.9.6) become the measurement-update equations at time $t = 0$:

$$x_0^0 = x_0^{-1} + K_0(y_0 - C(x_0^{-1}, 0)) \tag{7.9.7}$$

$$P_0^0 = (1 - K_0C_0)P_0^{-1}. \tag{7.9.8}$$

Now compute A_0 as

$$A_0 = A'_x(x_0^0, 0) \tag{7.9.9}$$

and (see (7.7.6))

$$P_1^0 = A_0^2 P_0^0 + Q_0. \tag{7.9.10}$$

A prediction of the state at $t = 1$ may be

$$x_1^0 = A(x_0^0, 0), \tag{7.9.11}$$

and we take it as the reference trajectory:

$$\overline{x}_1 = x_1^0. \tag{7.9.12}$$

From $\delta x_1 = x_1 - \overline{x}_1$ we get

$$(\delta x_1)^0 = x_1^0 - \overline{x}_1 = 0, \tag{7.9.13}$$

$$(\delta x_1)^1 = x_1^1 - \overline{x}_1 = x_1^1 - x_1^0. \tag{7.9.14}$$

Moreover, according to (7.8.6) and (7.8.7):

$$\delta y_1 = y_1 - \overline{y}_1 = y_1 - C(\overline{x}_1, 1) = y_1 - C(x_1^0, 1),$$

and thus

$$\delta y_1 = y_1 - C(x_1^0, 1). \tag{7.9.15}$$

At this point it is possible to compute

$$C_1 = C'_x(\overline{x}_1, 1) = C'_x(x_1^0, 1) \tag{7.9.16}$$

and (see (7.7.1))

$$K_1 = \frac{C_1}{C_1^2 P_1^0 + R_1} P_1^0. \tag{7.9.17}$$

From (7.7.8) we get

$$(\delta x_1)^1 = (\delta x_1)^0 + K_1(\delta y_1 - C_1(\delta x_1)^0);$$

together with (7.9.13), (7.9.14) and (7.9.15), this leads to

$$x_1^1 = x_1^0 + K_1(y_1 - C(x_1^0, 1)). \tag{7.9.18}$$

From (7.7.9) we infer

$$P_1^1 = (1 - K_1C_1)P_1^0. \tag{7.9.19}$$

Clearly (7.9.18) and (7.9.19) are the measurement-update equations at time $t = 1$.

Continuing by induction, we obtain the *extended Kalman filter*, described as follows.

Measurement update:

$$x_{t+1}^{t+1} = x_{t+1}^{t} + K_{t+1}[y_{t+1} - C(x_{t+1}^{t}, t)] \tag{7.9.20}$$

$$P_{t+1}^{t+1} = (1 - K_{t+1}C_{t+1})P_{t+1}^{t} \tag{7.9.21}$$

where

$$K_{t+1} = P_{t+1}^{t} \frac{C_{t+1}}{C_{t+1}^2 P_{t+1}^{t} + R_{t+1}} \tag{7.9.22}$$

$$C_{t+1} = C'_x(x_{t+1}^{t}, t+1). \tag{7.9.23}$$

Time update:

$$x_{t+1}^{t} = A(x_t^t, t) \tag{7.9.24}$$

$$P_{t+1}^{t} = A_t^2 P_t^t + Q_t \tag{7.9.25}$$

where

$$A_t = A'_x(x_t^t, t). \tag{7.9.26}$$

The extended Kalman filter is initialized by supplying the *a priori* estimate x_0^{-1} and covariance P_0^{-1}.

7.10 Problems

1. Let $Ey_t = 0,\ Ey_t y_\tau = a^{|t-\tau|},\ t, \tau = 0, 1, \ldots$

   (i) Find the innovation process by using the Gram-Schmidt procedure.

   (ii) Let $z_0 = y_0,\ z_t = y_t - ay_{t-1},\ t = 1, 2, \ldots$
   Prove that $\{z_0, z_1, \ldots, z_t\}$ is an orthogonal set and

   $$Span\{z_0, \ldots, z_t\} = Span\{y_0, \ldots, y_t\},\ t = 0, 1, \ldots$$

2. Let $Eu_t = m,\ E(u_t - m)(u_\tau - m) = b^2 a^{|t-\tau|},\ t, \tau \geq 0.$
   Prove that $au_{t-1} + m(1-a)$ is the linear least-squares predictor of u_t given $\{u_0, u_1, \ldots, u_{t-1}\}$.

CHAPTER 8

Markov chains

8.1 Markov chains

Let r be a positive integer and $S = \{1, 2, \ldots, r\}$.

Definition 8.1.1 *A* Markov chain *with state space S is a sequence of S-valued random variables $(X(n))_{n\geq 0}$ such that*

$$\begin{aligned} &P\left(X(n+1) = i_{n+1} | X(n) = i_n, \ldots, X(1) = i_1,\ X(0) = i_0\right) = \\ &\quad = P\left(X(n+1) = i_{n+1} | X(n) = i_n\right) \end{aligned}$$

for all $n \geq 0$ and all $i_0, i_1, \ldots, i_{n+1} \in S$.

Definition 8.1.2 *Denote $P(X(0) = i) = p_0(i),\ i = 1, \ldots, r$. Then $(p_0(1), \ldots, p_0(r))$ is called the* initial probability vector *of the Markov chain.*

Let us remark that $p_0(i) \geq 0,\ i = 1, \ldots, r$, and $p_0(1) + \cdots + p_0(r) = P(X(0) = 1) + \cdots + P(X(0) = r) = P(X(0) \in \{1, \ldots, r\}) = 1$.

Hence the initial probability vector has nonnegative components, the sum of which equals 1.

Definition 8.1.3 *A Markov chain is called homogeneous if for all $i, j \in S$ the conditional probability*

$$P(X(n+1) = j | X(n) = i)$$

is the same for all $n \geq 0$.

In the sequel we shall consider only homogeneous Markov chains and use the notation

$$P(X(n+1)=j \mid X(n)=i)=p(i,j) \text{ for all } n \geq 0.$$

Definition 8.1.4 *The numbers* $p(i,j) \in [0,1]$ *are called the* one-step transition probabilities *of the Markov chain.*

Definition 8.1.5 *The matrix*

$$T=\begin{bmatrix} p(1,1) & p(1,2) & \dots & p(1,r) \\ p(2,1) & p(2,2) & \dots & p(2,r) \\ \dots & \dots & \dots & \dots \\ p(r,1) & p(r,2) & \dots & p(r,r) \end{bmatrix}$$

is called the transition matrix *of the Markov chain.*

Definition 8.1.6 *A square matrix with nonnegative elements is called a* stochastic matrix *if the sum of the elements of each row equals* 1*.*

Proposition 8.1.7 *The transition matrix of a Markov chain is a stochastic matrix.*

Proof. Let $i \in \{1,\dots,r\}$ and $n \geq 1$ be given. Then

$$\begin{aligned} & p(i,1)+p(i,2)+\cdots+p(i,r)=P(X(n+1)=1|X(n)=i)+ \\ & +P(X(n+1)=2|X(n)=i)+\cdots+P(X(n+1)=r|X(n)=i)= \\ & =P(X(n+1)\in S|X(n)=i)=1. \end{aligned}$$

We conclude that T is a stochastic matrix.

☞ **Example 8.1.8** By using the one-step transition probabilities one can compute other conditional probabilities, as the following example shows:

$$
\begin{aligned}
&P(X(4)=j,\ X(5)=k,\ X(6)=h \mid X(3)=i)= \\
&=\frac{P(X(4)=j,\ X(5)=k,\ X(6)=h,\ X(3)=i)}{P(X(3)=i)}= \\
&=\frac{P(X(4)=j,\ X(3)=i)}{P(X(3)=i)}\cdot\frac{P(X(5)=k,\ X(4)=j,\ X(3)=i)}{P(X(4)=j, X(3)=i)}\cdot \\
&\cdot\frac{P(X(6)=h,\ X(5)=k,\ X(4)=j,\ X(3)=i)}{P(X(5)=k,\ X(4)=j,\ X(3)=i)}= \\
&=P(X(4)=j \mid X(3)=i)\cdot P(X(5)=k \mid X(4)=j,\ X(3)=i)\cdot \\
&\cdot P(X(6)=h \mid X(5)=k,\ X(4)=j,\ X(3)=i)= \\
&=P(X(4)=j \mid X(3)=i)\cdot P(X(5)=k \mid X(4)=j)\cdot \\
&\cdot P(X(6)=h \mid X(5)=k)=p(i,j)\cdot p(j,k)\cdot p(k,h).
\end{aligned}
$$

☞ **Example 8.1.9** An urn contains 8 white balls and 4 red balls. We have also an exterior white ball. At the moment $n=0$ we draw a ball from the urn and replace it by the exterior ball. We repeat this process at the moments $n=1,2,\ldots.$
Denote by $X(n)$ the colour of the ball drawn at the moment n; then $X(n)\in\{w,r\}$, hence the state space is $S=\{w,r\}$,

We have

$$
\begin{aligned}
&P(X(0)=w)=\frac{2}{3},\quad P(X(0)=r)=\frac{1}{3}, \\
&P(X(1)=w \mid X(0)=w)=\frac{2}{3}\quad ,P(X(1)=w \mid X(0)=r)=\frac{3}{4}, \\
&P(X(1)=r \mid X(0)=w)=\frac{1}{3}\quad ,P(X(1)=r \mid X(0)=r)=\frac{1}{4}.
\end{aligned}
$$

The content of the urn immediately before the moment $n+1$ is determined only by the state $X(n)$; this content is:

- 8 white and 4 red balls, if $X(n)=w$
- 9 white and 3 red balls, if $X(n)=r$.

Consequently, for all $n \geq 0$ we have

a) $P(X(n+1) = w|\ X(n) = w,\ X(n-1) = \text{arbitrary},\dots,X(0) = \text{arbitrary})$ $= P(X(n+1) = w \mid X(n) = w) = \dfrac{2}{3};$

b) $P(X(n+1) = w|\ X(n) = r,\ X(n-1) = \text{arbitrary},\dots,X(0) = \text{arbitrary})$ $= P(X(n+1) = w \mid X(n) = r) = \dfrac{3}{4};$

c) $P(X(n+1) = r|\ X(n) = w,\ X(n-1) = \text{arbitrary},\dots,X(0) = \text{arbitrary})$ $= P(X(n+1) = r \mid X(n) = w) = \dfrac{1}{3};$

d) $P(X(n+1) = r|\ X(n) = r,\ X(n-1) = \text{arbitrary},\dots,X(0) = \text{arbitrary})$ $= P(X(n+1) = r \mid X(n) = r) = \dfrac{1}{4};$

This shows that $(X(n))_{n\geq 0}$ is a homogeneous Markov chain with initial probability vector $(\frac{2}{3}, \frac{1}{3})$ and transition matrix

$$T = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix}.$$

8.2 The Chapman-Kolmogorov equations

Let $n \geq 0,\ i, j \in S$ be given. We want to compute

$$P(X(n+2) = j \mid X(n) = i).$$

Thus we have

$$
\begin{aligned}
&P(X(n+2)=j \mid X(n)=i)= \\
&=\sum_{k=1}^{r} P(X(n+2)=j,\ (X(n+1)=k \mid X(n)=i)= \\
&=\sum_{k=1}^{r} \frac{P(X(n+2)=j,\ X(n+1)=k,\ X(n)=i)}{P(X(n)=i)}= \\
&=\sum_{k=1}^{r} \frac{P(X(n+2)=j,\ X(n+1)=k,\ X(n)=i)}{P(X(n+1)=k,\ X(n)=i)}\cdot \\
&\cdot \frac{P(X(n+1)=k,\ X(n)=i)}{P(X(n)=i)}= \\
&=\sum_{k=1}^{r} P(X(n+2)=j \mid (X(n+1)=k,\ X(n)=i)\cdot \\
&\cdot P(X(n+1)=k \mid X(n)=i)= \\
&=\sum_{k=1}^{r} P(X(n+2)=j \mid (X(n+1)=k)\ \cdot P(X(n+1)=k \mid X(n)=i)= \\
&=\sum_{k=1}^{r} p(i,k)p(k,j).
\end{aligned}
$$

Denoting

$$p(2,i,j)=P(X(n+2)=j \mid X(n)=i), \quad n \geq 0$$

we have

$$p(2,i,j)=\sum_{k=1}^{r} p(i,k)p(k,j), \quad i,j \in S.$$

Consider now the matrix

$$(p(2,i,j))_{i,j\in S}=\begin{bmatrix} p(2,1,1) & \dots & p(2,1,r) \\ \dots & \dots & \dots \\ p(2,r,1) & \dots & p(2,r,r) \end{bmatrix}.$$

We conclude that

$$(p(2,i,j))_{i,j\in S}=T^2.$$

Moreover, denoting for $m \geq 1$

$$p(m,i,j)=P(X(n+m)=j \mid X(n)=i), \quad n \geq 0,$$

it can be proved by induction that

$$(p(m,i,j))_{i,j\in S} = T^m, \quad m \geq 1.$$

The numbers $p(m,i,j)$ are call *m-step transition probabilities.* Obviously $p(1,i,j) = p(i,j),\ i,j \in S$.

Since $T^{m+n} = T^m \cdot T^n$, we obtain the *Chapman-Kolmogorov equations*:

$$p(m+n,i,j) = \sum_{k=1}^{r} p(m,i,k)p(n,k,j) \quad m,n \geq 1,\ i,j \in S.$$

☞ **Example 8.2.1**

$$\begin{aligned}
&P(X(4)=j,\ X(6)=k,\ X(9)=h,\ \mid X(2)=i) = \\
&= \frac{P(X(4)=j,\ X(6)=k,\ X(9)=h,\ X(2)=i)}{P(X(6)=k,\ X(4)=h,\ X(2)=i)} \cdot \\
&\cdot \frac{P(X(6)=k,\ X(4)=j,\ X(2)=i)}{P(X(4)=j,\ X(2)=i)} \cdot \frac{P(X(4)=j,\ X(2)=i)}{P(X(2)=i)} = \\
&= P(X(9)=h \mid X(6)=k) \cdot P(X(6)=k \mid X(4)=j) \cdot P(X(4)=j \mid X(2)=i) \\
&= p(3,k,h)p(2,j,k)p(2,i,j).
\end{aligned}$$

8.3 Absolute probabilities

Definition 8.3.1 *The numbers*

$$p_n(i) = P(X(n)=i),\ n \geq 0,\ i \in S$$

are called the absolute probabilities *of the Markov chain.*

The n-th probability vector $(p_n(1),\ldots,p_n(r))$ can be expressed in terms of the initial probability vector and the transition matrix:

Theorem 8.3.2 *For all $n \geq 1$ we have*

$$[p_n(1) \quad p_n(2) \quad \ldots \quad p_n(r)] = [p_0(1) \quad p_0(2) \quad \ldots \quad p_0(r)] \cdot T^n.$$

Proof. Let $i \in S$. Then

$$\begin{aligned}
p_n(i) &= P(X(n)=i) = \sum_{j=1}^{r} P(X(n)=i,\ X(0)=j) = \\
&= \sum_{j=1}^{r} P(X(n)=i \mid X(0)=j) \cdot P(X(0)=j) = \sum_{j=1}^{r} p(n,j,i)p_0(j).
\end{aligned}$$

So we have

$$p_n(i) = \sum_{j=1}^{r} p_0(j)p(n,j,i), \quad n \geq 1, \ i \in S.$$

This yields the following relation involving matrices:

$$\begin{aligned} &[p_n(1) \quad p_n(2) \quad \dots \quad p_n(r)] = \\ &= [p_0(1) \quad p_0(2) \quad \dots \quad p_0(r)] \cdot \begin{bmatrix} p(n,1,1) & \dots & p(n,1,r) \\ \dots & \dots & \dots \\ p(n,r,1) & \dots & p(n,r,r) \end{bmatrix}. \end{aligned}$$

Since the last matrix is T^n, the proof is finished.

The n-th probability vector is important: it describes, probabilistically, the state of the system at the moment n. So it is important to know T^n; to this end it is useful to have information about the eigenvalues and the eigenvectors of the stochastic matrix T.

Proposition 8.3.3 *Every stochastic matrix has the eigenvalue* 1*. The vector*

$$v = \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}$$

is an eigenvector associated with this eigenvalue.

Proof. Let

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1r} \\ a_{21} & a_{22} & \dots & a_{2r} \\ \dots & \dots & \dots & \dots \\ a_{r1} & a_{r2} & \dots & a_{rr} \end{bmatrix}$$

be a stochastic matrix. We have

$$\sum_{j=1}^{r} a_{ij} = 1 \quad , i = 1, \dots, r.$$

It follows immediately that

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1r} \\ a_{21} & a_{22} & \dots & a_{2r} \\ \dots & \dots & \dots & \dots \\ a_{r1} & a_{r2} & \dots & a_{rr} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}.$$

This means that v is an eigenvector of A, associated to the eigenvalue 1.

Proposition 8.3.4 *If A is stochastic, then A^n is stochastic for all $n \geq 1$.*

Proof. Since the elements of A are nonnegative, those of A^n will be also nonnegative.

With the above notation we have $Av = v$ and by induction $A^n v = v,\ n \geq 1$; it follows that the elements on each row of A^n have the sum equal to 1. We conclude that A^n is a stochastic matrix.

The following example suggests an important property which will be discussed in the next section.

☞ **Example 8.3.5** Consider a Markov chain with the transition matrix

$$T = \begin{bmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{6} & \frac{1}{3} & \frac{1}{2} \end{bmatrix}.$$

We know that $v_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ is an eigenvector associated to the eigenvalue $\lambda_1 = 1$. The reader will check that the other eigenvalues are $\lambda_2 = \frac{1}{3}$ and $\lambda_3 = \frac{1}{6}$, with associated eigenvectors

$$v_2 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix},\ v_3 = \begin{bmatrix} 2 \\ -3 \\ 2 \end{bmatrix}.$$

The transition matrix from the canonical basis of $\mathbb{R}^3$ to the basis $\{v_1, v_2, v_3\}$ is:

$$P = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 0 & -3 \\ 1 & -1 & 2 \end{bmatrix}.$$

The matrix T and its diagonal form are connected by:

$$T = P \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{6} \end{bmatrix} \cdot P^{-1}.$$

We obtain

$$T^n = P \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & \dfrac{1}{3^n} & 0 \\ 0 & 0 & \dfrac{1}{6^n} \end{bmatrix} \cdot P^{-1}.$$

Consequently,

$$\lim_{n\to\infty} T^n = P \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \cdot P^{-1},$$

and it is easy to verify that

$$\lim_{n\to\infty} T^n = \frac{1}{10} \cdot \begin{bmatrix} 3 & 4 & 3 \\ 3 & 4 & 3 \\ 3 & 4 & 3 \end{bmatrix}.$$

8.4 Regular Markov chains. The limit distribution

Definition 8.4.1 *We say that a stochastic matrix T is* regular *if there exists $k \in \{1, 2, \dots\}$ such that all the elements of T^k are strictly positive.*

Definition 8.4.2 *A Markov chain is called* regular *if its transition matrix T is regular.*

We now state, without proof, the following fundamental theorem for regular Markov chains.

Theorem 8.4.3 *Let T be the transition matrix of a regular Markov chain. Then there exists*

$$B = \lim_{n\to\infty} T^n.$$

The matrix B is stochastic and all its rows are identical.

Remark 8.4.4 An illustration for this result is provided by Example 8.3.5, where

$$B = \begin{bmatrix} \dfrac{3}{10} & \dfrac{4}{10} & \dfrac{3}{10} \\ \dfrac{3}{10} & \dfrac{4}{10} & \dfrac{3}{10} \\ \dfrac{3}{10} & \dfrac{4}{10} & \dfrac{3}{10} \end{bmatrix}.$$

Returning to the general case of a regular Markov chain, we have

$$\lim_{n\to\infty} T^n = B = \begin{bmatrix} b_1 & b_2 & \dots & b_r \\ b_1 & b_2 & \dots & b_r \\ \dots & \dots & \dots & \dots \\ b_1 & b_2 & \dots & b_r \end{bmatrix},$$

where $b_1 \geq 0, \dots, b_r \geq 0, \quad b_1 + \dots + b_r = 1$.

Theorem 8.4.5 *If the Markov chain is regular, then*

$$\lim_{n\to\infty} (p_n(1), \dots, p_n(r)) = (b_1, \dots, b_r).$$

Proof. We know that

$$[p_n(1) \quad p_n(2) \quad \dots \quad p_n(r)] = [p_0(1) \quad p_0(2) \quad \dots \quad p_0(r)] \cdot T^n$$

and $p_0(1) + \dots + p_0(r) = 1$.
It follows that

$$\lim_{n\to\infty} [p_n(1) \quad p_n(2) \quad \dots \quad p_n(r)] =$$

$$= [p_0(1) \quad p_0(2) \quad \dots \quad p_0(r)] \cdot \begin{bmatrix} b_1 & b_2 & \dots & b_r \\ b_1 & b_2 & \dots & b_r \\ \dots & \dots & \dots & \dots \\ b_1 & b_2 & \dots & b_r \end{bmatrix} =$$

$$= [b_1 \quad b_2 \quad \dots \quad b_r].$$

This proves the theorem.

Definition 8.4.6 *The vector*

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_r \end{bmatrix}$$

is called the limit distribution *of the regular Markov chain.*

Remark 8.4.7 The limit distribution b is obviously connected with $\lim_{n\to\infty} T^n$ and $\lim_{n\to\infty} (p_n(1), \dots, p_n(r))$. A method for finding b is described in the next theorem.

Theorem 8.4.8 *Let T be the transition matrix of a regular Markov chain. Then 1 is an eigenvalue of the matrix T^t. The limit distribution b is the unique eigenvector of T^t associated with this eigenvalue, having nonnegative components whose sum equals 1.*

Proof. Since T is stochastic, it is easy to verify that 1 is an eigenvalue of T^t. On the other hand, from

$$\lim_{n\to\infty} T^n = B \text{ and } \lim_{n\to\infty} T^{n-1} \cdot T = B$$

we deduce that $B \cdot T = B$. This implies $T^t \cdot B^t = B^t$, i.e.,

$$T^t \cdot \begin{bmatrix} b_1 & b_1 & \dots & b_1 \\ b_2 & b_2 & \dots & b_2 \\ \dots & \dots & \dots & \dots \\ b_r & b_r & \dots & b_r \end{bmatrix} = \begin{bmatrix} b_1 & b_1 & \dots & b_1 \\ b_2 & b_2 & \dots & b_2 \\ \dots & \dots & \dots & \dots \\ b_r & b_r & \dots & b_r \end{bmatrix}.$$

It follows that

$$T^t \cdot \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_r \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_r \end{bmatrix},$$

i.e. $T^t \cdot b = b$. Thus the limit distribution b is an eigenvector of T^t associated with the eigenvalue 1.
Since B is stochastic, the components of b are nonnegative and have the sum equal to 1.

It remains to prove that b is the *unique* eigenvector of T^t associated with 1, having nonnegative components with sum 1.

Let

$$c = \begin{bmatrix} c_1 \\ c_2 \\ \dots \\ c_r \end{bmatrix}$$

be another vector with the same properties.

From $T^t c = c$ it follows by induction that $(T^t)^n c = c,\ n \geq 1$. Letting $n \to \infty$ we get $B^t c = c$, i.e.,

$$\begin{bmatrix} b_1 & b_1 & \dots & b_1 \\ b_2 & b_2 & \dots & b_2 \\ \dots & \dots & \dots & \dots \\ b_r & b_r & \dots & b_r \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \dots \\ c_r \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \dots \\ c_r \end{bmatrix}.$$

By hypothesis we have also $c_1 + \cdots + c_r = 1$, so that the above equation yields $b_1 = c_1, \dots,\ b_r = c_r$, i.e., $b = c$. This concludes the proof.

Definition 8.4.9 *A matrix T is called* doubly stochastic *if both T and T^t are stochastic matrices.*

Theorem 8.4.10 *If the transition matrix of a regular Markov chain is doubly stochastic, then the limit distribution is the vector*

$$\begin{bmatrix} \frac{1}{r} \\ \dots \\ \frac{1}{r} \end{bmatrix}.$$

Proof. According to Th.8.4.8, the limit distribution is the unique eigenvector of T^t associated with the eigenvalue 1, having nonnegative components with sum 1.

Since T^t is supposed to be stochastic, we apply Prop. 8.3.3 to conclude that it has the eigenvector

$$\begin{bmatrix} \frac{1}{r} \\ \dots \\ \frac{1}{r} \end{bmatrix}$$

associated with the eigenvalue 1. Thus

$$b = \begin{bmatrix} \frac{1}{r} \\ \dots \\ \frac{1}{r} \end{bmatrix},$$

and this ends the proof.

8.5 An example

Let us return to Example 8.1.9. The Markov chain presented there has the transition matrix

$$T = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ & \\ \frac{3}{4} & \frac{1}{4} \\ & \end{bmatrix}$$

and the initial probability vector $(\frac{2}{3}, \frac{1}{3})$. Obviously the Markov chain is regular. Then there exists the limit distribution

$$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

According to Th.8.4.8, b is uniquely determined by the conditions

(i) $$\begin{bmatrix} \frac{2}{3} & \frac{3}{4} \\ \frac{1}{3} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

(ii) $b_1 \geq 0,\ b_2 \geq 0,\ b_1 + b_2 = 1.$

Indeed, we get immediately $b_1 = \frac{9}{13}$ and $b_2 = \frac{4}{13}$.

A first conclusion is:

$$\lim_{n\to\infty} P(X(n) = w) = \frac{9}{13} \quad ; \lim_{n\to\infty} P(X(n) = r) = \frac{4}{13}.$$

If we want to know explicitly the probabilities $P(X(n) = w)$ and $P(X(n) = r)$, we need T^n.

The eigenvalues of T are $\lambda_1 = 1$ and $\lambda_2 = -\frac{1}{12}$, and the associated eigenvectors are

$$v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 4 \\ -9 \end{bmatrix}.$$

Denoting

$$P = \begin{bmatrix} 1 & 4 \\ 1 & -9 \end{bmatrix}$$

we have

$$T = P \cdot \begin{bmatrix} 1 & 0 \\ 0 & -\frac{1}{12} \end{bmatrix} \cdot P^{-1},$$

and consequently

$$T^n = P \cdot \begin{bmatrix} 1 & 0 \\ 0 & (-\frac{1}{12})^n \end{bmatrix} \cdot P^{-1}.$$

This yields

$$T^n = \frac{1}{13} \begin{bmatrix} 9 + 4(-\frac{1}{12})^n & 4 - 4(-\frac{1}{12})^n \\ 9 - 9(-\frac{1}{12})^n & 4 + 9(-\frac{1}{12})^n \end{bmatrix}.$$

Now we have

$$B = \lim_{n\to\infty} T^n = \frac{1}{13}\begin{bmatrix} 9 & 4 \\ 9 & 4 \end{bmatrix} = \begin{bmatrix} \dfrac{9}{13} & \dfrac{4}{13} \\ \dfrac{9}{13} & \dfrac{4}{13} \end{bmatrix},$$

from which we deduce again that $b_1 = \dfrac{9}{13}$, $b_2 = \dfrac{4}{13}$.
On the other hand,

$$[p_n(w) \quad p_n(r)] = [p_0(w) \quad p_0(r)] \cdot T^n = [\frac{2}{3} \quad \frac{1}{3}] \cdot T^n$$

It follows that the absolute probabilities are

$$P(X(n) = w) = p_n(w) = \frac{9}{13} - \frac{1}{39}\left(-\frac{1}{12}\right)^n,$$
$$P(X(n) = r) = p_n(r) = \frac{4}{13} + \frac{1}{39}\left(-\frac{1}{12}\right)^n, \quad n \geq 0.$$

8.6 Problems

1. Over a long period the weather on a certain island is either wet or fine (W or F) with day-to-day transition probabilities given by the matrix

$$T = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix},$$

i.e., $p(W, W) = 0.8$; $p(W, F) = 0.2$; $p(F, W) = 0.4$; $p(F, F) = 0.6$.

(i) If the weather is wet on Monday, what is the probability that it is wet on Wednesday of the same week?

(ii) If it is wet on Tuesday with probability 0.5, what is the probability that it is fine on Thursday of the same week?

2. Prove that the matrix

$$T = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

is regular and find $\lim_{n\to\infty} T^n$.

3. Find the limit distribution for the regular Markov chain having the transition matrix

$$T = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}.$$

4. An engineer either walks or uses his car. If he walks one day, the probability that he uses his car on the next is 0.4. The probability that he drives on two successive days is 0.8. If he walks on Sunday, what is the probability that he drives

   a) on Wednesday,

   b) on Friday,

   c) on Sunday ten weeks later?

CHAPTER 9

Hidden Markov Models

9.1 The elements of a hidden Markov model

Consider a homogeneous Markov chain with state space $\{1, 2, \ldots, r\}$. The initial probability vector will be denoted by

$$\pi = (\pi_1, \ldots, \pi_r) \quad , \pi_i = P(X(0) = i).$$

We shall denote the one-step transition probabilities $p(i, j)$ by a_{ij}:

$$a_{ij} = P(X(n+1) = j \mid X(n) = i), \quad i, j = 1, \ldots, r.$$

The transition matrix will be $A = (a_{ij})_{i,j=1,\ldots,r}$.

Let $\{v_1, \ldots, v_M\}$ be a set of symbols. Each state generates randomly a symbol. Let $b_j(k)$ be the probability that the state j generates the symbol $v_k, \quad j = 1, \ldots, r;\ k = 1, \ldots, M$. Consider the matrix

$$B = (b_j(k))_{j=1,\ldots,r;\, k=1,\ldots,M}.$$

☞ **Example 9.1.1** In a room there are r urns. Within each urn there are balls of colours $\{v_1, \ldots, v_M\}$. The probability of extracting a ball of colour v_k from the urn j is $b_j(k)$.

At the moment $n = 0$ an urn is chosen at random according to the initial probability vector π, and a ball is chosen at random from this urn. Denote by O_0 the observed color of the ball; then the ball is put into the urn from which it was selected.

At the moment $n = 1$ a new urn is chosen at random, according to the transition matrix A, and a ball is chosen at random from this urn; its observed color is denoted by O_1, and the process goes on.

We do not know the values of the parameters r, M, π, B, A; all that we know is the sequence of the observed symbols (in this case, colors) $O_0O_1O_2\dots$

Now we are in a position to describe the elements of a hidden Markov model.

a) The transition matrix $A = (a_{ij})_{i,j=1,\dots,r}$. It includes also the parameter r. Generally, A and r are unknown, but in some problems r may be specified.

b) The matrix $B = (b_j(k))_{j=1,\dots,r;\, k=1,\dots,M}$ describing the observation symbol probability distribution in state j. It includes the parameter M. As before, B and M are generally unknown, but sometimes M may be specified.

c) The unknown initial probability vector π.

The model is "hidden" in the sense that all that we know is the observed sequence of symbols $O_0O_1O_2\dots$, where $O_i \in \{v_1, \dots, v_M\},\ i = 0, 1, 2, \dots$

The elements of a hidden Markov model are denoted simply by $\lambda = (A, B, \pi)$.

9.2 Three problems for hidden Markov models

Problem 9.2.1 Suppose we are given the observation sequence $O = O_0O_1\dots O_T$. Then we choose a model $\lambda = (A, B, \pi)$. The problem is to compute $P(O \mid \lambda)$, that is, the conditional probability of the observation sequence O given the model λ.

This means to estimate the parameters (A, B, π) based upon observations. If $P(O \mid \lambda)$ is very small, the parameters of the model λ are not useful. Thus the solution to Problem 9.2.1 is important when we are trying to choose among several competing models; this solution allows us to choose the model which best matches the observations.

Problem 9.2.2 Given the observation sequence $O = O_0O_1\dots O_T$ and the model $\lambda = (A, B, \pi)$, we try to find the state sequence $X(0), X(1), \dots, X(T)$ which produced O. Of course, generally it is not possible to uncover the exact sequence of states; the problem is to find one which best matches the real state sequence, according to a given optimality criterion.

Problem 9.2.3 Given $O = O_0O_1\dots O_T$ and $\lambda = (A, B, \pi)$, we want to determine a method to optimize the parameters of the model, i.e., to adjust them in order to maximize the probability $P(O \mid \lambda)$. This means, after all, to describe in an optimal way how a given observation sequence comes about.

9.3 Solving Problem 9.2.1: Baum-Welch method

This means to compute $P(O \mid \lambda)$, i.e., the conditional probability of the observation sequence $O = O_0 O_1 \dots O_T$, given the model $\lambda = (A, B, \pi)$.

Fix a state sequence $Q = q_0 q_1 \dots q_T$, where $q_i \in \{1, \dots, r\}$, $i = 0, 1, \dots, T$. Then we have

$$P(O, Q \mid \lambda) = \pi_{q_0} b_{q_0}(O_0) a_{q_0 q_1} b_{q_1}(O_1) \dots a_{q_{T-1} q_T} b_{q_T}(O_T).$$

To compute this probability we need $2T + 1$ multiplications. There are r^{T+1} state sequences, and $P(O \mid \lambda)$ is the sum of all the terms of the above form, corresponding to all these r^{T+1} sequences.

In order to compute this sum we need $r^{T+1}(2T + 1)$ multiplications and $r^{T+1} - 1$ additions; a total of $2(T+1)r^{T+1} - 1$ operations. This is a very large number, even for small values of r and T. (For $r = 5$ and $T = 99$ we need about 10^{72} operations).

Fortunately a more efficient procedure exists; it is called the *forward-backward algorithm* or the *Baum-Welch method.*

Denote

$$\alpha_n(j) = P(O_0 O_1 \dots O_n\ ,\ X(n) = j \mid \lambda), \quad n = 0, 1, \dots,\ j = 1, 2, \dots, r.$$

Step 1. To compute $\alpha_0(j) = P(O_0, X(0) = j \mid \lambda) = \pi_j b_j(O_0)$, $j = 1, \dots, r$, we need r multiplications.

Step 2. Let us remark that

$$\begin{aligned} \alpha_{n+1}(j) &= P(O_0 O_1 \dots O_{n+1}\ ,\ X(n+1) = j \mid \lambda) = \\ &= P\left(\bigcup_{i=1}^{r} (O_0 O_1 \dots O_n,\ X(n) = i,\ X(n+1) = j,\ O_{n+1}) | \lambda \right) \\ &= \sum_{i=1}^{r} \alpha_n(i) a_{ij} b_j(O_{n+1}) = b_j(O_{n+1}) \sum_{i=1}^{r} \alpha_n(i) a_{ij}, \end{aligned}$$

for $0 \le n \le T - 1, \quad 1 \le j \le r$.

Given n and j, we need $r + 1$ multiplications and $r - 1$ additions in order to compute $\alpha_{n+1}(j)$; this gives $2r$ operations. Since n may assume T values and j may assume r values, the total number of operations requested at Step 2 is $2Tr^2$.

Step 3. We have

$$P(O \mid \lambda) = P\left(\bigcup_{j=1}^{r} (O_0 O_1 \dots O_T\ ,\ X(T) = j) \mid \lambda \right) = \sum_{j=1}^{r} \alpha_T(j).$$

Here we need $r-1$ additions.

So the total number of operations involved in calculating $P(O \mid \lambda)$ by the Baum-Welch method is $r + 2Tr^2 + r - 1 = 2Tr^2 + 2r - 1$. This number is much smaller than that requested by the direct method; for $r = 5, T = 99$, it is about 5000, versus 10^{72} for the direct method.

In order to solve Problem 9.2.2 and Problem 9.2.3 it is useful to introduce

$$\beta_n(i) = P(O_{n+1} \dots O_T \mid X(n) = i, \lambda)$$

for $n = T-1, T-2, \dots, 0;\ i = 1, \dots, r$.
The probability $\beta_n(i)$ can be efficiently computed in two steps.
Step 1. $\beta_T(i) = 1, \quad i = 1, \dots, r.$
Step 2.

$$\begin{aligned}
&\beta_n(i) = P(O_{n+1} \dots O_T \mid X(n) = i, \lambda) = \\
&= P\left(\bigcup_{j=1}^{r} (O_{n+1} \dots O_T,\ X(n+1) = j) \mid X(n) = i, \lambda\right) = \\
&= \sum_{j=1}^{r} P(O_{n+1} \dots O_T,\ X(n+1) = j \mid X(n) = i, \lambda) = \sum_{j=1}^{r} a_{ij} b_j(O_{n+1}) \beta_{n+1}(j).
\end{aligned}$$

9.4 Solving Problem 9.2.2: Viterbi algorithm

There are several possible ways of solving Problem 9.2.2, according to the chosen optimality criterion. We shall describe two such ways.

I.) Suppose that $\lambda = (A, B, \pi)$ and the observation sequence $O = O_0 O_1 \dots O_T$ are given. Given a fixed moment of time $n,\ 0 \le n \le T$, the optimality criterion is to choose the state which is most likely at moment n.

Since λ is fixed, we shall not mention it in the notation of conditional probabilities. Let

$$\gamma_n(j) = P(X(n) = j \mid O),\ j = 1, \dots, r.$$

Our problem is to find $\underset{1 \le j \le r}{argmax} \gamma_n(j)$. Denote by A_n the event $O_0 O_1 \dots O_n$ and by B_n the event $O_{n+1} \dots O_T$. Then $O = A_n \cap B_n$ and we have

$$\begin{aligned}
&\gamma_n(j) = P(X(n) = j \mid A_n \cap B_n) = P(A_n \cap (X(n) = j) \cap B_n)/P(O) = \\
&= P(A_n \cap (X(n) = j)) P(B_n \mid A_n \cap (X(n) = j))/P(O).
\end{aligned}$$

Due to the Markovian character of the process,

$$P(B_n) \mid A_n \cap (X(n) = j)) = P(B_n \mid X(n) = j) = \beta_n(j).$$

So we have

$$\gamma_n(j) = \alpha_n(j)\beta_n(j)/P(O).$$

Now let us remark that $P(O)$ is independent of j.

Indeed,

$$\sum_{i=1}^{r} \gamma_n(i) = 1,$$

which implies

$$\sum_{i=1}^{r} \alpha_n(i)\beta_n(i)/P(O) = 1,$$

i.e.,

$$P(O) = \sum_{i=1}^{r} \alpha_n(i)\beta_n(i).$$

We conclude that

$$\underset{1 \le j \le r}{argmax} \gamma_n(j) = \underset{1 \le j \le r}{argmax} \alpha_n(j)\beta_n(j).$$

Thus our problem is reduced to an optimization problem, namely to find $\underset{1 \le j \le r}{argmax} \alpha_n(j)\beta_n(j)$; this problem can be solved since we have efficient procedures to compute $\alpha_n(j)$ and $\beta_n(j)$.

II.) The most widely used optimality criterion is to find the best state sequence $Q = q_0 q_1 \ldots q_T$ for the given $\lambda = (A, B, \pi)$ and $O = O_0 O_1 \ldots O_T$. This can be done by using the *Viterbi algorithm*. We want to maximize $P(Q, \mid O, \lambda)$. Since $P(Q \mid O, \lambda) = P(Q,\ O|\lambda)/P(O \mid \lambda)$, our problem is to maximize $P(Q, O \mid \lambda)$.
Let us denote

$$\delta_n(i) = \max_{q_0, \cdots, q_{n-1}} P(q_0 \ldots q_{n-1}, q_n = i,\ O_0 \ldots O_n \mid \lambda).$$

Then we have

$$\delta_{n+1}(j) = \left(\max_{1 \le i \le r} \delta_n(i) a_{ij} \right) b_j(O_{n+1}).$$

Indeed,

$$\delta_{n+1}(j) = \max_{q_0,\cdots,q_n} P(q_0 \dots q_n, q_{n+1} = j,\ O_0 \dots O_{n+1} \mid \lambda) =$$

$$= \max_{1\leq i\leq r} \max_{q_0,\cdots,q_{n-1}} P(q_0 \dots q_{n-1}, q_n = i, q_{n+1} = j,\ O_0 \dots O_n, O_{n+1} \mid \lambda) =$$

$$= \max_{1\leq i\leq r} \max_{q_0,\cdots,q_{n-1}} P(q_0 \dots q_{n-1}, q_n = i,\ O_0 \dots O_n \mid \lambda) a_{ij} b_j(O_{n+1}) =$$

$$= \max_{1\leq i\leq r} \delta_n(i) a_{ij} b_j(O_{n+1}) = (\max_{1\leq i\leq r} \delta_n(i) a_{ij}) b_j(O_{n+1}).$$

The optimal state sequence $Q^* = q_0^* \dots q_T^*$ is that for which

$$P(Q^*, O \mid \lambda) = \max_Q P(Q, O \mid \lambda).$$

Remark that

$$\max_Q P(Q, O \mid \lambda) = \max_{1\leq i\leq r} \delta_T(i).$$

This means that $q_T^* = \underset{1\leq i\leq r}{argmax} \delta_T(i)$. Denoting by j this determined value, we have

$$\delta_T(j) = \max_{1\leq i\leq r} \delta_{T-1}(i) a_{ij}) b_j(O_{n+1}),$$

which entails

$$q_{T-1}^* = \underset{1\leq i\leq r}{argmax} \delta_{T-1}(i) a_{ij}.$$

By induction we find the values of $q_{T-2}^*, \dots, q_1^*$.

Denoting $q_1^* = j$ we have

$$\delta_1(j) = \left(\max_{1\leq i\leq r} \delta_0(i) a_{ij}\right) b_j(O_1)$$

and this yields

$$q_0^* = \underset{1\leq i\leq r}{argmax} \delta_0(i) a_{ij}.$$

On the other hand, by definition we have

$$\delta_0(i) = \pi_i b_i(O_0), \quad , i = 1, \dots, r.$$

Now the Viterbi algorithm can be completely described as follows:

1. Compute (by a forward procedure)

$$\delta_0(i) = \pi_i b_i(O_0), \quad i = 1, \dots, r,$$

$$\delta_n(j) = \left(\max_{1\leq i\leq r} \delta_{n-1}(i) a_{ij}\right) b_j(O_n), \quad n = 1, \dots, T; j = 1, \dots, r,$$

$$a_n(j) := \underset{1\leq i\leq r}{argmax} \delta_{n-1}(i) a_{ij}, \quad n = 1, \dots, T; j = 1, \dots, r.$$

2. Compute (by a backward procedure)

$$q_T^* = \underset{1 \leq i \leq r}{argmax} \delta_T(i),$$

$$q_T^* = a_{n+1}(q_{n+1}^*), \quad n = T-1, \ldots, 1, 0.$$

9.5 Solving Problem 9.2.3: Baum-Welch algorithm

Given the observation sequence $O = O_0 O_1 \ldots O_T$ we want to maximize $P(O|\lambda)$ as a function of λ.
We shall describe the iterative procedure of Baum-Welch (called also the *expectation-modification method)*; starting from a given model $\lambda = (A, B, \pi)$ one generates a sequence of models $\lambda_1, \lambda_2, \ldots$ whose limit is a point of local maximum for the function $P(O|\cdot)$.

Denote

$$Z_n(i,j) = P(X(n) = i,\ X(n+1) = j \mid O, \lambda),$$

for $n = 0, 1, 2, \ldots, T-1; \quad i, j = 1, \ldots, r$. Then we have

$$\begin{aligned} Z_n(i,j) =& \frac{P(X_n = i,\ X(n+1) = j,\ O|\lambda)}{P(O|\lambda)} = \\ &= \frac{\alpha_n(i) a_{ij} b_j(O_{n+1}) \beta_{n+1}(j)}{P(O|\lambda)}. \end{aligned}$$

Since $\sum\limits_{i=1}^{r} \sum\limits_{j=1}^{r} Z_n(i,j) = 1$, we deduce

$$P(O|\lambda) = \sum_{i,j=1}^{r} \alpha_n(i) a_{ij} b_j(O_{n+1}) \beta_{n+1}(j).$$

Thus we have:

$$Z_n(i,j) = \frac{\alpha_n(i) a_{ij} b_j(O_{n+1}) \beta_{n+1}(j)}{\sum_{i,j=1}^{r} \alpha_n(i) a_{ij} b_j(O_{n+1}) \beta_{n+1}(j)}.$$

Recall that

$$\gamma_n(i) = P(X(n) = i \mid O, \lambda).$$

Clearly

$$\gamma_n(i) = \sum_{j=1}^{r} Z_n(i,j).$$

Denote by $\nu_i(T)$ the number of times that the state i is visited. Then $\nu_i(T)$ is a random variable:

$$\nu_i(T) = \begin{pmatrix} 1 & 0 \\ \gamma_0(i) & 1-\gamma_0(i) \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ \gamma_1(i) & 1-\gamma_1(i) \end{pmatrix} + \cdots + \begin{pmatrix} 1 & 0 \\ \gamma_T(i) & 1-\gamma_T(i) \end{pmatrix}$$

with expectation

$$E\nu_i(T) = \sum_{n=0}^{T} \gamma_n(i).$$

This means that:

$$\sum_{n=0}^{T} \gamma_n(i) = \text{ The expected number of times in state } i.$$

The number of transitions from state i to state j is also a random variable τ_{ij}:

$$\tau_{ij} = \begin{pmatrix} 1 & 0 \\ Z_0(i,j) & 1-Z_0(i,j) \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ Z_1(i,j) & 1-Z_1(i,j) \end{pmatrix} + \\ \cdots + \begin{pmatrix} 1 & 0 \\ Z_{T-1}(i,j) & 1-Z_{T-1}(i,j) \end{pmatrix}.$$

Consequently,

$$E\tau_{ij} = \sum_{n=0}^{T-1} Z_n(i,j) = \text{ The expected number of transitions} \\ \text{from state } i \text{ to state } j.$$

The number of transitions from state i to an arbitrary state is a random variable μ_i:

$$\mu_i = \begin{pmatrix} 1 & 0 \\ \gamma_0(i) & 1-\gamma_0(i) \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ \gamma_1(i) & 1-\gamma_1(i) \end{pmatrix} + \\ \cdots + \begin{pmatrix} 1 & 0 \\ \gamma_{T-1}(i) & 1-\gamma_{T-1}(i) \end{pmatrix}$$

and thus

$E\mu_i = \sum\limits_{n=0}^{T-1} \gamma_n(i) =$ The expected number of transitions from state i.

Now we shall consider a new model $\lambda_1 = (\overline{A}, \overline{B}, \overline{\pi})$ where

$$\overline{\pi}_i = \gamma_0(i) = \text{expected number of times in state } i \text{ at time } n = 0,$$

$\overline{a_{ij}} = \left(\sum\limits_{n=0}^{T-1} Z_n(i,j)\right) / \left(\sum\limits_{n=0}^{T-1} \gamma_n(i)\right) =$ (expected number of transitions from i to j) / (expected number of transitions from i),

$\overline{b_j(k)} = \left(\sum\limits_{n=0,\ldots,T\ ;\ O_n=v_k} \gamma_n(j)\right) / \left(\sum\limits_{n=0}^{T} \gamma_n(j)\right) =$ (expected number of times in state j with symbol v_k) / (expected number of times in state j).

Now we pass from λ_1 to λ_2 in the same way we passed from λ to λ_1, and so on. It can be proved that the resulting sequence $\lambda_1, \lambda_2, \ldots$ has a limit which is a point of local maximum for the function $P(O|\cdot)$.

9.6 Problems

1. Evaluate the number of operations in computing $\beta_n(i)$, $n = 0, 1, \ldots, T$; $i = 1, \ldots, r$.
2. Give detailed probabilistic interpretations of the computations used in solving Problems 9.2.1, 9.2.2, 9.2.3.

CHAPTER 10

Brownian Motion. Itô Integral

10.1 Brownian Motion

A *Brownian Motion* is a family $B_t : \Omega \longrightarrow \mathbb{R},\ t \geq 0$, of random variables such that:

1) $B_0 = 0$;

2) $B_t - B_s$ is a normal variable $N(0, t-s)$, for all $t > s \geq 0$;

3) $B_{t_1}, B_{t_2} - B_{t_1}, \ldots, B_{t_k} - B_{t_{k-1}}$ are independent random variables for all $t_k > \cdots > t_1 \geq 0$;

4) $P\left(\{\omega \in \Omega \mid \text{the mapping } t \to B_t(\omega) \text{ is continuous on } [0, +\infty)\}\right) = 1$.

In particular, B_t is normal $N(0, t)$; it has the density

$$f(x) = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t},\ x \in \mathbb{R}.$$

It follows that $EB_t = 0$ and $VarB_t = t$.

The moments of odd orders vanish:

$$E(B_t^{2k+1}) = \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{+\infty} x^{2k+1} e^{-x^2/2t} dx = 0,\ k = 0, 1, 2, \ldots$$

Let us consider the moments of even orders:

$$E(B_t^{2k}) = \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{+\infty} x^{2k} e^{-x^2/2t} dx = \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{+\infty} x^{2k-1} (-te^{-x^2/2t})' dx$$

$$= \frac{1}{\sqrt{2\pi t}} (-tx^{2k-1} e^{-x^2/2t})|_{-\infty}^{+\infty} + (2k-1)t \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{+\infty} x^{2k-2} e^{-x^2/2t} dx =$$

$$= (2k-1)tE(B_t^{2k-2}).$$

So we have the recurrence formula

$$E(B_t^{2k}) = (2k-1)tE(B_t^{2k-2}),\ k = 1, 2, \ldots$$

which implies, in particular,

$$E(B_t^2) = t;\ E(B_t^4) = 3t^2.$$

Moreover, by similar considerations we find

$$E((B_t - B_s)^2) = t - s\ ;\ E((B_t - B_s)^4) = 3(t-s)^2.$$

10.2 The Itô integral

Let $X_t : \Omega \longrightarrow \mathbb{R},\ t \geq 0$, be random variables defined on the same probability space as B_t. We shall suppose that X_t depends only on the values $B_s,\ 0 \leq s \leq t$.

Let $\Delta : 0 = t_0 < t_1 < \cdots < t_k = T$ be a partition of the interval $[0, T]$. Denote $\|\Delta\| = \max\limits_{j=1,\ldots,k}(t_j - t_{j-1})$. Consider the "Riemann sum"

$$\sigma(\Delta) = \sum_{j=1}^{k} X_{t_{j-1}}(B_{t_j} - B_{t_{j-1}}).$$

It is important that the integrand is evaluated at the left-hand time point of the interval $[t_{j-1}, t_j]$ over which the increment $B_{t_j} - B_{t_{j-1}}$ is taken. With this restriction on how the approximating sums are set up, one can show that as the partition becomes finer (i.e., $\|\Delta\| \to 0$), the approximating sums $\sigma(\Delta)$ converge in mean square to a random variable, which we call the *Itô integral* and denote by $\int\limits_0^T X_t dB_t$.

☞ **Example 10.2.1** Let us compute $\int\limits_0^T B_t dB_t$.

First of all,

$$\sigma(\Delta) = \sum_{j=1}^{k} B_{t_{j-1}}(B_{t_j} - B_{t_{j-1}}).$$

Using the fact that

$$x(y-x) = \frac{1}{2}(y^2 - x^2 - (y-x)^2)$$

we get

$$\sigma(\Delta) = \frac{1}{2}\sum_{j=1}^{k}(B_{t_j}^2 - B_{t_{j-1}}^2) - \frac{1}{2}\sum_{j=1}^{k}(B_{t_j} - B_{t_{j-1}})^2 =$$
$$= \frac{1}{2}(B_T^2 - B_0^2) - \frac{1}{2}\sum_{j=1}^{k}(B_{t_j} - B_{t_{j-1}})^2.$$

We shall show that $\sum_{j=1}^{k}(B_{t_j} - B_{t_{j-1}})^2$ converges in mean square to the constant T, i.e.,

$$E(\sum_{j=1}^{k}(B_{t_j} - B_{t_{j-1}})^2 - T)^2 \longrightarrow 0 \text{ as } \|\Delta\| \to 0.$$

Indeed,

$$E(\sum_{j=1}^{k}(B_{t_j} - B_{t_{j-1}})^2 - T)^2 = E(\sum_{j}(B_{t_j} - B_{t_{j-1}})^4 +$$
$$+ 2\sum_{i<j}(B_{t_i} - B_{t_{i-1}})^2(B_{t_j} - B_{t_{j-1}})^2 - 2T\sum_{j}(B_{t_j} - B_{t_{j-1}})^2 + T^2).$$

Since $B_{t_i} - B_{t_{i-1}}$ and $B_{t_j} - B_{t_{j-1}}$ are independent, we get

$$E(\sum_{j=1}^{k}(B_{t_j} - B_{t_{j-1}})^2 - T)^2 = \sum_{j} 3(t_j - t_{j-1})^2 +$$
$$+ 2\sum_{i<j}(t_i - t_{i-1})(t_j - t_{j-1}) - 2T\sum_{j}(t_j - t_{j-1}) + T^2 =$$
$$= 2\sum_{j}(t_j - t_{j-1})^2 + (\sum_{j}(t_j - t_{j-1}))^2 - 2T^2 + T^2 =$$
$$= 2\sum_{j}(t_j - t_{j-1})^2 \leq 2\|\Delta\|\sum_{j}(t_j - t_{j-1}) = 2T\|\Delta\| \longrightarrow 0$$

as $\|\Delta\| \longrightarrow 0$.

Returning to $\sigma(\Delta)$ we see that

$$\sigma(\Delta) \longrightarrow \frac{1}{2}B_T^2 - \frac{1}{2}T$$

which means that

$$\int_0^T B_t dB_t = \frac{1}{2}B_T^2 - \frac{1}{2}T.$$

☞ **Example 10.2.2** Let us prove that

$$\int_0^T s dB_s = TB_T - \int_0^T B_s ds.$$

Indeed, we have

$$\sigma(\Delta) = \sum_{j=1}^{k} t_{j-1}(B_{t_j} - B_{t_{j-1}}) = \sum_{j=1}^{k}(t_j B_{t_j} - t_{j-1}B_{t_{j-1}}) -$$

$$- \sum_{j=1}^{k} B_{t_j}(t_j - t_{j-1}) = TB_T - \sum_{j=1}^{k} B_{t_j}(t_j - t_{j-1}) \longrightarrow$$

$$\longrightarrow TB_T - \int\limits_0^T B_s ds.$$

10.3 Problems

1. Give an intuitive interpretation of the properties of the family $(B_t)_{t\geq 0}$.
2. Find $E((B_t - B_s)^n)$.
3. Find the expected values and the variances of $\int\limits_0^T s dB_s$ and $\int\limits_0^T B_s ds$.

CHAPTER 11

Stochastic differential equations. Itô's formula

11.1 Definitions and main results

Let $\beta, \sigma : \mathbb{R} \longrightarrow \mathbb{R}$ be continuous functions. By a solution of the stochastic differential equation

$$dY_t = \beta(Y_t)dt + \sigma(Y_t)dB_t \quad ,t \geq 0 \tag{11.1.1}$$

we mean a family of random variables $(Y_t)_{t\geq 0}$ such that

$$Y_t = Y_0 + \int_0^t \beta(Y_s)ds + \int_0^t \sigma(Y_s)dB_s \quad ,t > 0. \tag{11.1.2}$$

Suppose that Y_t satisfies (11.1.1), or equivalently (11.1.2).

Let $f(t, y)$ be a continuous function with continuous partial derivatives f'_t, f'_y, f''_{yy}.

Itô's formula reads as follows:

$$df(t, Y_t) = f'_t(t, Y_t)dt + f'_y(t, Y_t)dY_t + \frac{1}{2}f''_{yy}(t, Y_t)dY_tdY_t \tag{11.1.3}$$

where we have to observe the following rules:

$$dtdt = 0;\ dtdB_t = dB_tdt = 0;\ dB_tdB_t = dt. \tag{11.1.4}$$

Consequently we have $dY_t dY_t = \sigma^2(Y_t)dt$ and Itô's formula (11.1.3) can be written as:

$$\begin{aligned} df(t,Y_t) &= f'_t(t,Y_t)dt + f'_y(t,Y_t)\beta(Y_t)dt+ \\ &+ f'_y(t,Y_t)\sigma(Y_t)dB_t + \frac{1}{2}f''_{yy}(t,Y_t)\sigma^2(Y_t)dt. \end{aligned}$$

Finally we get

$$df(t,Y_t) = \left(f'_t(t,Y_t) + f'_y(t,Y_t)\beta(Y_t) + \frac{1}{2}f''_{yy}(t,Y_t)\sigma^2(Y_t)\right)dt + f'_y(t,Y_t)\sigma(Y_t)dB_t. \tag{11.1.5}$$

Equivalently,

$$\begin{aligned} f(T,Y_T) = f(0,Y_0) &+ \int_0^T \left(f'_t(t,Y_t) + \beta(Y_t)f'_y(t,Y_t) + \frac{1}{2}\sigma^2(Y_t)f''_{yy}(t,Y_t)\right)dt+ \\ &+ \int_0^T f'_y(t,Y_t)\sigma(Y_t)dB_t, \quad T>0. \end{aligned} \tag{11.1.6}$$

11.2 Examples

☞ **Example 11.2.1** If $Y_t = B_t$ then $\beta = 0$ and $\sigma = 1$ are constant functions; (11.1.5) becomes

$$df(t,B_t) = \left(f'_t(t,B_t) + \frac{1}{2}f''_{yy}(t,B_t)\right)dt + f'_y(t,B_t)dB_t. \tag{11.2.1}$$

☞ **Example 11.2.2** If $f(t,y) = \varphi(y)$, (11.1.5) becomes

$$d\varphi(Y_t) = \left(\varphi'(Y_t)\beta(Y_t) + \frac{1}{2}\varphi''(Y_t)\sigma^2(Y_t)\right)dt + \varphi'(Y_t)\sigma(Y_t)dB_t. \tag{11.2.2}$$

☞ **Example 11.2.3** If $Y_t = B_t$ and $f(t,y) = \varphi(y)$, we get

$$d\varphi(B_t) = \varphi'(B_t)dB_t + \frac{1}{2}\varphi''(B_t)dt. \tag{11.2.3}$$

☞ **Example 11.2.4** In the context of the preceding example, let $\varphi(y) = y^2$. Then, according to (11.2.3),

$$dB_t^2 = 2B_t dB_t + dt. \tag{11.2.4}$$

The integral version of this equation is

$$B_T^2 = B_0^2 + 2\int_0^T B_t dB_t + \int_0^T dt.$$

This yields

$$\int_0^T B_t dB_t = \frac{1}{2}B_T^2 - \frac{1}{2}T,$$

which was also proved in Example 10.2.1.

☞ **Example 11.2.5** For $f(t, x, y)$, Itô's formula is:

$$df(t, X_t, Y_t) = f_t' dt + f_x' dX_t + f_y' dY_t + \frac{1}{2} f_{xx}'' dX_t dX_t + \\ + f_{xy}'' dX_t dY_t + \frac{1}{2} f_{yy}'' dY_t dY_t.$$

In particular, for $f(t, x, y) = xy$ we get

$$d(X_t Y_t) = Y_t dX_t + X_t dY_t + dX_t dY_t.$$

☞ **Example 11.2.6** Consider the stochastic differential equation

$$dY_t = \sigma(Y_t) dB_t + \frac{1}{2}\sigma(Y_t)\sigma'(Y_t) dt$$

where $\sigma \in C^1(I),\ \sigma > 0$ on the interval I.
Let $\varphi \in C^1(I)$ with $\varphi' = \dfrac{1}{\sigma}$ on I. Denote $X_t = \varphi(Y_t)$.

Then according to (11.2.2) with $\beta(y) = \frac{1}{2}\sigma(y)\sigma'(y)$,

$$dX_t = d\varphi(Y_t) = \frac{1}{\sigma}\sigma dB_t + \left(\frac{1}{\sigma}\frac{1}{2}\sigma\sigma' - \frac{1}{2}\frac{\sigma'}{\sigma^2}\sigma^2\right) dt,$$

that is, $dX_t = dB_t$. This implies $X_t = B_t + X_0$, i.e., $\varphi(Y_t) = B_t + \varphi(Y_0)$. We conclude that

$$Y_t = \varphi^{-1}(B_t + \varphi(Y_0)).$$

In particular, consider the problem

$$\begin{cases} dY_t = Y_t dB_t + \frac{1}{2} Y_t dt, \\ Y_0 = 1. \end{cases}$$

Then $\sigma(y) = y,\ y \in I = (0, +\infty),\ \varphi'(y) = \frac{1}{y},\ \varphi(y) = \log y$. We have $\varphi(Y_0) = 0$ and consequently $Y_t = exp(B_t),\ t \geq 0$.

☞ **Example 11.2.7** We shall prove that the solution of the stochastic problem

$$\begin{cases} dY_t = rY_t dt + \alpha Y_t dB_t, \\ Y_0 = y_0 \end{cases}$$

$(r, \alpha, y_0 \in \mathbb{R})$ is given by

$$Y_t = y_0 exp\left(\left(r - \frac{1}{2}\alpha^2\right) t + \alpha B_t\right),\ t \geq 0.$$

Indeed, $Y_t = f(t, B_t)$, where

$$f(t, y) = y_0 exp\left(\left(r - \frac{1}{2}\alpha^2\right) t + \alpha y\right).$$

According to (11.2.1),

$$\begin{aligned} dY_t = df(t, B_t) &= \left(\left(r - \frac{1}{2}\alpha^2\right) Y_t + \frac{1}{2}\alpha^2 Y_t\right) dt + \alpha Y_t dB_t = \\ &= rY_t dt + \alpha Y_t dB_t. \end{aligned}$$

11.3 Problems

Apply Itô's formula in order to compute:

1. dB_t^n.
2. $d\sin(tB_t)$.
3. $d(e^{B_t}\cos B_t)$.
4. $d(t^2 B_t \log(1 + B_t^2))$.
5. Let $X_t = re^{-rt}\int\limits_0^t e^{rs}dBs$. Prove that

$$X_T = rB_T - r\int_0^T X_s ds.$$

CHAPTER 12

The continuous Kalman filter

12.1 The filtering problem

For some systems the state $X_t \in \mathbb{R}$ at time t is subject to a differential equation

$$\frac{dX_t}{dt} = b(t, X_t), \quad t \geq 0.$$

But in many situations X_t is also subject to some random environmental effects, so that it satisfies an equation of the form

$$\frac{dX_t}{dt} = b(t, X_t) + \sigma(t, X_t)W_t$$

where $(W_t)_{t\geq 0}$ is a family of random variables representing the "noise".

Let us write this equation as

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)W_t dt.$$

From a mathematical point of view, as well as from many practical points of view, it is convenient to treat $W_t dt$ as being dB_t, where $(B_t)_{t\geq 0}$ is a Brownian Motion; thus we are lead to the stochastic equation

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t.$$

Now suppose that the state X_t is observed continuously and the observations $H_t \in \mathbb{R}$ are of the form

$$H_t = c(t, X_t) + \gamma(t, X_t)N_t$$

where N_t is a "noise" independent of B_t and X_0.

Define

$$Z_t = \int_0^t H_s ds, \quad t \geq 0.$$

Then $dZ_t = H_t dt = c(t, X_t)dt + \gamma(t, X_t)N_t dt$.

As before, $N_t dt$ is interpreted as being dV_t, where $(V_t)_{t \geq 0}$ is a Brownian Motion independent of B_t and X_0. We conclude that

$$\begin{cases} dZ_t = c(t, X_t)dt + \gamma(t, X_t)dV_t \\ Z_0 = 0. \end{cases}$$

In what follows Z_s are considered as our observations instead of H_t.

The *filtering problem* is to find the best estimate Y_t of X_t based on the observations Z_s, $0 \leq s \leq t$.

From an intuitive point of view, the problem is to filter the noise away from the observations in the best possible way.

Our optimality criterion will be to find Y_t whose value depends only on Z_s, $0 \leq s \leq t$, such that $E((X_t - Y_t)^2)$ is as small as possible.

12.2 The continuous linear filter

We consider the linear filtering problem, when the system equation is

$$dX_t = F(t)X_t dt + C(t)dB_t$$

and the observation equation:

$$dZ_t = G(t)X_t dt + D(t)dV_t, \quad Z_0 = 0.$$

Under suitable hypotheses it can be proved that the best estimate Y_t satisfies the stochastic differential equation

$$dY_t = \left(F(t) - \frac{G^2(t)S(t)}{D^2(t)}\right)Y_t dt + \frac{G(t)S(t)}{D^2(t)}dZ_t, \quad Y_0 = E(X_0),$$

where $S(t) = E((X_t - Y_t)^2)$ is the solution of the deterministic Riccati equation

$$S'(t) = -\frac{G^2(t)}{D^2(t)}S^2(t) + 2F(t)S(t) + C^2(t), \; S(0) = Var(X_0).$$

☞ **Example 12.2.1** (Noisy observations of a constant process.) Suppose that

$$dX_t = 0,\ E(X_0) = 0,\ E(X_0^2) = a^2,$$

$$dZ_t = X_t dt + m dV_t,\ Z_0 = 0,\ m = const. > 0.$$

Remark that in this case

$$H_t = \frac{dZ_t}{dt} = X_t + m\frac{dV_t}{dt} = X_t + mW_t, \quad W_t = \text{noise}.$$

We have $F(t) = C(t) = 0,\ G(t) = 1,\ D(t) = m$. The Riccati equation is

$$S'(t) = -\frac{1}{m^2}S^2(t), \quad S(0) = a^2.$$

In fact, this is an equation with separable variables:

$$\frac{dS}{S^2} = -\frac{1}{m^2}dt,$$

so that

$$\frac{1}{S} = \frac{1}{m^2}(t + c), \quad c = \text{const}.$$

From $S(0) = a^2$ we infer that $c = \frac{m^2}{a^2}$, i.e.,

$$S(t) = \frac{a^2 m^2}{a^2 t + m^2}, \quad t \geq 0.$$

It follows that Y_t satisfies the equation

$$dY_t = -\frac{a^2}{a^2 t + m^2}Y_t dt + \frac{a^2}{a^2 t + m^2}dZ_t, \quad Y_0 = 0.$$

The solution is

$$Y_t = \frac{a^2}{a^2 t + m^2}Z_t, \quad t \geq 0.$$

Indeed, according to Itô's formula (11.1.3),

$$\begin{aligned} dY_t = d\left(\frac{a^2}{a^2 t + m^2}Z_t\right) &= -\frac{a^4}{(a^2 t + m^2)^2}Z_t dt + \frac{a^2}{a^2 t + m^2}dZ_t \\ = -\frac{a^2}{a^2 t + m^2}Y_t dt + \frac{a^2}{a^2 t + m^2}dZ_t. \end{aligned}$$

☞ **Example 12.2.2** (Noisy observations of a Brownian Motion). Consider now the system equation

$$dX_t = cdB_t,\ c = \text{const.} > 0,\ E(X_0) = 0,\ E(X_0^2) = a^2$$

and the observation equation

$$dZ_t = X_t dt + mdV_t,\ m = \text{const.} > 0,\ Z_0 = 0.$$

The Riccati equation is

$$S'(t) = -\frac{1}{m^2}S^2(t) + c^2, \quad S(0) = a^2.$$

It can be written as

$$\frac{m^2 dS}{m^2c^2 - S^2} = dt$$

or, equivalently,

$$-m^2 \int \frac{dS}{S^2 - m^2c^2} = t + t_0, \quad t_0 = \text{const.}$$

This leads to

$$-m^2 \frac{1}{2mc} \log |\frac{S - mc}{S + mc}| = t + t_0,$$

that is,

$$\log |\frac{mc + S}{mc - S}| = \frac{2ct}{m} + t_1, \quad t_1 = \text{const.}$$

Finally we get

$$|\frac{mc + S}{mc - S}| = e^{2ct/m}K, \quad K = \text{const.}$$

Since $S(0) = a^2$, we have for $a^2 \neq mc$,

$$K = |\frac{mc + a^2}{mc - a^2}|.$$

(i) Let $a^2 < mc$. Then in a neighborhood of $t = 0$ we have

$$\frac{mc + S}{mc - S} = Ke^{2ct/m},$$

which implies

$$S(t) = mc\frac{Ke^{2ct/m} - 1}{Ke^{2ct/m} + 1}.$$

Since for this $S(t)$ we have obviously $S(t) < mc$, it will be the solution for all $t \geq 0$.

(ii) If $a^2 > mc$, we find analogously

$$S(t) = mc\frac{Ke^{2ct/m}+1}{Ke^{2ct/m}-1}, \quad t \geq 0.$$

(iii) For $a^2 = mc$ it is easy to verify that the Riccati equation has the solution $S(t) = mc,\ t \geq 0$.

Remark that in all the cases $\lim_{t\to\infty} S(t) = mc$.

For the sake of simplicity in what follows we take $a = 0$ and $m = c = 1$. Then $K = 1$ and

$$S(t) = \frac{e^{2t}-1}{e^{2t}+1} = \tanh t.$$

The best estimate Y_t satisfies

$$dY_t = -(\tanh t)Y_t dt + (\tanh t)dZ_t, \quad Y_0 = 0.$$

This can be written as

$$(\sinh t)Y_t dt + (\cosh t)dY_t = (\sinh t)dZ_t.$$

According to Itô's formula we have also

$$d((\cosh t)Y_t) = (\sinh t)Y_t dt + (\cosh t)dY_t.$$

Thus we arrive at

$$d((\cosh t)Y_t) = (\sinh t)dZ_t,$$

which implies

$$(\cosh t)Y_t - (\cosh 0)Y_0 = \int_0^t (\sinh s)dZ_s.$$

Since $Y_0 = 0$, this leads to

$$Y_t = \frac{1}{\cosh t}\int_0^t (\sinh s)dZ_s.$$

But for Z_t we have the interpretation

$$Z_t = \int_0^t H_s ds,$$

i.e., $dZ_t = H_t dt$. Thus the best estimation Y_t is given by

$$Y_t = \frac{1}{\cosh t}\int_0^t (\sinh s)H_s ds.$$

☞ **Example 12.2.3** (Estimation of a parameter).

Suppose that $X_t = X_0$ is independent of t; we want to estimate it based on observations Z_t satisfying

$$dZ_t = X_0 M(t)dt + N(t)dV_t, \quad Z_0 = 0.$$

Since $dX_t = 0$, we have $F(t) = C(t) = 0$; moreover, $G(t) = M(t)$ and $D(t) = N(t)$ are known functions.

The Riccati equation is

$$S'(t) = -\frac{M^2(t)}{N^2(t)} S^2(t), \quad S(0) = Var(X_0).$$

From

$$\frac{dS}{S^2} = -\frac{M^2(t)}{N^2(t)} dt$$

we infer

$$\frac{1}{S(t)} = \int_0^t \frac{M^2(s)}{N^2(s)} ds + \frac{1}{S(0)}.$$

The best estimate Y_t satisfies

$$dY_t = \frac{M(t)S(t)}{N^2(t)}(dZ_t - M(t)Y_t dt), \quad Y_0 = E(X_0).$$

By using Itô's formula we find

$$d\left(\frac{1}{S(t)} Y_t\right) = -\frac{S'(t)}{S^2(t)} Y_t dt + \frac{1}{S(t)} dY_t = \frac{M^2(t)}{N^2(t)} Y_t dt + \frac{1}{S(t)} dY_t =$$
$$= \frac{M^2(t)}{N^2(t)} Y_t dt + \frac{M(t)}{N^2(t)} dZ_t - \frac{M^2(t)}{N^2(t)} Y_t dt = \frac{M(t)}{N^2(t)} dZ_t.$$

Thus

$$d\left(\frac{1}{S(t)} Y_t\right) = \frac{M(t)}{N^2(t)} dZ_t,$$

which yields

$$\frac{1}{S(t)} Y_t - \frac{1}{S(0)} Y_0 = \int_0^t \frac{M(s)}{N^2(s)} dZ_s.$$

Finally we get

$$Y_t = \left(\frac{E(X_0)}{Var(X_0)} + \int_0^t \frac{M(s)}{N^2(s)} dZ_s\right) \Big/ \left(\frac{1}{Var(X_0)} + \int_0^t \frac{M^2(s)}{N^2(s)} ds\right).$$

☞ **Example 12.2.4** (Noisy observations of a population growth).
Consider the simple population growth model

$$dX_t = rX_t dt,\ r > 0,\ E(X_0) = b > 0,\ Var(X_0) = a^2.$$

Suppose that the observations satisfy

$$dZ_t = X_t dt + m dV_t,\ m = const. > 0,\ Z_0 = 0.$$

Thus $F(t) = r,\ C(t) = 0,\ G(t) = 1,\ D(t) = m$. The corresponding Riccati equation becomes

$$S'(t) = 2rS(t) - \frac{1}{m^2}S^2(t), \quad S(0) = a^2.$$

In fact, this is a Bernoulli equation.

Denoting $H(t) = \frac{1}{S(t)}$ we obtain the linear equation:

$$H' + 2rH = \frac{1}{m^2}$$

with solution

$$H(t) = Ke^{-2rt} + \frac{1}{2rm^2}, \quad K = const.$$

Then

$$S(t) = \frac{2rm^2}{1 + 2rm^2 Ke^{-2rt}}\ .$$

Since $S(0) = a^2$ we get

$$S(t) = \frac{2rm^2}{1 + e^{-2rt}\left(\frac{2rm^2}{a^2} - 1\right)}\ .$$

For the sake of simplicity we consider the case when $S(t)$ is constant; the value of the constant will be $S(0) = a^2$. On the other hand clearly we must have $2rm^2 = a^2$, and then indeed

$$S(t) = 2rm^2 = a^2.$$

The best estimate Y_t satisfies

$$dY_t = -rY_t dt + 2r dZ_t, \quad Y_0 = b.$$

But

$$\begin{aligned} d\left(e^{rt}Y_t\right) &= re^{rt}Y_t dt + e^{rt}dY_t = \\ &= e^{rt}\left(dY_t + rY_t dt\right) = 2re^{rt}dZ_t, \end{aligned}$$

i.e.,

$$d\left(e^{rt}Y_t\right) = 2re^{rt}dZ_t.$$

This implies:

$$e^{rt}Y_t - Y_0 = \int_0^t 2re^{rs}dZ_s.$$

Thus we obtain

$$Y_t = e^{-rt}\left(b + 2r\int_0^t e^{rs}dZ_s\right).$$

Since $Z_t = \int_0^t H_s ds$, we have also $dZ_t = H_t dt$, i.e.,

$$Y_t = e^{-rt}\left(b + 2r\int_0^t e^{rs}H_s ds\right).$$

(i) Suppose that $H_s = H_0,\ 0 \le s \le t$. Then

$$Y_t = be^{-rt} + 2H_0\left(1 - e^{-rt}\right).$$

For large t we have $Y_t \approx 2H_0$.

(ii) Suppose now that $H_s = H_0 e^{\alpha s}$. Then

$$Y_t = e^{-rt}\left(b + 2rH_0\frac{e^{(r+\alpha)t} - 1}{r+\alpha}\right) \approx \frac{2rH_0}{r+\alpha}e^{\alpha t}$$

for large t.

In particular, if $\alpha = r$ we have $Y_t \approx H_t$ for large t; if $\alpha = r$ and $H_0 = b$, then $Y_t = H_t$ for all $t \ge 0$.

12.3 Problems

1. Give intuitive interpretations of X_t, H_t, Z_t and the equations defining them.
2. Compare Example 12.2.1 with Example 5.2.1.
3. In Example 12.2.2 investigate different cases with respect to the parameters a, m, c.

CHAPTER 13

Gaussian families

13.1 Multidimensional normal variables

Let $f : \Omega \longrightarrow \mathbb{R}^n$ be a continuous random variable on the probability space $(\Omega, \mathscr{F}, P)$, with density $p : \mathbb{R}^n \longrightarrow \mathbb{R}$.
Let $x, y \in \mathbb{R}^n$,

$$x = \begin{pmatrix} x_1 \\ \dots \\ x_n \end{pmatrix}, \; y = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}.$$

We shall use the canonical inner-product in $\mathbb{R}^n$ defined by

$$\langle x, y \rangle = y^t x = x_1 y_1 + \dots + x_n y_n.$$

Definition 13.1.1 *The function* $\varphi_f : \mathbb{R}^n \longrightarrow \mathbb{C}$*,*

$$\varphi_f(s) = Ee^{i<s,f>} = \int_{\mathbb{R}^n} e^{i<s,x>} p(x)dx, \quad s \in \mathbb{R}^n,$$

is called the characteristic function *of* f*.*

Theorem 13.1.2 *Two random variables* $f, g : \Omega \longrightarrow \mathbb{R}^n$ *have the same distribution function if and only if they have the same characteristic function, i.e.,* $F_f = F_g \Longleftrightarrow \varphi_f = \varphi_g$*.*

Now let $B \in M_n(\mathbb{R})$ be a symmetric, positive definite matrix. Then:

(i) There exists a non singular matrix $A \in M_n(\mathbb{R})$ such that $B = A \cdot A^t$.

(ii) B has real positive eigenvalues $\lambda_1, \ldots, \lambda_n > 0$. The corresponding eigenvectors can be chosen to form an orthonormal basis $\{v_1, \ldots, v_n\}$. The eigenvalues of B^{-1} are $\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}$, and $B^{-1}v_i = \frac{1}{\lambda_i}v_i,\ i = 1, \ldots, n$.
If $x \in \mathbb{R}^n,\ x = x_1v_1 + \cdots + x_nv_n$, then

$$B^{-1}x = x_1\frac{1}{\lambda_1}v_1 + \cdots + x_n\frac{1}{\lambda_n}v_n,$$

so that $\langle B^{-1}x, x\rangle = \frac{x_1^2}{\lambda_1} + \cdots + \frac{x_n^2}{\lambda_n}$.

(iii) $detB = \lambda_1 \ldots \lambda_n$.

Definition 13.1.3 *Let $m \in \mathbb{R}^n$, and $B \in M_n(\mathbb{R})$ be symmetric and positive definite. We say that $f : \Omega \longrightarrow \mathbb{R}^n$ is a* normal variable with parameters m and B *if f has the density*

$$p(x) = \frac{1}{\sqrt{(2\pi)^n detB}}\ exp\left[-\frac{1}{2}\langle B^{-1}(x-m)\ ,\ x-m\rangle\right].$$

In this case we write $f \in N^n(m, B)$.

Let us find the characteristic function of the random variable $f \in N^n(m, B)$. Representing B as $A \cdot A^t$ we have $B^{-1} = (A^{-1})^t \cdot A^{-1}$. Then

$$\begin{aligned}
&\varphi_f(s) = Ee^{i<s,f>} = \int_{\mathbb{R}^n} e^{i<s,x>}p(x)dx = \\
&= \frac{1}{\sqrt{(2\pi)^n detB}}\int_{\mathbb{R}^n} exp\left[i\langle s, x\rangle - \frac{1}{2}\langle (A^{-1})^tA^{-1}(x-m),\ x-m\rangle\right]dx = \\
&\frac{1}{\sqrt{(2\pi)^n detB}}\int_{\mathbb{R}^n} exp\left[i\langle s, x\rangle - \frac{1}{2}\langle A^{-1}(x-m),\ A^{-1}(x-m)\rangle\right]dx.
\end{aligned}$$

Denote $y = A^{-1}(x-m)$; then $x = Ay + m$. The Jacobian $\dfrac{dx}{dy}$ is $detA$, so that $|\dfrac{dx}{dy}| = |detA| = \sqrt{detB}$. We deduce:

$$\begin{aligned}
\varphi_f(s) &= \frac{1}{\sqrt{(2\pi)^n}} \int_{\mathbb{R}^n} exp\left[i\langle s, Ay+m\rangle - \frac{1}{2}\langle y,y\rangle\right] dy = \\
&= \frac{e^{i<s,m>}}{\sqrt{(2\pi)^n}} \int_{\mathbb{R}^n} exp\left[-\frac{1}{2}\langle A^t s, A^t s\rangle - \frac{1}{2}\langle y - iA^t s, y - iA^t s\rangle\right] dy \\
&= (2\pi)^{-n/2} exp\left[i\langle s,m\rangle - \frac{1}{2}\langle A^t s, A^t s\rangle\right] \cdot \\
&\cdot \int_{\mathbb{R}^n} exp\left[-\frac{1}{2}\sum_{k=1}^{n}(y_k - i(A^t s)_k)^2 dy_1 \dots dy_k\right] = \\
&(2\pi)^{-n/2} exp\left[i\langle s,m\rangle - \frac{1}{2}\langle A^t s, A^t s\rangle\right] \cdot \prod_{k=1}^{n} \int_{\mathbb{R}} exp\left(-\frac{1}{2}u_k^2\right) du_k,
\end{aligned}$$

where $u_k = y_k - i(A^t s)_k$.

Since $\int_{\mathbb{R}} e^{-t^2/2} dt = \sqrt{2\pi}$, we get

$$\begin{aligned}
\varphi_f(s) &= exp\left[i\langle s,m\rangle - \frac{1}{2}\langle A^t s, A^t s\rangle\right] = \\
&= exp\left[i\langle s,m\rangle - \frac{1}{2}\langle AA^t s, s\rangle\right] = \\
&= exp\left[i\langle s,m\rangle - \frac{1}{2}\langle Bs, s\rangle\right].
\end{aligned}$$

So we have

Theorem 13.1.4 *$f \in N^n(m,B)$ if and only if*

$$\varphi_f(s) = exp\left[i\langle s,m\rangle - \frac{1}{2}\langle Bs,s\rangle\right], \ s \in \mathbb{R}^n.$$

With similar calculations it can be proved that

$$Ef = m, \quad C_f = B. \quad \text{(See Section 6.4).}$$

13.2 Independent normal variables

Theorem 13.2.1 *Let f and g be independent random variables, $f \in N^n(u,A)$, $g \in N^n(v,B)$. Then $f+g \in N^n(u+v, A+B)$.*

Proof. Let's compute the characteristic function of $f+g$. For $s \in \mathbb{R}^n$ we have

$$\begin{aligned}
&\varphi_{f+g}(s) = Ee^{i<s,f+g>} = E\left(e^{i<s,f>}e^{i<s,g>}\right) = \\
&= Ee^{i<s,f>}Ee^{i<s,g>} = \varphi_f(s)\varphi_g(s) = \\
&= e^{i<s,u>-\frac{1}{2}\langle As,s\rangle}e^{i<s,v>-\frac{1}{2}\langle Bs,s\rangle} = \\
&= e^{i<s,u+v>-\frac{1}{2}\langle (A+B)s,s\rangle}.
\end{aligned}$$

This means that $f+g \in N^n(u+v, A+B)$.

The same method of the characteristic function will be used in order to prove:

Theorem 13.2.2 *Let $f \in N^n(m,B)$ and $a \in \mathbb{R}^p, C \in M_{p,n}(\mathbb{R})$. Then $a + Cf \in N^p(a+Cm, CBC^t)$.*

Proof.

$$\begin{aligned}
&\varphi_{a+Cf}(s) = Ee^{i<s,a+Cf>} = \\
&= e^{i<s,a>}Ee^{i<s,Cf>} = e^{i<s,a>}Ee^{i<C^ts,f>} = \\
&= e^{i<s,a>}\varphi_f(C^ts) = e^{i<s,a>}e^{i<C^ts,m>-\frac{1}{2}\langle BC^ts,C^ts\rangle} \\
&= e^{i(<s,a>+<s,Cm>)-\frac{1}{2}\langle CBC^ts,s\rangle},
\end{aligned}$$

which proves the theorem.

It can be also proved that:

Theorem 13.2.3 *Let f and g be independent. Then $f \in N^n(a,A)$, $g \in N^p(b,B)$ iff*

$$\begin{pmatrix} f \\ g \end{pmatrix} \in N^{n+p}\left(\begin{pmatrix} a \\ b \end{pmatrix}, \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \right).$$

We know that if the real random variables X, Y are independent, then $C(X,Y) = 0$, i.e., X and Y are uncorrelated. The next result shows that for *normal* variables a converse holds true.

Theorem 13.2.4 *Let $f = \begin{pmatrix} f_1 \\ \dots \\ f_n \end{pmatrix}$ be a random variable, where $f_i : \Omega \longrightarrow \mathbb{R}$, $i = 1, \dots, n$. The following statements are equivalent:*

(1) $f_1, \dots, f_n$ are normal and independent;

(2) f is normal and $C(f_i, f_j) = 0,\ i \neq j$;

(3) $f \in N^n(m, B)$ with B diagonal.

Proof. (1)$\Longrightarrow$(2). Let $f_1, \dots, f_n$ be normal and independent. By Th.13.2.3, f is normal. Moreover, $C(f_i, f_j) = 0$ for $i \neq j$, since f_i and f_j are independent.
(2)$\Longrightarrow$ (3). If f is normal, then $f \in N^n(m, B)$ for some $m \in \mathbb{R}^n$ and $B = C_f$. Since $C(f_i, f_j) = 0$ for $i \neq j$, B is a diagonal matrix.
(3) $\Longrightarrow$(1).(Sketch) Let $f \in N^n(m, B)$ with B a diagonal matrix.

We have
$P(\{f_1 \in A_1\} \cap \dots \cap \{f_n \in A_n\}) = P(f \in A_1 \times \dots \times A_n) = \int\limits_{A_1 \times \dots \times A_n} p(x)dx$,
where p is the density of f. Since B is diagonal, it is not difficult to put the last integral under the form

$$\int_{A_1} p_1(x_1)dx_1 \dots \int_{A_n} p_n(x_n)dx_n,$$

where p_j is the density of $f_j,\ j = 1, \dots, n$.
This means that
$P(\{f_1 \in A_1\} \cap \dots \cap \{f_n \in A_n\}) = P(f_1 \in A_1) \dots P(f_n \in A_n)$, which shows that $f_1, \dots, f_n$ are independent.

Since $f \in N^n(m, B)$, with $B = diag(b_1, \dots, b_n)$, Th.13.2.3 shows that $f_i \in N^1(m_i, b_i),\ i = 1, \dots, n$.

13.3 Gaussian families

Definition 13.3.1 *Let T be an arbitrary set. A family $(X_t)_{t \in T}$ of random variables $X_t : \Omega \longrightarrow \mathbb{R}$ is called* Gaussian *if for any subset $\{t_1, \dots, t_n\} \subset T$, $n \geq 1$, the random variable $(X_{t_1}, \dots, X_{t_n})^t$ is normal.*

Theorem 13.3.2 *Let $f_1, \dots, f_n : \Omega \longrightarrow \mathbb{R},\ f = (f_1, \dots, f_n)^t$.*

a) Let $m = (m_1, \dots, m_n)^t \in \mathbb{R}^n$ and $A = (a_{ij})_{i,j=1,\dots,n}$. If $f \in N^n(m, A)$, then each linear combination $g = c_1 f_1 + \dots + c_n f_n$ is in $N(m_0, a_0)$, where $m_0 = \sum_{i=1}^n c_i m_i$ and $a_0 = \sum\limits_{i,j=1}^{n} c_i c_j a_{ij}$.

b) If each linear combination of $f_1, \dots, f_n$ is normal, then $f \in N^n(m, A)$, where $m_j = Ef_j$ and $a_{ij} = C(f_i, f_j),\ i, j = 1, \dots, n$.

c) If each linear combination of $f_1, \dots, f_n$ is normal, then the family $\{f_1, \dots, f_n\}$ is Gaussian.

d) If f is normal, then the family $\{f_1, \dots, f_n\}$ is Gaussian.

Proof. a) For $s \in \mathbb{R}^n$ we have $\varphi_f(s) = e^{i<m,s>-\frac{1}{2}\langle As,s\rangle}$, i.e.,

$$Eexp\left(i\sum_{k=1}^{n} s_k f_k\right) = exp\left[\left(i\sum_{k=1}^{n} s_k m_k\right) - \frac{1}{2}\sum_{k,l=1}^{n} a_{kl} s_k s_l\right].$$

Now, for $t \in \mathbb{R}$,

$$\varphi_g(t) = Ee^{itg} = Eexp\left[\left(\sum_{k=1}^{n} c_k f_k\right) it\right] =$$

$$= Eexp\left(i\sum_{k=1}^{n}(tc_k) f_k\right) = exp\left[i\sum_{k=1}^{n} tc_k m_k - \frac{1}{2}\sum_{k,l=1}^{n} a_{kl}(tc_k)(tc_l)\right] =$$

$$= exp\left[itm_0 - \frac{1}{2}t^2 a_0\right],$$

which means that $g \in N(m_0, a_0)$.

b) Suppose that each linear combination $g = \sum_{i=1}^{n} c_i f_i = \langle f, c\rangle$, $c \in \mathbb{R}^n$, is normal.
Then, for $t \in \mathbb{R}$ we have:

$$\varphi_g(t) = Ee^{itg} = e^{itEg - \frac{1}{2}t^2 Var(g)}.$$

On the other hand,

$$Eg = \sum_{i=1}^{n} c_i E f_i = \langle m, c\rangle,$$

$$Var(g) = E(g - Eg)^2 = E(\sum_{i=1}^{n} c_i (f_i - Ef_i))^2 =$$

$$= \sum_{i,j=1}^{n} c_i c_j E\left[(f_i - Ef_i)(f_j - Ef_j)\right] = \sum_{i,j=1}^{n} c_i c_j C(f_i, f_j) =$$

$$= \sum_{i,j=1}^{n} c_i c_j a_{ij} = \langle Ac, c\rangle.$$

Thus we infer:

$$\varphi_g(t) = exp\left[it\langle m, c\rangle - \frac{1}{2}t^2\langle Ac, c\rangle\right].$$

Now let us remark that

$$\varphi_f(c) = Ee^{i<f,c>} = Ee^{ig} = \varphi_g(1) = e^{i\langle m,c\rangle - \frac{1}{2}\langle Ac,c\rangle}.$$

This means that $f \in N^n(m, A)$.

c) Suppose that each linear combination of $f_1, \ldots, f_n$ is normal. Let $\{f_{j_1}, \ldots, f_{j_k}\} \subset \{f_1, \ldots, f_n\}$. Then each linear combination of $f_{j_1}, \ldots, f_{j_k}$ is a linear combination of $f_1, \ldots, f_n$, and hence is normal.
According to b), $(f_{j_1}, \ldots, f_{j_k})^t$ is normal. This means that $\{f_1, \ldots, f_n\}$ is Gaussian.

d) This is a consequence of a) and c).

An immediate consequence of Th.13.3.2 is

Corollary 13.3.3 *A family $(X_t)_{t\in T}$ is Gaussian if and only if each linear combination of variables from the family is normal.*

Definition 13.3.4 *Let $X_t : \Omega \longrightarrow \mathbb{R}, t \in T$, be an arbitrary family of random variables with $E(X_t^2) < \infty$, $t \in T$.*
The functions $m : T \longrightarrow \mathbb{R}$, $m(t) = EX_t$, respectively
$K : T \times T \longrightarrow \mathbb{R}$, $K(s,t) = C(X_s, X_t)$, are called the mean*, respectively the* covariance *of the family.*

Remark 13.3.5 K is symmetric (i.e., $K(s,t) = K(t,s)$) and positive semidefinite. Indeed, for $c_1, \ldots, c_n \in \mathbb{R}$ and $\{t_1, \ldots, t_n\} \subset T$ we have
$\sum\limits_{k,l=1}^{n} c_k c_l K(t_k, t_l) = \sum\limits_{k,l=1}^{n} c_k c_l E((X_{t_k} - EX_{t_k})((X_{t_l} - EX_{t_l}) =$
$= E\left(\sum\limits_{j=1}^{n} c_j (X_{t_j} - EX_{t_j})\right)^2 \geq 0.$

Theorem 13.3.6 *Let T be a set, $m : T \longrightarrow \mathbb{R}$ a function, and $K : T \times T \longrightarrow \mathbb{R}$ a symmetric, positive semidefinite function. Then there exists a probability space $(\Omega, \mathscr{F}, P)$ and a Gaussian family $(X_t)_{t\in T}$ on Ω such that m is the mean and K the covariance of this family.*

13.4 Problems

1. Prove assertions (i), (ii), (iii) in Section 13.1.
2. Prove that for f in Theorem 13.1.4 we have $Ef = m$, $C_f = B$.
   As in Definition 13.1.1, let $\varphi_X(s) = Ee^{isX}$ be the characteristic function of the random variable X. Find $\varphi_X(s)$ if X is

3. Uniformly distributed in $[a, b]$;
4. Binomial with parameters n and p;
5. Poisson with parameter λ;
6. Cauchy.

CHAPTER 14

Estimation. Bayes techniques

14.1 Unbiased estimates

Consider a random variable X whose probability density function depends on a parameter θ. Let $X_1, \ldots, X_n$ be a random sample of size n of X; then $X_1, \ldots, X_n$ are independent and have the same distribution as X. Let $t(X_1, \ldots, X_n)$ be a function, called a *statistic.*We shall use this function as an *estimate* (or *estimator*) of θ; that is, for a given set of observational values $x_1, \ldots, x_n$ the number $t(x_1, \ldots, x_n)$ will be considered as an estimate of θ.

Definition 14.1.1 *The statistic $t(X_1, \ldots, X_n)$ is called an unbiased estimate of θ if $Et = \theta$ for all θ.*

☞ **Example 14.1.2** Let μ be the expectation of X. The statistic

$$t(X_1, \ldots, X_n) = \frac{X_1 + \cdots + X_n}{n}$$

is an unbiased estimate of μ, since

$$Et = \frac{1}{n}(EX_1 + \cdots + EX_n) = \frac{1}{n}(n\mu) = \mu.$$

☞ **Example 14.1.3** Let μ be the expected value and σ^2 the variance of X. We have seen that

$$\overline{X} := \frac{X_1 + \cdots + X_n}{n}$$

is an unbiased estimator of μ. Define the *sample variance* by

$$S^2 := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

It is natural to consider S^2 as an estimator of σ^2. On the other hand,

$$\begin{aligned} E(S^2) &= \frac{1}{n} E \sum_{i=1}^{n} \left((X_i - \mu) - (\overline{X} - \mu) \right)^2 = \\ &= \frac{1}{n} E \left(\sum_{i=1}^{n} (X_i - \mu)^2 - 2(\overline{X} - \mu) \sum_{i=1}^{n} (X_i - \mu) + n(\overline{X} - \mu)^2 \right) = \\ &= \frac{1}{n} \sum_{i=1}^{n} E((X_i - \mu)^2 - \frac{2}{n} E \left((\overline{X} - \mu) n(\overline{X} - \mu) \right) + E(\overline{X} - \mu)^2 = \\ &= \frac{1}{n} n\sigma^2 - E(\overline{X} - \mu)^2 = \sigma^2 - Var\overline{X}. \end{aligned}$$

We have also (see Section 6.1)
$Var\overline{X} = Var\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n^2}(VarX_1 + \cdots + VarX_n) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$,
and so

$$E(S^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

This means that S^2 is not an unbiased estimator of σ^2. In order to overcome the bias in S^2 it is sufficient to multiply S^2 by $\frac{n}{n-1}$, since then

$$E\left(\frac{n}{n-1}S^2\right) = \frac{n}{n-1}E(S^2) = \sigma^2.$$

Thus the statistic

$$t(X_1, \ldots, X_n) := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is an unbiased estimator of σ^2.

14.2 Maximum likelihood estimators

Let $f(x;\theta)$ be the probability density function of the random variable X, where θ is the parameter to be estimated. Let $x_1, \ldots, x_n$ be random sample values of X. The function

$$L(x_1, \ldots, x_n; \theta) := \prod_{i=1}^{n} f(x_i; \theta)$$

is called the *likelihood function*.

Definition 14.2.1 *A* maximum likelihood *estimate of θ is an estimate that maximizes the function L as a function of θ.*

☞ **Example 14.2.2** Consider the exponential density

$$f(x;\theta) = \begin{cases} \theta e^{-\theta x} & , \ x > 0 \\ 0 & , \ x \leq 0. \end{cases}$$

The likelihood function is

$$L = \theta e^{-\theta x_1} \ldots \theta e^{-\theta x_n} = \theta^n e^{-\theta(x_1 + \cdots + x_n)}.$$

As a function of θ, L has a maximum for $\theta = \dfrac{n}{x_1 + \cdots + x_n}$. Thus the maximum likelihood estimator of θ is the statistic

$$\frac{n}{X_1 + \cdots + X_n}.$$

☞ **Example 14.2.3** Consider now the uniform density

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & , \ 0 \leq x \leq \theta, \\ 0 & , \ \text{otherwise.} \end{cases}$$

The likelihood function is here

$$L = \prod_{i=1}^{n} f(x_i, \theta) = \left(\frac{1}{\theta}\right)^n, \quad 0 \leq x_i \leq \theta, \ i = 1, \ldots, n.$$

It is maximized by $\theta = \max\{x_1, \ldots, x_n\}$, hence the maximum likelihood estimator is in this case the statistic $\max\{X_1, \ldots, X_n\}$.

14.3 Bayes techniques of estimation

Let X be a continuous random variable with density function $f(x;\theta)$. Let $X_1, \ldots, X_n$ be a random sample of size n and $t = t(X_1, \ldots, X_n)$ an estimate of θ based on this sample. We introduce a *cost function* (or *loss function*) $L(\theta, t)$ that measures the economic loss in claiming that the value of the parameter to be estimated is t when it is actually θ; usual cost functions are $L(\theta, t) = c|\theta - t|$ or $L(\theta, t) = c(\theta - t)^2$ for suitable constants c.

To decide whether $t = (X_1, \ldots, X_n)$ is a good estimator, one considers the expected value of the loss function:

$$R(\theta, t) = E(L(\theta, t)) = \int_{\mathbb{R}^n} L(\theta, t) \prod_{i=1}^{n} f(x_i;\theta) dx_1 \ldots dx_n.$$

$R(\theta, t)$ is called the *risk function.* An estimator that makes the risk small is considered a good estimator.

The Bayes approach is to introduce a probability density for the parameter θ and then calculate the expected value of $R(\theta, t)$ with respect to this density function.

Let $\Pi(\theta)$ denote the probability density; with respect to it, the expectation of $R(\theta, t)$ is

$$r(\Pi, t) := \int_{\mathbb{R}} R(\theta, t)\Pi(\theta)d\theta = \int_{\mathbb{R}^{n+1}} L(\theta, t) \prod_{i=1}^{n} f(x_i; \theta)\Pi(\theta)dx_1 \ldots dx_n d\theta.$$

The number $r(\Pi, t)$ is called the *mean risk.*
Now the problem is to find an estimating function $t(x_1, \ldots, x_n)$ that minimizes the mean risk; if such a function exists, it is called the *Bayes solution* to the estimation problem corresponding to the density function $\Pi(\theta)$.

☞ **Example 14.3.1** Suppose that only one observation of X is made, and the loss function is $L(\theta, t) = (t - \theta)^2$. Then the mean risk is

$$r(\Pi, t) = \iint_{\mathbb{R}^2} (t - \theta)^2 f(x; \theta)\Pi(\theta)dxd\theta.$$

Let $g(x, \theta)$ be the joint density of the random variables X and θ. Since $f(x; \theta)$ is the conditional density function of X with fixed θ, we have

$$f(x; \theta) = \frac{g(x, \theta)}{\Pi(\theta)}.$$

Thus $g(x, \theta) = \Pi(\theta)f(x; \theta)$. On the other hand,

$$g(x, \theta) = h(x)g(\theta \mid x)$$

where $h(x) = \int_{\mathbb{R}} g(x, \theta)d\theta$ is the marginal density function of X, and $g(\theta|x)$ is the conditional density function of θ with fixed X.

Consequently the mean risk can be written as
$r(\Pi, t) = \iint\limits_{\mathbb{R}^2} (t - \theta)^2 g(x; \theta)dxd\theta = \iint\limits_{\mathbb{R}^2} (t - \theta)^2 h(x)g(\theta|x)dxd\theta =$
$= \int\limits_{\mathbb{R}} h(x) \left(\int\limits_{\mathbb{R}} (\theta - t)^2 g(\theta|x)d\theta \right) dx.$

Let x be fixed. The derivative of the function $t \longrightarrow \int\limits_{\mathbb{R}} (\theta - t)^2 g(\theta|x)d\theta$ is $-2\int\limits_{\mathbb{R}} (\theta - t)g(\theta|x)d\theta$. This derivative vanishes for

$$t = \int_{\mathbb{R}} \theta g(\theta|x)d\theta = E(\theta|x).$$

It is easy to see that this t minimizes the function $\int\limits_{\mathbb{R}}(\theta - t)^2 g(\theta|x)d\theta$; consequently, $t(x) = E(\theta|x)$ minimizes the mean risk $r(\Pi, t)$.
Thus we have:

Theorem 14.3.2 *Under the above assumptions, the Bayes estimate of θ is $E(\theta|x)$, where x is the observed value of X.*

Now the problem is to find the conditional expected value of the parameter θ when X is fixed.

☞ **Example 14.3.3** Let $\alpha > 0$, $\beta > 0$, $a \in \mathbb{R}$ be given. Suppose that the random variable X is normally distributed with parameters θ and α, i.e.,

$$f(x;\theta) = \frac{1}{\alpha\sqrt{2\pi}} e^{-(x-\theta)^2/2\alpha^2}, \quad x \in \mathbb{R}.$$

Moreover, suppose that the parameter θ which will be estimated is a normal variable with parameters a and β, i.e.,

$$\Pi(\theta) = \frac{1}{\beta\sqrt{2\pi}} e^{-(\theta-a)^2/2\beta^2}.$$

As in Example 14.3.1, the joint density of X and θ is

$$g(x,\theta) = \Pi(\theta) f(x;\theta) = \frac{1}{2\pi\alpha\beta} e^{-(x-\theta)^2/2\alpha^2 - (\theta-a)^2/2\beta^2}.$$

The marginal density function $h(x)$ is given by

$$h(x) = \int_{\mathbb{R}} g(x,\theta)d\theta = \frac{1}{\sqrt{2\pi(\alpha^2+\beta^2)}} e^{-(x-a)^2/2(\alpha^2+\beta^2)}.$$

Consequently,

$$g(\theta|x) = \frac{g(x,\theta)}{h(x)} = \frac{\sqrt{\alpha^2+\beta^2}}{\alpha\beta\sqrt{2\pi}} exp\left(-\frac{\alpha^2+\beta^2}{2\alpha^2\beta^2}\left(\theta - \frac{\beta^2 x + \alpha^2 a}{\alpha^2+\beta^2}\right)^2\right).$$

This is a normal density with parameters

$$\frac{\beta^2 x + \alpha^2 a}{\alpha^2+\beta^2} \quad \text{and} \quad \frac{\alpha\beta}{\sqrt{\alpha^2+\beta^2}}.$$

The corresponding expected value is

$$E(\theta|x) = \frac{\beta^2 x + \alpha^2 a}{\alpha^2+\beta^2}.$$

We conclude that the Bayes solution for this estimation problem is

$$t(X) = \frac{\beta^2 X + \alpha^2 a}{\alpha^2+\beta^2}.$$

14.4 Problems

1. Given the probability density function

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} \quad , \ x \in \mathbb{R},$$

find the maximum likelihood estimator of θ based on a sample of size n.

2. For $\theta > -1$ consider the density

$$f(x;\theta) = \begin{cases} (1+\theta)x^{\theta} & , \ 0 < x < 1 \\ 0 & , \ \text{otherwise.} \end{cases}$$

Find the maximum likelihood estimate based on n observations.

3. A random variable has a normal distribution with parameters 0 and θ. Find the maximum likelihood estimator based on a sample of size n.

4. Let $L(\theta, t) = |\theta - t|$,

$$f(x;\theta) = \begin{cases} \theta e^{-\theta x} & , \ x > 0 \\ 0 & , \ \text{otherwise} \end{cases} \quad , \quad \Pi(\theta) = \begin{cases} e^{-\theta} & , \ \theta > 0 \\ 0 & , \ \text{otherwise.} \end{cases}$$

   i) Determine the risk function and the mean risk for the estimator $t = X$.

   ii) The same problem when $t = 2X$.

5. Consider a random variable X with density

$$f(x;\theta) = \begin{cases} \frac{\theta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\theta x} & , \ x > 0 \\ 0 & , \ x \le 0, \end{cases}$$

where $\alpha > 0$ and $\theta > 0$. Determine the Bayes solution for estimating the parameter θ if θ possesses the density

$$\Pi(\theta) = \frac{\beta^{a}}{\Gamma(a)} \theta^{a-1} e^{-\beta\theta} \quad , \beta > 0, \ a > 0.$$

CHAPTER 15

Testing statistical hypotheses

15.1 Statistical hypotheses. Tests

Let X be a random variable with density $f(x;\theta)$ depending on an (unknown) parameter θ. An assertion of type $\theta = \theta_0$ about the value of the parameter will be called a *statistical hypothesis*.

For example, consider a density of the form

$$f(x;\theta) = \begin{cases} \theta e^{-\theta x} & , x > 0 \\ 0 & , x \leq 0. \end{cases}$$

The assertion $\theta = 3$ is a statistical hypothesis.
A *test* of a statistical hypothesis is a procedure for deciding whether to accept or reject it.

In the above example suppose that θ can assume only the value 3 or the value 1. Denote by H_0 the hypothesis $\theta = 3$ and by H_1 the hypothesis $\theta = 1$.

Our problem is to test the hypothesis H_0 against the alternative hypothesis H_1. For the sake of simplicity, suppose that a single observation is made on the random variable X.

To construct a test, we have to decide what values of X should be selected for accepting H_0 and what values for rejecting H_0 (and thus accepting H_1). The rejection values form the *critical region* of the test.

In our example, let $(2, +\infty)$ be the critical region. This is an arbitrary choice; is it a wise one? Let us consider its consequences.

Denote by x the result of the observation on X.

1. If H_0 is actually true and $x > 2$, H_0 will be incorrectly rejected. This is called an error of type I. The *size* of the type I error is the probability that the sample point will fall in the critical region when H_0 is true.This probability is denoted by α and is called also *the size of the critical region.*

   In our example,
$$\alpha = \int_2^\infty 3e^{-3x}dx = e^{-6}.$$

2. If H_0 is actually false and $x \leq 2$, H_0 will be incorrectly accepted. This is an error of type II. The size of it is the probability (denoted by β) that the sample point will fall in the noncritical region when H_0 is false.

   In our example,
$$\beta = \int_0^2 e^{-x}dx = 1 - e^{-2}.$$

3. If H_0 is true and $x \leq 2$, the decision of accepting H_0 is correct; the corresponding probability is $1 - \alpha$.

4. If H_0 is false and $x > 2$, the decision of rejecting H_0 is correct; the probability is $1 - \beta$.

In order to determine good tests, the following simple principle is often applied: among all tests possessing the same size α, choose one for which β is as small as possible.

Returning to the above example, let $a = -\frac{1}{3}\log(1 - e^{-6})$. Consider a new test, with $(0, a)$ as critical region. The size of this critical region is

$$\alpha_1 = \int_0^a 3e^{-3x}dx = 1 - e^{-3a} = e^{-6}.$$

Thus both tests have the same α. For the new test the size of the type II error is
$$\beta_1 = \int_a^\infty e^{-x}dx = e^{-a} = (1 - e^{-6})^{1/3}.$$

Since $\beta < \beta_1$, the old test is better than the new one.

15.2 Best tests. Neyman-Pearson Lemma

Among all tests having the same size α of the critical region, a best test is defined as a test that minimizes the size β of the type II error. The Neyman-Pearson Lemma can be used to construct such a best test.

Let $f(x;\theta)$ be the density of a random variable X, and $x_1,\dots,x_n$ the values of a random sample of size n of X. We want to test the hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta = \theta_1$. We choose a critical region $A \subset \mathbb{R}^n$; this means that if $(x_1,\dots,x_n) \in A$, the hypothesis H_0 will be rejected (and, consequently, H_1 will be accepted).

Theorem 15.2.1 *(Neyman-Pearson Lemma). Suppose that $A \subset \mathbb{R}^n$ is a critical region of size α, and there exists a constant c such that*

(i) $\dfrac{f(t_1,\theta_1)\dots f(t_n,\theta_1)}{f(t_1,\theta_0)\dots f(t_n,\theta_0)} > c \quad , \ (t_1,\dots,t_n) \in A$

(ii) $\dfrac{f(t_1,\theta_1)\dots f(t_n,\theta_1)}{f(t_1,\theta_0)\dots f(t_n,\theta_0)} \le c \quad , \ (t_1,\dots,t_n) \in \mathbb{R}^n \diagdown A.$

Then A is a best critical region of size α.

In order to illustrate the usefulness of this result, consider the following example.

☞ **Example 15.2.2** Let X be a random variable with density

$$f(x;\theta) = \begin{cases} \theta e^{-\theta x} & , \ x > 0, \\ 0 & , \ x \le 0. \end{cases}$$

Let $0 < \theta_1 < \theta_0$. We want to test the hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta = \theta_1$, based on a random sample of size n.

According to Theorem 15.2.1, A is the region in which

$$\frac{\theta_1^n e^{-\theta_1(t_1+\cdots+t_n)}}{\theta_0^n e^{-\theta_0(t_1+\cdots+t_n)}} > c.$$

This inequality is equivalent to

$$t_1 + \cdots + t_n > \frac{1}{\theta_0 - \theta_1} \log\left(c\left(\frac{\theta_0}{\theta_1}\right)^n\right).$$

Thus

$$A = \left\{(t_1,\dots,t_n) \in \mathbb{R}^n : t_1,\dots,t_n \ge 0, \ t_1 + \cdots + t_n > \frac{1}{\theta_0 - \theta_1} \log\left(c\left(\frac{\theta_0}{\theta_1}\right)^n\right)\right\}.$$

The size of this critical region is

$$\int_A \theta_0^n e^{-\theta_0(x_1+\cdots+x_n)} dx_1 \dots dx_n.$$

Given $\alpha \in (0,1)$, we want the above integral to be equal to α; this provides an equation for determining c. With the corresponding value of c, the best critical region A is completely determined.

For the concrete example examined in the preceding section, $\theta_0 = 3$, $\theta_1 = 1$, $n = 1$. We infer that $A = \left(\frac{1}{2}\log 3c, +\infty\right)$. Given $\alpha = e^{-6}$, we have

$$\int_{\frac{1}{2}\log 3c}^{\infty} 3e^{-3x} dx = e^{-6},$$

which entails $3c = e^4$, and finally $A = (2, +\infty)$.

15.3 Problems

1. Let X be normally distributed with mean 0 and variance σ^2. Determine the nature of a best critical region for testing the hypothesis $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma = \sigma_1 > \sigma_0$, based on a random sample of size n.

2. Consider the density

$$f(x;\theta) = \begin{cases} (1+\theta)x^{\theta} & , \ 0 < x < 1, \\ 0 & , \ \text{otherwise.} \end{cases}$$

Determine the nature of a best critical region based on a random sample of size n for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 < \theta_0$.

CHAPTER 16

Financial Mathematics: The Black-Scholes equation

16.1 Option pricing

Consider a riskless security - a bond, or a savings bank account-with price X_0 at time $t = 0$ and $X(t)$ at time t. If interest is paid continuously at a constant rate r, then

$$dX(t) = rX(t)dt \tag{16.1.1}$$

which leads to

$$X(t) = X_0 e^{rt}, \quad t \geq 0. \tag{16.1.2}$$

Now consider a risky asset, for example a share with stock price S_t satisfying the stochastic equation

$$dS_t = \mu S_t dt + \sigma S_t dB_t$$

where B_t is a Brownian Motion, and $\mu > 0$, $\sigma > 0$ are real numbers called *the mean of returns*, respectively *the volatility* for investing in the stock.

From Example 11.2.7 we know that

$$S_t = S_0 exp\left((\mu - \frac{1}{2}\sigma^2)t + \sigma B_t\right). \tag{16.1.3}$$

A *call option* is an option to buy a share at a certain time T, for a certain price K. If $S_T > K$, the option is exercised, making a profit of $S_T - K$; if $S_T \leq K$, the option is not exercised, and the holder of it receives 0.

So, if we denote by Y_t the price of the option, then

$$Y_T = (S_T - K)^+. \tag{16.1.4}$$

An option to sell a share at time T, for a price K, is called a *put option*; its price Z_t satisfies

$$Z_T = (K - S_T)^+. \tag{16.1.5}$$

Option pricing means to determine Y_0 and Z_0. The starting point of this theory was the dissertation of Louis Bachelier, Théorie de la Spéculation, Paris, 1900.

A widely accepted solution of the option pricing problem was given in 1973 by the Black-Scholes formula. For their contributions to this domain, Myron Scholes and Robert Merton were awarded the Nobel prize for economics in 1997; sadly, Fisher Black had passed away in 1995.

16.2 The Black-Scholes equation

Consider Y_t as a function of S_t and t:

$$Y_t = u(S_t, t), \quad t \geq 0. \tag{16.2.1}$$

At time t an investor sells a call option and buys M shares, so that the value of his portfolio at time t is

$$P_t = MS_t - Y_t. \tag{16.2.2}$$

The change in the portfolio over the time interval $(t, t + dt)$ is

$$dP_t = MdS_t - dY_t. \tag{16.2.3}$$

Using Itô's formula we deduce

$$\begin{aligned} dP_t &= MdS_t - u'_s dS_t - u'_t dt - \frac{1}{2} u''_{ss} dS_t dS_t = \\ &= (M - u'_s) dS_t - (u'_t + \frac{1}{2}\sigma^2 S_t^2 u''_{ss}) dt. \end{aligned}$$

If we take

$$M = u'_s, \tag{16.2.4}$$

we get

$$dP_t = -\left(u'_t + \frac{1}{2}\sigma^2 S_t^2 u''_{ss}\right) dt. \tag{16.2.5}$$

According to (16.2.5), the portfolio P_t has no volatility over the interval $(t, t + dt)$; hence it must obey a law similar to (16.1.1), i.e.,

$$dP_t = rP_t dt. \tag{16.2.6}$$

Indeed, otherwise there would be an opportunity for making money at no risk (an *arbitrage opportunity*): comparing the rates of return for $X(t)$ and P_t, it would be possible to borrow money at the cheapest rate and lend it out at the more expensive one.

From (16.2.5) and (16.2.6) we deduce

$$\frac{1}{2}\sigma^2 S_t^2 u''_{ss} + u'_t = -rP_t. \tag{16.2.7}$$

Now (16.2.7) and (16.2.2) yield

$$u'_t = -\frac{1}{2}\sigma^2 S_t^2 u''_{ss} - rMS_t + rY_t. \tag{16.2.8}$$

Taking into account (16.2.1) and (16.2.4), we get

$$u'_t(S_t, t) = -\frac{1}{2}\sigma^2 S_t^2 u''_{ss}(S_t, t) - rS_t u'_s(S_t, t) + ru(S_t, t). \tag{16.2.9}$$

According to (16.1.4), we have also

$$u(S_T, T) = (S_T - K)^+. \tag{16.2.10}$$

From (16.2.9) and (16.2.10) we deduce that the function $u(x, t)$ satisfies:

$$u'_t(x, t) = -\frac{1}{2}\sigma^2 x^2 u''_{xx}(x, t) - rxu'_x(x, t) + ru(x, t), \tag{16.2.11}$$

$$u(x, T) = (x - K)^+. \tag{16.2.12}$$

(16.2.11) is called the *Black-Scholes equation.* It can be proved that the solution of the problem (16.2.11)-(16.2.12) is the function

$$\begin{aligned} u(x, t) = xF&\left(\frac{\log(x/K) + (r + \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}}\right) - \\ - Ke^{-r(T-t)}F&\left(\frac{\log(x/K) + (r - \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}}\right), \end{aligned} \tag{16.2.13}$$

where

$$F(x) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2} dy = \frac{1}{2} + \Phi(x). \tag{16.2.14}$$

(see Example 2.3.4).

As mentioned before, we are interested in pricing the call option at time $t = 0$.

So we have

$$\begin{aligned} Y_0 = & u(S_0, 0) = S_0 F\left(\frac{\log(S_0/K) + (r + \sigma^2/2)T}{\sigma\sqrt{T}}\right) - \\ & - Ke^{-rT} F\left(\frac{\log(S_0/K) + (r - \sigma^2/2)T}{\sigma\sqrt{T}}\right). \end{aligned} \tag{16.2.15}$$

(16.2.15) is called the *Black-Scholes formula.* It is responsible for a huge amount of options traded on stock exchanges all over the world.

Now let

$$Z_t = v(S_t, t), \quad t \geq 0$$

be the price of a put option.

By similar considerations we find that the function $v(x,t)$ is the solution of the problem

$$v'_t(x,t) = -\frac{1}{2}\sigma^2 x^2 v''_{xx}(x,t) - rxv'_x(x,t) + rv(x,t), \tag{16.2.16}$$

$$v(x,T) = (K - x)^+. \tag{16.2.17}$$

It can be proved that this solution is

$$\begin{aligned} v(x,t) = & Ke^{-r(T-t)} F\left(-\frac{\log(x/K) + (r - \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}\right) - \\ & - xF\left(-\frac{\log(x/K) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}\right). \end{aligned} \tag{16.2.18}$$

Consequently, the price of a put option at $t = 0$ is

$$\begin{aligned} Z_0 = & v(S_0, 0) = Ke^{-rT} F\left(-\frac{\log(x/K) + (r - \sigma^2/2)T}{\sigma\sqrt{T}}\right) - \\ & - S_0 F\left(-\frac{\log(x/K) + (r + \sigma^2/2)T}{\sigma\sqrt{T}}\right). \end{aligned} \tag{16.2.19}$$

16.3 Call-Put parity: direct proofs and financial proof

We know that the price of a call option at time $t \in [0,T]$ is

$$Y_t = u(S_t, t) \tag{16.3.1}$$

and the price of a put option is

$$Z_t = v(S_t, t). \tag{16.3.2}$$

We shall prove that

$$Y_t - Z_t = S_t - Ke^{-r(T-t)}. \tag{16.3.3}$$

This is called *the call-put parity formula.*

A first direct proof of it can be given by using (16.3.1), (16.3.2), (16.2.13) and (16.2.18); this is the content of Problem 1 below.

A second proof runs as follows. Let

$$w(x,t) := u(x,t) - v(x,t). \tag{16.3.4}$$

From (16.2.11) and (16.2.16) we deduce that

$$w_t'(x,t) = -\frac{1}{2}\sigma^2 x^2 w_{xx}''(x,t) - rxw_x'(x,t) + rw(x,t). \tag{16.3.5}$$

On the other hand, (16.2.12) and (16.2.17) yield:

$$w(x,T) = x - K. \tag{16.3.6}$$

It is easy to verify that the solution of the problem (16.3.5)-(16.3.6) is

$$w(x,t) = x - Ke^{-r(T-t)}. \tag{16.3.7}$$

Now we have

$$Y_t - Z_t = u(S_t,t) - v(S_t,t) = w(S_t,t) = S_t - Ke^{-r(T-t)},$$

and this proves the call-put parity formula.

There is also a proof of this formula based on financial arguments. Indeed, let us consider the portfolio

$$Q_t = Y_t - Z_t + Ke^{-r(T-t)}.$$

We have

$$Q_T = Y_T - Z_T + K = (S_T - K)^+ - (K - S_T)^+ + K = S_T - K + K = S_T.$$

Hence $Q_T = S_T$; to avoid arbitrage opportunities, it is necessary to have $Q_t = S_t$ for all $t \in [0,T]$, which leads immediately to (16.3.3).

16.4 Problems

1. Prove the call-put parity formula (16.3.3) by using (16.3.1), (16.3.2), (16.2.13) and (16.2.18).

2. Consider the problem (16.2.11)-(16.2.12). Obviously the function $u(\cdot, T)$ is convex. Prove that the function $u(\cdot, t)$ is also convex, for all $t \in [0, T]$.

3. The share price is \$20 today, the interest rate $r = 0.05$, the volatility $\sigma = 0.1$. A call option has strike price $K = \$22$ and maturity date $T = 1$ (that is, one year). Find the present price of the option.

4. Prove that the solution of the problem (16.2.11)-(16.2.12) is also given by

$$u(x,t) = \frac{e^{-r(T-t)}}{\sqrt{2\pi(T-t)}} \int_{\mathbb{R}} \left(x e^{\sigma s + (r - \sigma^2/2)(T-t)} - K \right)^{+} e^{-s^2/2(T-t)} ds.$$

Bibliography

[1] G. Ciucu, C. Tudor, *Probabilităţi şi procese stocastice*, vol.I, II, Ed. Academiei, Bucureşti, 1978.

[2] G. Ciucu, V. Craiu, I. Săcuiu, *Probleme de teoria probabilităţilor,* Ed. Tehnică, Bucureşti, 1974.

[3] S. Dineen, *Probability Theory in Finance*, Graduate Studies in Mathematics 70, American Math. Society, 2005.

[4] M. Iosifescu, *Lanţuri Markov finite şi aplicaţii*, Ed. Tehnică, Bucureşti, 1977.

[5] C. Jalobeanu, I. Raşa, *Mathcad. Probleme de calcul numeric şi statistic*, Editura Albastră, Cluj, 1995.

[6] C. Jalobeanu, I. Raşa, *Incertitudine şi decizie. Statistică şi probabilităţi aplicate în management*, Editura U.T. Pres, 2001.

[7] A. Mitrea, *Fundamente de Teoria Probabilităţilor*, U.T.Pres, 2003.

[8] T. K. Moon, Wynn C. Stirling, *Mathematical Methods and Algorithms for Signal Processing*, Prentice Hall, 2000.

[9] R. Munteanu, Gh. Todoran, *Teoria şi practica prelucrării datelor de măsurare*, Editura Mediamira, 1997.

[10] B. Øksendal, *Stochastic Differential Equations*, Springer-Verlag, 2000.

[11] T. G. Potra, *Probabilităţi şi statistică matematică. Procese stochastice*, Transilvania Press, 2004.

[12] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE*, 77(1989), 257-286.

[13] I. Raşa, *Teoria probabilităţilor şi aplicaţii*, Lito U.T.C.N, 1994.

[14] H. Stark, J. W. Woods, *Probability, random proceses and estimation theory for engineers*, Prentice Hall, 1986.