Appendix F

Concepts from probability theory

This appendix provides a ready reference for many of the basic concepts from elementary probability theory.

F.1 Probability space

Let Ω denote the set of all **elementary events** where the choice of these events very much depends on the specific problem of interest. Consider the experiment of **tossing a single coin**. In this case Ω consists of two elements each denoting the outcome of a single toss namely, the coin falling **head** or **tail**. Consider the problem of simultaneously **throwing two dice**. Recall that each dice has six faces labelled with integers from 1 through 6. Hence Ω contains 36 pairs $\{(i, j)\}$ where i and j take independently values from 1 through 6, where i denotes the number on the top face of the first dice and j on the second dice. If we are considering the **price of a technology stock**, Ω will consists of all factors that directly or indirectly affect the raise and fall of the prices of these stocks which includes status of the national economy, foreign competition for the products, weather, acts of terrorism, foreign exchange rate, quality of accounting practices, and the credibility of the upper management, to mention a few.

Let $\mathcal{P}$ denote an **assignment of probability** to each of the elementary events in Ω. Thus, if ω denotes an elementary event, then it is required that

$$\mathcal{P}(\omega) \geq 0 \quad \text{and} \quad \sum_{\omega \in \Omega} \mathcal{P}(\omega) = 1. \tag{F.1.1}$$

In the case of a coin tossing experiment

$$\mathcal{P}(head) = p \qquad \text{and} \qquad \mathcal{P}(tail) = q$$

where p and q are nonnegative real numbers and $p + q = 1$. In the case of a single throw of one dice, let p_i denote the probability that this dice will fall with its face marked i where

$$p_i \geq 0 \qquad \text{and} \qquad \sum_{i=1}^{6} p_i = 1. \tag{F.1.2}$$

Thus, in the case of a single throw of two dices simultaneously, Ω consists of 36 pairs of the form (i, j) that is,

$$\Omega = \{(i, j) | 1 \leq i \leq 6 \text{ and } 1 \leq j \leq 6\}.$$

Let q_{ij} be the probability assigned to the elementary event (i, j). Then

$$q_{ij} \geq 0 \quad \text{and} \quad q_{ij} = p_i p_j \quad \text{with} \quad \sum_{i=1}^{6} \sum_{j=1}^{6} q_{ij} = 1. \tag{F.1.3}$$

Any subset A of Ω (denoted by $A \subseteq \Omega$) is called an **event**. Let A and B be two events. Then $A \cup B$ (read as A **union** B) denotes the combined event denoting the occurrence of the **event** A **or event** B **or both**. Similarly, $A \cap B$ or equivalently AB (read as A **intersection** B) denotes the combined event denoting the **simultaneous** occurrence of the **events** A **and** B. It can be verified that

$$\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(AB). \tag{F.1.4}$$

Let $\mathcal{F}$ denote the **set of events** of interest to us. Then the triple $(\Omega, \mathcal{F}, \mathcal{P})$ is called the **probability space**. The set Ω can be finite or infinite. When Ω is finite, $\mathcal{F}$ consists of all subsets of Ω and $\mathcal{F}$ is the **power set**[†] **of** Ω**.** The term $\mathcal{P}$ is known as the **probability measure**. It is required that $\mathcal{P}$ satisfies the following technical conditions:

(1) If ϕ is the null set, then $\mathcal{P}(\phi) = 0$.
(2) For any subset $\mathcal{S} \subseteq \Omega$, $\mathcal{P}(\mathcal{S}) \geq 0$.
(3) If $\{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \ldots\}$ is an **infinite** collection of subsets of Ω that are **mutually disjoint**[‡], then

$$\mathcal{P}\left[\bigcup_{i=0}^{\infty} \mathcal{S}_i\right] = \sum_{i=0}^{\infty} \mathcal{P}(\mathcal{S}_i). \tag{F.1.5}$$

As an example, the probability space $(\Omega, \mathcal{F}, \mathcal{P})$ for the single throw of two dices is such that Ω has 36 **elementary events**, $\mathcal{F}$ has $2^{36} = 68\,719\,476\,736$ events and $\mathcal{P}$ is the probability measure given by (F.1.3). Now what is the probability of obtaining the sum of six in one single simultaneous throw of two dices? Here we want $i + j = 6$ and this could happen in **five mutually exclusive** ways as

$$i + j = 6 = (1 + 5) \text{ or } (2 + 4) \text{ or } (3 + 3) \text{ or } (4 + 2) \text{ or } (5 + 1).$$

Thus,

$$\mathcal{P}[i + j = 6] = p_3^2 + 2p_1 p_5 + 2p_2 p_4. \tag{F.1.6}$$

[†] If $\Omega = \{a, b, c\}$, then the **power set** of Ω consists of $2^3 = 8$ subsets given by $\{\phi, a, b, c, ab, bc, ac, abc\}$ where for simplicity the subset $\{a, b\}$ is denoted as ab, and ϕ denotes the null set. In general, if Ω contains k elements, then the power set of Ω contains 2^k elements each of which is a subset of Ω.

[‡] Two subsets $\mathcal{A}$ and $\mathcal{B}$ of Ω are said to be **mutually disjoint or exclusive** if $\mathcal{A}$ and $\mathcal{B}$ **do not** have any common elementary event. That is, the **intersection** of $\mathcal{A}$ and $\mathcal{B}$ is a null set.

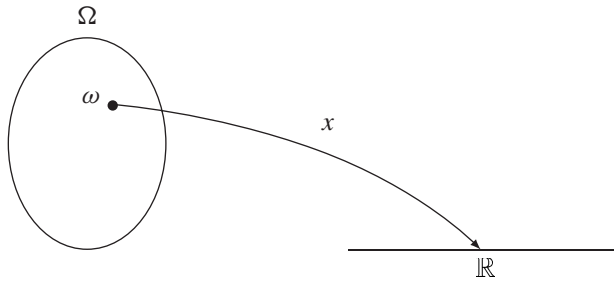

Fig. F.2.1 A view of the probability space and random variable $x : \Omega \to \mathbb{R}$.

F.2 Random variables and vectors

A **real-valued random variable** x is a function that maps Ω to the real line, $\mathbb{R}$. That is, $x : \Omega \to \mathbb{R}$. Thus, for each elementary event ω, $x(\omega)$ is a real number. Let $\mathcal{P}_x$ denote the **induced** probability measure on $\mathbb{R}$. To define $\mathcal{P}_x$ we must first relate events in $(\Omega, \mathcal{F}, \mathcal{P})$ to intervals in $\mathbb{R}$. Let $B = [a, b]$ denote a closed interval in $\mathbb{R}$. Let $\mathcal{S} \in \mathcal{F}$ be both such that

$$\mathcal{S} = \{\omega \in \Omega | x(\omega) \in B\} = x^{-1}(B). \tag{F.2.1}$$

That is, x is such that it maps the elements of $\mathcal{S}$ onto the interval B. Then by definition,

$$\mathcal{P}_x(B) = \mathcal{P}(x^{-1}(B)). \tag{F.2.2}$$

The induced probability measure $\mathcal{P}_x$ so defined satisfies the following conditions:

$$\mathcal{P}_x(\phi) = \mathcal{P}\left[x^{-1}(\phi)\right] = \mathcal{P}(\phi) = 0$$

and

$$\mathcal{P}_x(\mathbb{R}) = \mathcal{P}\left[x^{-1}(\mathbb{R})\right] = \mathcal{P}(\Omega) = 1.$$

Given $(\Omega, \mathcal{F}, \mathcal{P})$ we can now define a new triple $(x, \mathcal{B}, \mathcal{P}_x)$ where the values that the random variable $x(\omega)$ takes as ω is varied in Ω provides a new representation of the elementary events in Ω and $\mathcal{B}$ denotes the set of all intervals of the real line for which $\mathcal{P}_x$ is defined using (F.2.1).

The relation between these two representations of the probability spaces is given in Figure F.2.1.

In the coin tossing experiment $\Omega = \{\text{head, tail}\}$ and $x : \Omega \to \mathbb{R}$ is defined by

$$\begin{aligned} x(\text{head}) &= 1 \ \text{ with probability } p \\ x(\text{tail}) &= 0 \ \text{ with probability } q \end{aligned}$$

where $p + q = 1$. When $p = 1/2 = q$, it is called a **fair** coin, otherwise it is a biased coin. This random variable has a special name called **Bernoulli** random

variable. In the case of the dice, $\Omega = \{$six faces of the dice numbered 1 through 6$\}$. Define $x : \Omega \to \mathbb{R}$ as

$$x(\text{face marked } i) = i \text{ with probability } p_i$$

for $i = 1$ to 6. Here again, if $p_i = 1/6$ for each i, then it is called a **balanced** or **fair** dice, otherwise it is said to be biased.

Remark F.2.1 Elements of $\mathcal{B}$ which are all intervals of the real line are known as **Borel sets** and $\mathcal{B}$ itself is called the Borel σ-field (pronounced as Borel sigma field). Any closed subset of a real line is a Borel set. The advantage of working with $(x, \mathcal{B}, \mathcal{P}_x)$ is that in addition to performing all the operations on $(x, \mathcal{F}, \mathcal{P})$, since the real line is endowed with a rich **topological** structure, we can readily bring to bear all the mathematical tools at the disposal of probabilistic analysis. As an example, $\mathcal{P}_x$ can be defined in terms of a probability distribution function whose derivative corresponds to the notion of probability density function as described below.

Let x be a (continuous) random variable. The **cumulative distribution** of x is the function $F : \mathbb{R} \to \mathbb{R}$ where for any real number a

$$\begin{aligned} F(a) &= \mathcal{P}_x\,[x \le a] \\ &= \textstyle\int_{-\infty}^{a} f(x)\mathrm{d}x \end{aligned} \tag{F.2.3}$$

where $f(x) \ge 0$ is called the **probability density function** of x. That is, the induced probability measure can be defined using the probability density function $f(x)$ satisfying the following conditions:

(1) $f(x) \ge 0$ for all $x \in \mathbb{R}$ and $F(a) \ge a$ for all $a \in \mathbb{R}$.
(2) $F(a)$ is an increasing function of a with

$$\lim_{a \to -\infty} F(a) = 0 \quad \text{and} \quad \lim_{a \to \infty} F(a) = 1.$$

(3)

$$\begin{aligned} \textstyle\int_a^b f(x)\mathrm{d}x &= F(b) - F(a) \\ &= \mathcal{P}_x\,[a < x \le b] \\ &= \mathcal{P}\left[x^{-1}(B)\right] \qquad \text{with} \qquad B = (a, b]. \end{aligned}$$

In the special case when x takes only finitely many values, that is $x \in \{b_1, b_2, \ldots, b_k\}$ where b_i's are real and $k < \infty$, it is called a **discrete random** variable. In this case, the probability distribution of x is given by

$$\left.\begin{aligned} &\mathcal{P}_x\,[x = b_i] = p_i > 0 \\ \text{and} \qquad & \\ &F(a) = \mathcal{P}_x\,[x \le a] \\ &\quad = \textstyle\sum_j p_j \end{aligned}\right\} \tag{F.2.4}$$

where the summation is taken over all those j's for which $b_j \le a$.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ denote a **random vector** of size n where each component x_i is a random variable defined over the same underlying probability space $(\Omega, \mathcal{F}, \mathcal{P})$. That is, $x_i : \Omega \to \mathbb{R}$ for each $i = 1$ to n. Thus $x : \Omega^n \to \mathbb{R}^n$ where $\Omega^n = \Omega \times \Omega \times \cdots \times \Omega$, the n fold **cartesian product**[†] of Ω and $\mathbb{R}^n$ is the standard n-dimensional **Euclidean space** (Appendix A).

In the interest of simplicity in notation, in the following random vectors of size two, that is, $n = 2$ are used. Extension to general n is rather obvious.

Let $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ be a random vector and $f : \mathbb{R}^2 \to \mathbb{R}$ be the **joint probability density function** of the vector $\mathbf{x}$. Let $\mathbf{a} = (a_1, a_2)^{\mathrm{T}} \in \mathbb{R}^2$. Then

$$\left.\begin{aligned} F(\mathbf{a}) &= \mathrm{Prob}\,[x_1 \le a_1 \text{ and } x_2 \le a_2] \\ &= \int_{-\infty}^{a_1} \int_{-\infty}^{a_2} f(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \\ &= \int_{-\infty}^{a} f(x)\mathrm{d}x \end{aligned}\right\} \tag{F.2.5}$$

denotes the **joint cumulative probability distribution** of $\mathbf{x}$. Clearly,

$$\lim_{a_1 \to -\infty} F(\mathbf{a}) = 0 = \lim_{a_2 \to -\infty} F(\mathbf{a})$$

and

$$\lim_{a_1 \to \infty} \lim_{a_2 \to \infty} F(\mathbf{a}) = 1.$$

$$\mathrm{Prob}\,[x_1 \le a_1] = \int_{-\infty}^{a_1} \int_{-\infty}^{\infty} f(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \tag{F.2.6}$$

is called the **marginal distribution** of x_1. Similarly the marginal distribution of x_2 can be defined.

In the discrete case $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ with $x_1 \in \{b_1, b_2, \ldots, b_n\}$ and $x_2 \in \{c_1, c_2, \ldots, c_n\}$. Then

$$\mathcal{P}_x \left[x_1 = b_i \text{ and } x_2 = c_j\right] = p_{ij} \tag{F.2.7}$$

defines the joint distribution of $\mathbf{x}$. Then

$$\begin{aligned} F(\mathbf{a}) &= \mathcal{P}_x\,[x_1 = a_1 \text{ and } x_2 = a_2] \\ &= \textstyle\sum_i \sum_j p_{ij} \end{aligned}$$

where the sum is taken over all those i's and j's such that $b_i \le a_1$ and $c_j \le a_2$.

F.3 Expected value, variance and covariance

Let x be a given random variable. Let $f(x)$ denote the probability density of x when x is **continuous** and let x take the value x_i with probability p_i for $i = 1$ to k when $\mathbf{x}$ is **discrete**.

[†] If $\Omega_1 = \{a, b, c\}$ and $\Omega_2 = \{\alpha, \beta\}$, then the **cartesian product** $\Omega_1 \times \Omega_2$ is the set of all pairs of the form (x, y) with $x \in \Omega_1$ and $y \in \Omega_2$. Thus, $\Omega_1 \times \Omega_2 = \{(a, \alpha), (a, \beta), (b, \alpha), (b, \beta), (c, \alpha), (c, \beta)\}$

(a) The **expected value** or the **mean** of x is denoted by $E(x)$ and is defined by

$$E(x) = \begin{cases} \int_{-\infty}^{\infty} x f(x) \mathrm{d}x & \text{when } x \text{ is continuous} \\ \sum_{i=1}^{k} x_i p_i & \text{when } x \text{ is discrete} \end{cases} \tag{F.3.1}$$

The function $E(\cdot)$ so defined is called the expectation operator. Let x_1 and x_2 be two random variables and let a be real constant. It can be verified that

$$\left.\begin{array}{lll} E(x_1 + x_2) = E(x_1) + E(x_2) & - \text{Additivity} \\ E(ax_1) \quad = aE(x_1) & - \text{Homogeneity} \end{array}\right\} \tag{F.3.2}$$

from which it follows that $E(\cdot)$ is a **linear operator**. The following are easily verified.

$$\left.\begin{array}{l} E(x + a) \quad = E(x) + a \\ E\left[x - E(x)\right] = 0 \end{array}\right\} \tag{F.3.3}$$

(b) The **variance** of x denoted by σ^2 is defined by

$$\sigma^2 = \mathrm{Var}(x) = \begin{cases} \int_{-\infty}^{\infty} \left[x - E(x)\right]^2 f(x) \mathrm{d}x & - x \text{ continuous} \\ \sum_{i=1}^{k} \left[x_i - E(x)\right]^2 p_i & - x \text{ discrete} \end{cases} \tag{F.3.4}$$

After simplification of the r.h.s., it can be verified that

$$\sigma^2 = E(x^2) - \left[E(x)\right]^2 .$$

For any constant a, it follows that

$$\mathrm{Var}(ax) = a^2 \mathrm{Var}(x). \tag{F.3.5}$$

The square root of the variance is known as the **standard deviation** and is denoted by σ.

(c) **Normalizing a random variable** If x is a random variable with mean $\mu = E(x)$and variance σ^2, then

$$z = \tfrac{x - \mu}{\sigma} \tag{F.3.6}$$

is the normalized random variable with mean, $E(z) = 0$ and variance, $\mathrm{Var}(z) = 1$.

(d) **Mean, variance and covariance of random vectors** Let $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ be a random vector with the **joint probability density given** by $f(\mathbf{x}) = f(x_1, x_2)$.

Let $\mu = (\mu_1, \mu_2)^{\mathrm{T}}$ with $\mu_i = E(x_i)$ denote the **mean** of $\mathbf{x}$. Then

$$\mu = \int_{\mathbb{R}^2} \mathbf{x} f(\mathbf{x}) \mathrm{d}x = E(\mathbf{x}) \tag{F.3.7}$$

where

$$\begin{aligned}\mu_1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \\ &= \int_{-\infty}^{\infty} x_1 \left[\int_{-\infty}^{\infty} f(x_1, x_2)\mathrm{d}x_2\right]\mathrm{d}x_1 \\ &= \int_{-\infty}^{\infty} x_1 f_1(x_1)\mathrm{d}x_1\end{aligned}$$

where

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\mathrm{d}x_2 \tag{F.3.8}$$

is called the **marginal density** of x_1. Similarly μ_2 can be defined.

If $\mathbf{x}$ and $\mathbf{y}$ are two random vectors and if $\mathbf{a} = (a_1, a_2)^{\mathrm{T}}$, then it can be verified that

$$\left.\begin{aligned} E(\mathbf{x}+\mathbf{y}) &= E(\mathbf{x}) + E(\mathbf{y}) \\ E(\mathbf{x}+\mathbf{a}) &= E(\mathbf{x}) + \mathbf{a} \\ E(\mathbf{a}^{\mathrm{T}}\mathbf{x}) &= \textstyle\sum_{i=1}^{2} a_i E(x_i) \end{aligned}\right\} \tag{F.3.9}$$

Let σ_1^2 and σ_2^2 be the **variances** of x_1 and x_2. Then

$$\begin{aligned}\sigma_1^2 &= \int_{-\infty}^{\infty}(x_1-\mu_1)^2 f_1(x_1)\mathrm{d}x_1 = E(x_1-\mu_1)^2 \\ \sigma_2^2 &= \int_{-\infty}^{\infty}(x_2-\mu_2)^2 f_2(x_2)\mathrm{d}x_2 = E(x_2-\mu_2)^2.\end{aligned} \tag{F.3.10}$$

The **covariance** between x_1 and x_2 denoted by

$$\begin{aligned}\mathrm{Cov}(x_1, x_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1-\mu_1)(x_2-\mu_2) f(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \\ &= E\left[(x_1-\mu_1)(x_2-\mu_2)\right] \\ &= E\left[(x_2-\mu_2)(x_1-\mu_1)\right] \\ &= \mathrm{Cov}(x_2, x_1).\end{aligned} \tag{F.3.11}$$

It is convenient to denote the $\mathrm{Cov}(x_1, x_2)$ as σ_{12}^2. Clearly $\sigma_{12}^2 = \sigma_{21}^2$.

The **covariance matrix** Σ is defined by expected value of the outer product matrix as

$$\begin{aligned}\Sigma &= E\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^{\mathrm{T}}\right] \\ &= E\left[\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}(x_1-\mu_1)(x_2-\mu_2)\right] \\ &= \begin{bmatrix} E(x_1-\mu_1)^2 & E\left[(x_1-\mu_1)(x_2-\mu_2)\right] \\ E\left[(x_2-\mu_2)(x_1-\mu_1)\right] & E(x_2-\mu_2)^2 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{bmatrix}, \quad \text{a symmetric matrix.}\end{aligned} \tag{F.3.12}$$

The above development readily generalizes to the case of a random vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ of size n. Then $\mu = E(\mathbf{x})$ is the **mean vector** of size n

and

$$\Sigma = E\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^{\mathrm{T}}\right] \tag{F.3.13}$$

is an $n \times n$ **symmetric matrix** which is the **covariance matrix** of the vector $\mathbf{x}$.

(e) **Transformation of random vectors** Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$ a matrix. We consider two types of transformations of $\mathbf{x}$. First consider the **scalar product** $\mathbf{a}^{\mathrm{T}}\mathbf{x}$. It can be verified

$$E(\mathbf{a}^{\mathrm{T}}\mathbf{x}) = \mathbf{a}^{\mathrm{T}}E(\mathbf{x})$$

and

$$\begin{aligned}
\mathrm{Var}(\mathbf{a}^{\mathrm{T}}\mathbf{x}) &= E\left[\mathbf{a}^{\mathrm{T}}\mathbf{x} - \mathbf{a}^{\mathrm{T}}E(\mathbf{x})\right]^2 \\
&= E\left[\mathbf{a}^{\mathrm{T}}(\mathbf{x}-E(\mathbf{x}))\right]^2 \\
&= E\left[\mathbf{a}^{\mathrm{T}}(\mathbf{x}-E(\mathbf{x}))\mathbf{a}^{\mathrm{T}}(\mathbf{x}-E(\mathbf{x}))\right] \\
&= E\left[\mathbf{a}^{\mathrm{T}}(\mathbf{x}-E(\mathbf{x}))(\mathbf{x}-E(\mathbf{x}))^{\mathrm{T}}\mathbf{a}\right] \quad [\textit{since}\ \mathbf{a}^{\mathrm{T}}\mathbf{b} = \mathbf{b}^{\mathrm{T}}\mathbf{a}] \\
&= \mathbf{a}^{\mathrm{T}}E\left[(\mathbf{x}-E(\mathbf{x}))(\mathbf{x}-E(\mathbf{x}))^{\mathrm{T}}\right]\mathbf{a} \\
&= \mathbf{a}^{\mathrm{T}}\Sigma\,\mathbf{a}
\end{aligned} \tag{F.3.14}$$

which is a quadratic form in $\mathbf{a}$.

Now consider the **matrix transformation** of the vector $\mathbf{x}$, where $\mathbf{z} = \mathbf{A}\mathbf{x}$. Then, if $E(\mathbf{x}) = \mu$, we have

$$E(\mathbf{z}) = \mathbf{A}E(\mathbf{x}) = \mathbf{A}\mu$$

and

$$\begin{aligned}
\mathrm{Cov}(\mathbf{z}) = \Sigma_{\mathbf{z}} &= E\left[(\mathbf{z}-\mathbf{A}\mu)(\mathbf{z}-\mathbf{A}\mu)^{\mathrm{T}}\right] \\
&= E\left[\mathbf{A}(\mathbf{x}-\mu)(\mathbf{x}-\mu)^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\right] \\
&= \mathbf{A}E\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^{\mathrm{T}}\right]\mathbf{A}^{\mathrm{T}} \\
&= \mathbf{A}\Sigma_{\mathbf{x}}\mathbf{A}^{\mathrm{T}}
\end{aligned} \tag{F.3.15}$$

where $\Sigma_{\mathbf{x}}$ is the covariance matrix of $\mathbf{x}$.

Now let $\mathbf{x}$ and $\mathbf{y}$ be two random vectors of size n and $\mathbf{A}$ and $\mathbf{B}$ be two $n \times n$ matrices. Let $E(\mathbf{x}) = \mu_{\mathbf{x}}$ and $E(\mathbf{y}) = \mu_{\mathbf{y}}$. Then

$$\begin{aligned}
\mathrm{Cov}(\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{y}) &= E\left[(\mathbf{A}\mathbf{x}-\mathbf{A}\mu_{\mathbf{x}})(\mathbf{B}\mathbf{y}-\mathbf{B}\mu_{\mathbf{y}})^{\mathrm{T}}\right] \\
&= E\left[\mathbf{A}(\mathbf{x}-\mu_{\mathbf{x}})(\mathbf{y}-\mu_{\mathbf{y}})^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\right] \\
&= \mathbf{A}E\left[(\mathbf{x}-\mu_{\mathbf{x}})(\mathbf{y}-\mu_{\mathbf{y}})^{\mathrm{T}}\right]\mathbf{B}^{\mathrm{T}} \\
&= \mathbf{A}\,\mathrm{Cov}(\mathbf{x}, \mathbf{y})\mathbf{B}^{\mathrm{T}}.
\end{aligned} \tag{F.3.16}$$

(f) **Correlation matrix** Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$. Then ρ_{ij}, the **correlation coefficient** between x_i and x_j $(i \neq j)$ is given by

$$\begin{aligned}\rho_{ij} = \mathrm{Cov}(x_i, x_j) &= \frac{\mathrm{Cov}(x_i, x_j)}{\sqrt{\mathrm{Var}(x_i)}\sqrt{\mathrm{Var}(x_j)}} \\ &= \frac{\sigma_{ij}^2}{\sigma_i \sigma_j} \\ &= \frac{\sigma_{ji}^2}{\sigma_i \sigma_j} = \rho_{ji}. \end{aligned} \tag{F.3.17}$$

It can be verified that when $i \neq j$, the value of ρ_{ij} can be either positive or negative but its absolute value is always less than one, that is, $|\rho_{ij}| \leq 1$. When $i = j$

$$\rho_{ii} = \frac{\sigma_i^2}{\sigma_i \sigma_i} = 1 \text{ for all } i. \tag{F.3.18}$$

We thus obtain an $n \times n$ **symmetric matrix**

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} \\ \rho_{12} & 1 & \rho_{23} & \cdots & \rho_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{1n} & \rho_{2n} & \rho_{3n} & \cdots & 1 \end{bmatrix} \tag{F.3.19}$$

called the **correlation matrix**.

To see the relation between the covariance matrix Σ and the correlation matrix $\mathbf{R}$, first define a diagonal matrix $\mathbf{D}$ consisting of the diagonal elements of Σ as

$$\mathbf{D} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}. \tag{F.3.20}$$

Then define the square root of $\mathbf{D}$ as

$$\mathbf{D}^{\frac{1}{2}} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix}, \tag{F.3.21}$$

where the diagonal elements are the standard deviations. Then from the definition (F.3.17), it follows that

$$\mathbf{R} = \mathbf{D}^{-\frac{1}{2}} \Sigma \mathbf{D}^{-\frac{1}{2}} \tag{F.3.22}$$

where $\mathbf{D}^{-\frac{1}{2}}$ is the inverse of $\mathbf{D}^{\frac{1}{2}}$. It can also be verified that

$$\mathbf{R} = \mathrm{Cor}(\mathbf{z}) \quad \text{where } \mathbf{z} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{x} - \mu) \tag{F.3.23}$$

is the **normalized vector**.

When $\sigma_{ij}^2 = 0$ for $i \neq j$ it implies that $\rho_{ij} = 0$ and the random variables x_i and x_j are **uncorrelated** or **orthogonal**. When $\sigma_{ij}^2 = 0$ for all i and j, then all the components of $\mathbf{x}$ are **uncorrelated** and Σ becomes the diagonal matrix $\mathbf{D}$ in (F.3.20) and in this case $\mathbf{R}$ reduces to an **identity matrix.**

In the special case when $n = 2$, we have

$$\text{Cov}(\mathbf{x}) = \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\, \sigma_1\sigma_2 \\ \rho\, \sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{F.3.24}$$

and

$$\text{Cor}(\mathbf{x}) = \mathbf{R} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{F.3.25}$$

where $\rho = \sigma_{12}^2/\sigma_1\sigma_2$.

F.4 Independence

The notion of independence in probability theory is quite basic and it naturally arises in modelling many real world phenomenona. For example, in the tossing of a coin experiment, the result of the second toss does **not** depend on or is **independent** of the knowledge that we obtained a head or tail in the first toss. Similarly, in simultaneously throwing two dice, the result of one dice does **not** affect the other. Examples of this type abound in nature and the notion of **statistical independence** is the mathematical formalism of this naturally occurring notion of independence.

In terms of the original probability space $(\Omega, \mathcal{F}, \mathcal{P})$, two events A and B are said to be **independent** if

$$\mathcal{P}(AB) = \mathcal{P}(A)\mathcal{P}(B). \tag{F.4.1}$$

In terms of the random variables, if $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ and $f(\mathbf{x}) = f(x_1, x_2)$ is the joint density of $\mathbf{x}$ then x_1 and x_2 are **independent** if

$$f(x_1, x_2) = f_1(x_1) f_2(x_2). \tag{F.4.2}$$

Thus, if x_1 and x_2 are independent, then

$$\begin{aligned} E(x_1x_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2 f(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \\ &= \int_{-\infty}^{\infty} x_1 f_1(x_1)\mathrm{d}x_1 \int_{-\infty}^{\infty} x_2 f_2(x_2)\mathrm{d}x_2 \\ &= E(x_1)E(x_2) \end{aligned} \tag{F.4.3}$$

and

$$\begin{aligned} \text{Cov}(x_1x_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1 - E(x_1))(x_2 - E(x_2)) f(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \\ &= \int_{-\infty}^{\infty} [x_1 - E(x_1)]\, f_1(x_1)\mathrm{d}x_1 \int_{-\infty}^{\infty} [x_2 - E(x_2)]\, f_2(x_2)\mathrm{d}x_2 \\ &= 0 \quad \text{using (F.3.3).} \end{aligned} \tag{F.4.4}$$

Let x_1 and x_2 be two random variables. Then let $g(x_1)$ and $h(x_2)$ be two functions of x_1 and x_2 respectively. It can be verified that if x_1 and x_2 are independent, then so are the random variables $g(x_1)$ and $h(x_2)$. If x_1 and x_2 are independent, their covariance and hence their correlation is zero. That is, **independence implies uncorrelated**. The converse is not, in general, true. For, let x_1 and x_2 be two independent zero mean random variables. Let $z = x_1x_2$. Then $E(z) = E(x_1)E(x_2) = 0$. Then $E(zx_1) = E(x_1^2x_2) = E(x_1^2)E(x_2) = 0$ that is, z and x_1 are uncorrelated. Clearly, they are **not independent**. Notice that we have used the fact that if x_1 and x_2 are independent, then, as random variables, so are $g(x_1) = x_1^2$ and $h(x_2) = x_2$.

F.5 Conditional probability and Bayes' rules

We now move on to analyzing dependency among events. Let A and B be two events. It is often of interest to compute the probability of occurrence of the event A given that the event B has already occurred. This **conditional** probability of the occurrence of A given that B has occurred is denoted by $P(A|B)$. Likewise, one can define $P(B|A)$. It is well known that the joint probability $P(AB)$ of the simultaneous occurrence of events A and B can be expanded in two ways using the conditional probabilities as

$$P(AB) = P(A|B)P(B) = P(B|A)P(A) \tag{F.5.1}$$

where $P(A)$ and $P(B)$ denote the probabilities of the occurrence of the individual events A and B respectively. $P(A)$ and $P(B)$ are also known as **prior** probabilities. On rewriting, (F.5.1) becomes

$$P(B|A) = \frac{P(A|B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} \tag{F.5.2}$$

which is known as the **Bayes' rule**.

Let $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ be a random vector with $f(x_1, x_2)$ as the joint density. Then the conditional density of x_1 given x_2 denoted by $f(x_1|x_2)$ is given by

$$f(x_1|x_2) = \begin{cases} \frac{f(x_1,x_2)}{f_2(x_2)} & \text{when } f_2(x_2) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{F.5.3}$$

where $f_2(x_2)$ is the marginal density of x_2. It can be verified that

$$\begin{aligned} \int_{-\infty}^{\infty} f(x_1|x_2)\mathrm{d}x_1 &= \frac{1}{f_2(x_2)} \int_{-\infty}^{\infty} f(x_1, x_2)\mathrm{d}x_1 \\ &= 1. \end{aligned} \tag{F.5.4}$$

F.6 Conditional expectation

The conditional expectation of x_1 given x_2 given by

$$E\left[x_1|x_2 = a\right] = \int_{-\infty}^{\infty} x_1 f(x_1, x_2 = a)\mathrm{d}x_1. \tag{F.6.1}$$

As x_2 takes on different values in its domain, this conditional expectation will also take different values. In other words, the conditional expectation $E(x_1|x_2)$ is clearly a function of the conditioning random variable and hence it itself is a random variable. Sometimes it is useful to distinguish the conditional expectation operator by using the conditioning random variable as the subscript. Thus, $E(x_1|x_2)$ is also denoted as $E_{x_2}(x_1)$. Thus,

$$\begin{aligned} E\{E_{x_2}(x_1)\} &= E\{E\left[x_1|x_2\right]\} \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x_1 f(x_1|x_2)\mathrm{d}x_1\right] f_2(x_2)\mathrm{d}x_2 \\ &= \int_{-\infty}^{\infty} x_1 \left[\int_{-\infty}^{\infty} f(x_1|x_2) f_2(x_2)\mathrm{d}x_2\right]\mathrm{d}x_1 \\ &= \int_{-\infty}^{\infty} x_1 f_1(x_1)\mathrm{d}x_1 \\ &= E(x_1). \end{aligned} \tag{F.6.2}$$

Thus, the random variable $E_{x_2}(x_1)$ has the same expectation as x_1 and (F.6.2) is called the **law of iterated expectations**.

We now quote a standard result relating to the conditional expectation of normal variates. Let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{z} \in \mathbb{R}^m$ be jointly normal random vectors, that is,

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{z} \end{pmatrix} \sim N(m, \Sigma) \tag{F.6.3}$$

where

$$m = \begin{pmatrix} m_x \\ m_z \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_x & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_z \end{bmatrix}. \tag{F.6.4}$$

Then, the **conditional density** of $f(\mathbf{x}|\mathbf{z})$ of $\mathbf{x}$ given $\mathbf{z}$ is also normal and is given by

$$f(\mathbf{x}|\mathbf{z}) \sim N(\mu, \mathbf{A}) \tag{F.6.5}$$

where the **conditional mean**

$$\mu = E[\mathbf{x}|\mathbf{z}] = m_x + \Sigma_{xz}\Sigma_z^{-1}[\mathbf{z} - m_z] \tag{F.6.6}$$

and the **conditional covariance**

$$\mathbf{A} = \mathrm{Cov}(x|z) = \Sigma_x - \Sigma_{xz}\Sigma_z^{-1}\Sigma_{zx}. \tag{F.6.7}$$

F.7 Special distributions

In this section, we list the properties of some of the common distributions of interest in our analysis.

(a) **Bernoulli distribution** This is a **discrete** distribution wherein a random variable x takes on only two values:

$$x = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } q = 1 - p \end{cases} \tag{F.7.1}$$

It can be verified that

$$E(x) = p, \quad \text{Var}(x) = pq.$$

(b) **Binomial distribution** This is again a **discrete** distribution that models the distribution of the number of heads in n successive tosses of a coin that falls head with probability p and tail with probability $q = 1 - p$. Since the n tosses are independent, the random variable x, that denotes the total number of heads in n tosses, can take values from 0 through n according to the following rule.

$$\text{Prob}\,[x = k] = \binom{n}{k} p^k q^{n-k} \tag{F.7.2}$$

where $k = 0, 1, 2, \ldots, n$.

It can be verified that $E(x) = np$ and $\text{Var}(x) = npq$.

(c) **Univariate normal distribution** A **continuous** random variable x is said to be **normally** distributed with mean μ and variance σ^2, denoted by $x \sim N(\mu, \sigma^2)$ if the probability density function of x is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{F.7.3}$$

It can be verified that $E(x) = \mu$ and $\text{Var}(x) = \sigma^2$. When $\mu = 0$ and $\sigma^2 = 1$, x is called the standard normal distribution. The graph of the standard normal distribution

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right] \tag{F.7.4}$$

is given in Figure F.7.1. It can be verified that

$$\int_{-1}^{1} f(z)\mathrm{d}z = 0.6826 \quad \text{and} \quad \int_{-2}^{2} f(z)\mathrm{d}z = 0.9544, \tag{F.7.5}$$

that is, slightly over 95% of the total area of $f(z)$ is contained in the line segment from -2 to 2. Normal distribution was originally invented by Gauss and is also known as **Gaussian distribution**.

(d) **Multivariate normal distribution** Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ be a vector. The vector $\mathbf{x}$ is said to be normally distributed with mean vector μ and covariance

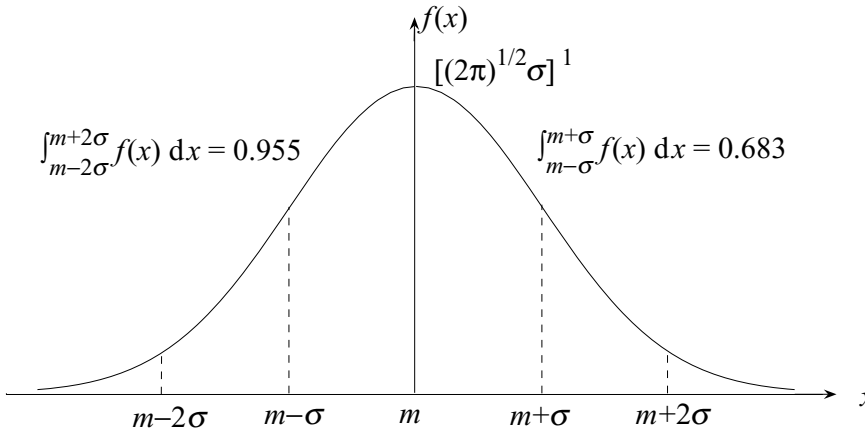

Fig. F.7.1 The plot of the standard normal distribution function.

matrix Σ denoted by $\mathbf{x} \sim N(\mu, \Sigma)$ if $f(\mathbf{x})$ is given by

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\tfrac{1}{2} (\mathbf{x} - \mu)^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \mu)\right] \tag{F.7.6}$$

where $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$ is a **symmetric** and **positive definite matrix** and $|\Sigma|$ is the **determinant** of Σ. The exponent

$$-\tfrac{1}{2} (\mathbf{x} - \mu)^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \mu) \tag{F.7.7}$$

is a quadratic form in $\mathbf{x}$ and Σ^{-1} denotes the inverse of Σ. It can be shown that $E(\mathbf{x}) = \mu$ and $\mathrm{Cov}(\mathbf{x}) = E\left[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^{\mathrm{T}}\right] = \Sigma$

To get a feel for this important distribution consider the case when $n = 2$. Then

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \\ \sigma_{12}^2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{F.7.8}$$

where ρ is the correlation coefficient between x_1 and x_2 (refer to (F.3.17)).

Hence

$$|\Sigma|^{\frac{1}{2}} = \sigma_1\sigma_2(1 - \rho^2)^{\frac{1}{2}}$$

$$\Sigma^{-1} = \tfrac{1}{\Sigma} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}$$

and

$$(\mathbf{x} - \mu)^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \mu) = (x_1 - \mu_1, x_2 - \mu_2) \Sigma^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}$$

$$= \frac{\sigma_2^2 (x_1 - \mu_1)^2 - 2\rho\sigma_1\sigma_2 (x_1 - \mu_1)(x_2 - \mu_2) + \sigma_2^2 (x_2 - \mu_2)^2}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)}.$$

In the special case when x_1 and x_2 are **uncorrelated**, that is, when $\rho = 0$, it follows that Σ reduces to a diagonal matrix given by

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$|\Sigma|^{\frac{1}{2}} = \sigma_1 \sigma_2$$

and

$$(\mathbf{x} - \mu)^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \mu) = \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2}.$$

Hence, the joint distribution becomes

$$\begin{aligned} f(x_1, x_2) &= \tfrac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\tfrac{(x_1-\mu_1)^2}{2\sigma_1^2} \right] \tfrac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\tfrac{(x_2-\mu_2)^2}{2\sigma_2^2} \right] \\ &= f_1(x_1) f_2(x_2). \end{aligned} \tag{F.7.9}$$

That is, x_1 and x_2 are **independent**. This result stating that if two normal variables are uncorrelated then they are also independent is one of the key distinguishing feature of this important class of distributions. Refer to Figure F.7.2 for contours of constant probability in the case of bivariate normal distribution for various values of ρ.

F.8 Functions of random variables

Let $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ be a random vector in $\mathbb{R}^2$ with $f(\mathbf{x})$ as its joint density. Let $\mathbf{y} \in \mathbb{R}^2$ be such that

$$\mathbf{y} = (y_1, y_2)^{\mathrm{T}} = (g_1(x_1), g_2(x_2))^{\mathrm{T}} = g(\mathbf{x}). \tag{F.8.1}$$

Clearly, $\mathbf{y}$ is a vector function of a random vector $\mathbf{x}$ and hence is also random. Let $h(\mathbf{y})$ be the distribution function of $\mathbf{y}$. The question is: given $f(\mathbf{x})$ and $\mathbf{y} = g(\mathbf{x})$, how to compute $h(\mathbf{y})$? In answering this question, recall that the Jacobian (Appendix C) of $g(\mathbf{x})$ is given by

$$D_g(\mathbf{x}) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{bmatrix}.$$

Let $|D_g(\mathbf{x})|$ denote the determinant of $D_g(\mathbf{x})$. Given $\mathbf{y} = (y_1, y_2)^{\mathrm{T}} \in \mathbb{R}^2$ define

$$S_{\mathbf{y}} = \{\mathbf{x} \mid g(\mathbf{x}) = \mathbf{y}\}.$$

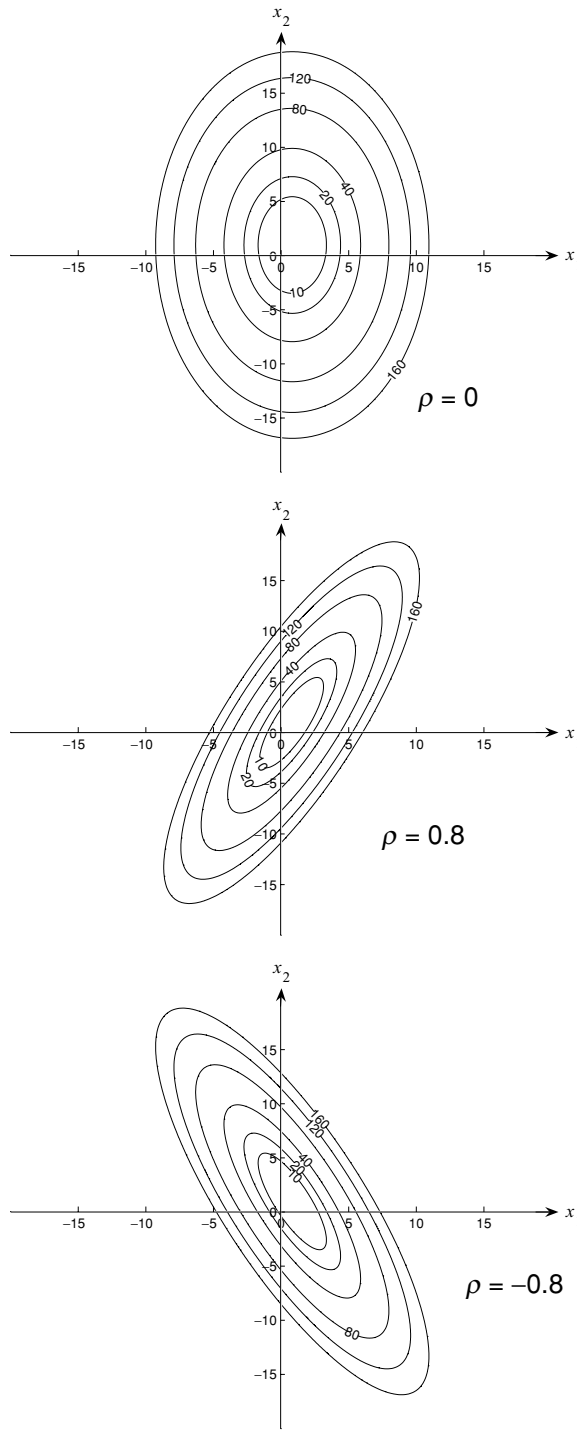

Fig. F.8.1 Contour plots of bivariate normal distribution for $\rho = 0$ and ± 0.8.

That is, $S_{\mathbf{y}}$ consists of all points in $\mathbb{R}^2$ that are mapped onto $\mathbf{y}$ by $g(\mathbf{x})$. Then, it can be shown (Papoulis(1984)) that

$$h(\mathbf{y}) = \Sigma_{\mathbf{x}\in S_{\mathbf{y}}} \frac{1}{|D_g(\mathbf{x})|} f(\mathbf{x}). \tag{F.8.2}$$

We now illustrate this using a couple of examples.

Example F.8.1 Let $\mathbf{y} = g(\mathbf{x}) = \mathbf{A}\mathbf{x}$ where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$ and $\mathbf{A} \in \mathbb{R}^{2\times 2}$ is a non-singular matrix. Let $f(\mathbf{x})$ and $h(\mathbf{y})$ be the probability density functions of $\mathbf{x}$ and $\mathbf{y}$ respectively. Then

$$y_1 = g_1(\mathbf{x}) = a_{11}x_1 + a_{12}x_2$$
$$y_2 = g_2(\mathbf{x}) = a_{21}x_1 + a_{22}x_2$$

and

$$D_g(\mathbf{x}) = \begin{bmatrix} a_{11} & a_{12} \\ \\ a_{21} & a_{22} \end{bmatrix} = \mathbf{A}$$

and

$$|D_g(\mathbf{x})| = |\mathbf{A}| \neq 0.$$

Since $\mathbf{A}$ is non-singular, the association between $\mathbf{x}$ and $\mathbf{y}$ is one-to-one and hence for each $\mathbf{x}$ there is a unique $\mathbf{y} = \mathbf{A}\mathbf{x}$. From (F.8.2), we obtain

$$h(\mathbf{y}) = \tfrac{1}{|\mathbf{A}|}\ f(\mathbf{A}^{-1}\mathbf{y}). \tag{F.8.3}$$

We now apply this derivation to the case of a linear transformation of a normal random vector. Let $\mathbf{x} \in \mathbb{R}^n$ with $\mathbf{x} \sim N(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})$, that is,

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_{\mathbf{x}}|^{\frac{1}{2}}} \exp\left[-(\mathbf{x}-\mu_x)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu_x)\right]. \tag{F.8.4}$$

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be non-singular and $\mathbf{b} \in \mathbb{R}^n$ define

$$\mathbf{y} = g(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$$

Then, $D_g = \mathbf{A}$ and

$$h(\mathbf{y}) = \tfrac{1}{|\mathbf{A}|}\ f(\mathbf{A}^{-1}(\mathbf{y}-\mathbf{b})) \tag{F.8.5}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_{\mathbf{x}}|^{\frac{1}{2}} |\mathbf{A}|} \exp\left[-\left[\mathbf{A}^{-1}(\mathbf{y}-\mathbf{b}) - \mu_x\right]^{\mathrm{T}} \Sigma_x^{-1} \left[\mathbf{A}^{-1}(\mathbf{y}-\mathbf{b}) - \mu_x\right]\right]. \tag{F.8.6}$$

But

$$\begin{aligned} &\left[\mathbf{A}^{-1}(\mathbf{y}-\mathbf{b}) - \mu_x\right]^{\mathrm{T}} \Sigma_x^{-1} \left[\mathbf{A}^{-1}(\mathbf{y}-\mathbf{b}) - \mu_x\right] \\ &\quad = \left[\mathbf{A}^{-1}(\mathbf{y} - (\mathbf{A}\mu_x + \mathbf{b}))\right]^{\mathrm{T}} \Sigma_x^{-1} \left[\mathbf{A}^{-1}(\mathbf{y} - (\mathbf{A}\mu_x + \mathbf{b})\right] \\ &\quad = (\mathbf{y}-\mu_y)^{\mathrm{T}} \Sigma_{\mathbf{y}}^{-1} (\mathbf{y}-\mu_y) \end{aligned} \tag{F.8.7}$$

where

$$\mu_{\mathbf{y}} = \mathbf{A}\mu_{\mathbf{x}} + \mathbf{b} \quad \text{and} \quad \Sigma_{\mathbf{y}} = \mathbf{A}\Sigma_{\mathbf{x}}\mathbf{A}^{\mathrm{T}}. \tag{F.8.8}$$

Similarly, it can be verified that

$$|\Sigma_{\mathbf{y}}| = |\mathbf{A}\Sigma_{\mathbf{x}}\mathbf{A}^{\mathrm{T}}| = |\mathbf{A}|^2|\Sigma_{\mathbf{x}}|. \tag{F.8.9}$$

Now, combining (F.8.7)–(F.8.9) with (F.5.6), it follows that

$$h(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_{\mathbf{y}}|^{\frac{1}{2}}} \exp\left[-(\mathbf{y}-\mu_{\mathbf{y}})^{\mathrm{T}}\Sigma_{\mathbf{y}}^{-1}(\mathbf{y}-\mu_{\mathbf{y}})\right] \tag{F.8.10}$$

that is,

$$\mathbf{y} \sim N(\mathbf{A}\mu_{\mathbf{x}} + b, \mathbf{A}\Sigma_{\mathbf{x}}\mathbf{A}^{\mathrm{T}}). \tag{F.8.11}$$

F.9 Distribution of quadratic forms of normal random vectors

Estimation of covariance of normal random vectors naturally leads to the analysis of the distribution of **quadratic forms of normal random vectors**. Since the notion of **chi-square** distribution plays a fundamental role in this analysis, we begin by describing this important family of distribution. A **non-negative** random variable z is said to be chi-square distributed with k-degrees of freedom, for some integer $(k \geq 1)$, denoted by $z \sim \chi^2(k)$, if the density function of z is given by

$$f(z) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} z^{\frac{k}{2}-1}\mathrm{e}^{-\frac{z}{2}}, \quad \text{for } z > 0 \tag{F.9.1}$$

where

$$\Gamma(r) = \int_0^\infty x^{r-1}\mathrm{e}^{-x}\mathrm{d}x \tag{F.9.2}$$

is the standard **gamma function**. It can be verified that

(a) $\Gamma(1) = 1$.
(b) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.
(c) $\Gamma(r+1) = r\Gamma(r)$ if r is a positive integer.

We now state several facts.

(P1) Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}} \sim N(0, \mathbf{I})$. Then

$$\mathbf{z} = \mathbf{x}^{\mathrm{T}}\mathbf{x} = \sum_{i=1}^{n} x_i^2 \sim \chi^2(n). \tag{F.9.3}$$

The mean and the variance of this random variable z with distribution $\chi^2(n)$ are given by

$$E(z) = n \quad \text{and} \quad \mathrm{Var}(z) = 2n.$$

(P2) If $z_1 \sim \chi^2(n_1)$ and $z_2 \sim \chi^2(n_2)$ are **independent**, then

$$z = z_1 + z_2 \sim \chi^2(n_1 + n_2). \tag{F.9.4}$$

(P3) Let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x} \sim N(0, \mathbf{I})$. Let $\mathbf{A}$ and $\mathbf{B}$ be two symmetric $n \times n$ matrices which are **idempotent**, that is $\mathbf{A}^2 = \mathbf{A}$ and $\mathbf{B}^2 = \mathbf{B}$. Then as random variables $\mathbf{z}_1 = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ and $\mathbf{z}_2 = \mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}$ are **independent** if $\mathbf{AB} = 0$. First recall that since $\mathbf{A}$ and $\mathbf{B}$ are symmetric and idempotent it follows that

$$\mathbf{A} = \mathbf{A}^2 = \mathbf{A}\mathbf{A} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$$

and

$$\mathbf{B} = \mathbf{B}^2 = \mathbf{B}\mathbf{B} = \mathbf{B}^{\mathrm{T}}\mathbf{B}.$$

Hence

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{x})^{\mathrm{T}}(\mathbf{A}\mathbf{x}) = \mathbf{x}_a^{\mathrm{T}}\mathbf{x}_a$$
$$\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{x} = (\mathbf{B}\mathbf{x})^{\mathrm{T}}(\mathbf{B}\mathbf{x}) = \mathbf{x}_b^{\mathrm{T}}\mathbf{x}_b$$

where

$$\mathbf{x}_a = \mathbf{A}\mathbf{x} \quad \text{and} \quad \mathbf{x}_b = \mathbf{B}\mathbf{x},$$

which are linear transformations of the normal random vector $\mathbf{x}$. Now applying the results in Example F.8.1 to $\mathbf{x}_a$ and $\mathbf{x}_b$, it follows that

$$\mathbf{x}_a \sim N(0, \mathbf{A}\mathbf{I}\mathbf{A}^{\mathrm{T}}) = N(0, \mathbf{A})$$
$$\mathbf{x}_b \sim N(0, \mathbf{B}\mathbf{I}\mathbf{B}^{\mathrm{T}}) = N(0, \mathbf{B}).$$

The covariance between $\mathbf{x}_a$ and $\mathbf{x}_b$ is then given by

$$\begin{aligned}
\mathrm{Cov}(\mathbf{x}_a, \mathbf{x}_b) &= E\left[\mathbf{x}_a\mathbf{x}_b^{\mathrm{T}}\right] \\
&= E\left[\mathbf{A}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{B}\right] \\
&= \mathbf{A}E\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right]\mathbf{B} \\
&= \mathbf{A}\mathbf{B} \\
&= 0.
\end{aligned}$$

That is, if $\mathbf{AB} = 0$, then $\mathbf{x}_a$ and $\mathbf{x}_b$ are uncorrelated normal random vectors and hence are independent by the results of Section F.7.

(P4) Let $\mathbf{A}$ be a symmetric and idempotent matrix with $\mathbf{A} \neq \mathbf{I}$, the identity matrix. Then it is well known (Appendix B) that the eigenvalues of $\mathbf{A}$ are either 0 or 1 and the number of non-zero eigenvalues is equal to the rank of $\mathbf{A}$. Since $\mathbf{A}$ is symmetric there exists an orthogonal matrix $\mathbf{P}$ such that

$$\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathrm{T}}$$

where $\mathbf{P}\mathbf{P}^{\mathrm{T}} = \mathbf{I}$ and $\mathbf{\Lambda}$ is the diagonal matrix with k non-zero eigenvalues. Without loss of generality, let

$$\begin{aligned} \lambda_1 \quad &= \lambda_2 \quad = \cdots = \lambda_k = 1 \\ \lambda_{k+1} &= \lambda_{k+2} = \ldots = \lambda_n = 0. \end{aligned}$$

Clearly, such a matrix $\mathbf{A}$ is singular. Let $\mathbf{x} \sim N(0, \mathbf{I})$ and $\mathbf{z} = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$. Then

$$\begin{aligned} \mathbf{z} &= \mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathrm{T}}\mathbf{x} \\ &= (\mathbf{P}^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}\mathbf{\Lambda}(\mathbf{P}^{\mathrm{T}}\mathbf{x}) \\ &= \mathbf{y}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{y} \quad \text{where } \mathbf{y} = \mathbf{P}\mathbf{x} \\ &= \textstyle\sum_{i=1}^{k} y_i^2. \end{aligned} \tag{F.9.5}$$

Since $\mathbf{x} \sim N(0, \mathbf{I})$, by the results of section F.7, it follows that $\mathbf{y} \sim N(0, \mathbf{P}\mathbf{I}\mathbf{P}^{\mathrm{T}}) = N(0, \mathbf{I})$, that is, the components of $\mathbf{y}$ are standard normal variables. Combining this with (F.9.3), it follows $\mathbf{z}$ in (F.9.5) is $\chi^2(k)$, where the degree of freedom is equal to the rank of the matrix $\mathbf{A}$.

In closing, we present an application of these results. Let $\mathbf{x} \sim N(0, \mathbf{I})$. Then

$$\begin{aligned} \mathbf{x}^{\mathrm{T}}\mathbf{x} &= \textstyle\sum_{i=1}^{n} x_i^2 \\ &= \textstyle\sum_{i=1}^{n} (x_i - \bar{\mathbf{x}})^2 + n\bar{\mathbf{x}}^2 \end{aligned} \tag{F.9.6}$$

where

$$\bar{\mathbf{x}} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} x_i.$$

We now analyze the two terms on the r.h.s. of (F.9.6). Consider first

$$\begin{aligned} z_1 = n\bar{\mathbf{x}}^2 \quad &= n\left[\tfrac{1}{n} \textstyle\sum_{i=1}^{n} x_i\right]^2 \\ &= \mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x} \end{aligned} \tag{F.9.7}$$

where $\mathbf{B}$ is a **rank-one, symmetric, idempotent, outer-product** matrix given by

$$\mathbf{B} = \tfrac{1}{n}\, \iota\, \iota^{\mathrm{T}} \tag{F.9.8}$$

with $\iota = (1, 1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^n$. Similarly,

$$z_2 = \textstyle\sum_{i=1}^{n} (x_i - \bar{\mathbf{x}})^2 = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$$

where $\mathbf{A}$ is a symmetric matrix given by

$$\mathbf{A} = \mathbf{I} - \tfrac{1}{n}\, \iota\, \iota^{\mathrm{T}} = \mathbf{I} - \mathbf{B}.$$

It can be verified that $\mathbf{A}$ is **symmetric, idempotent** matrix whose rank is equal to its trace which is $(n - 1)$. Further it can be verified that

$$\mathbf{A}\mathbf{B} = (\mathbf{I} - \mathbf{B})\mathbf{B} = \mathbf{B} - \mathbf{B}^2 = \mathbf{B} - \mathbf{B} = 0.$$

Hence, by property (P3) proved above, z_1 and z_2 are **independent**. Further,

$$z_1 \sim \chi^2(1) \quad \text{and} \quad z_2 \sim \chi^2(n-1)$$

and hence

$$\mathbf{x}^{\mathrm{T}}\mathbf{x} = z_1 + z_2 \sim \chi^2(n)$$

which is to be expected from the definition.

F.10 Stationary random sequence and white noise

Let $\mathbf{x} = \{x_0, x_1, x_2, x_3, \ldots\}$ be a sequence of (scalar) random variables (also called a **time series**). Let $p(x_t)$ be the probability density function of x_t for $t = 0, 1, 2, \ldots$. In general, $p(x_t)$ may vary with time t. Then

$$E(x_t) = \mu_t = \int_{\mathbb{R}} x_t \, p(x_t) \mathrm{d}x_t \tag{F.10.1}$$

is the **mean** of x_t. Similarly,

$$\begin{aligned} & E[(x_t - \mu_t)(x_{t-j} - \mu_{t-j})] \\ &= \gamma_{t,j} \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} (x_t - \mu_t)(x_{t-j} - \mu_{t-j}) p(x_t, x_{t-j}) \mathrm{d}x_t \mathrm{d}x_{t-j} \end{aligned} \tag{F.10.2}$$

where $p(x_t, x_{t-j})$ is the joint density of x_t and x_{t-j} is called the **covariance** between x_t and x_{t-j}. Clearly, $\gamma_{t,0} = \sigma_t^2$ is the **variance** of x_t.

The sequence x is said to be **weakly** (or **second-order**) **stationary** if

$$E(x_t) \equiv \mu \tag{F.10.3}$$

and

$$E[(x_t - \mu_t)(x_{t-j} - \mu_{t-j})] = \gamma_j \tag{F.10.4}$$

that is, the mean and the covariance are **independent** of t. The plot of γ_j vs. j is called the **covariance function**. It can be verified $\gamma_j = \gamma_{-j}$, that is, γ_j is a symmetric function with $\gamma_0 = \sigma^2$, the common variance of x_t. If $\gamma_j \equiv 0$ for $j \neq 0$, then the sequence x is said to be **uncorrelated**. A weakly stationary random sequence that is uncorrelated is called a **stationary white noise**. In addition if each of the elements x_t of the sequence has a common normal distribution, say $N(0, \sigma^2)$, then it is called **Gaussian white noise**. All of these notions readily carry over to the sequence of random vectors as well.

Notes and references

The material covered in this appendix is rather standard in the first year graduate level course in probability theory. There are numerous excellent expositions of these topics. We particularly recommend the following: Brammer and Siffling (1989), Feller (1957), Sage and Melsa (1971), and Papoulis (1984). For a comprehensive discussion of time series refer to Hamilton (1994) and Harvey (1989).