Chapter 1

Fundamental Concepts of Probability Theory

1.1 Probability Space

1.1.1 The Probability Triplet

The probability space is formally defined through the probability triplet $(\Omega, \mathcal{B}, P)$ where,

- Ω: is the sample space, which contains all possible outcomes of an experiment.
- $\mathcal{B}$: is a set of subsets of Ω (a Borel field — a closed set under operations of: union, intersection and complement)
- P: is a scalar function defined on $\mathcal{B}$, called the probability function or probability measure.

Each set $B \in \mathcal{B}$ is called an event, that is, B is a collection of specific possible outcomes. In what follows, the mathematical details corresponding to the field $\mathcal{B}$ will be ignored (e.g., see Chung [26], for a detailed treatment). The values $\omega \in \Omega$ are the realizations, and for each set $B \in \mathcal{B}$, the function $P(B)$ defines the probability that the realization ω is in B. The quantity P is a probability function if it satisfies the following axioms:

1. $0 \leq P(B) \leq 1$, for all $B \in \mathcal{B}$
2. $P(\Omega) = 1$
3. $P(\bigcup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} P(B_i)$, for all disjoint sequences of $B_i \in \mathcal{B}$.

1.1.2 Conditional Probability

If A and B are two events and $P(B) \neq 0$, the conditional probability of A given B is defined as

$$P(A|B) \equiv P(A \cap B)/P(B) \tag{1.1}$$

The events A and B are statistically independent if $P(A|B) = P(A)$. Consequently, $P(A \cap B) = P(A)P(B)$.

Analogously,

$$P(B|A) = \frac{P(B \cap A)}{P(A)}, \tag{1.2}$$

for all $P(A) \neq 0$.

Combining the two relations above we have:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{1.3}$$

which is known as Bayes rule (or theorem) for probabilities. This relation is useful when we need to reverse the condition of events.

1.2 Random Variables

A scalar $\mathrm{x}(\omega)$ random variable (r.v.) is a function, whose value x is determined by the result ω of a random experiment. Note the typographical distinction between both quantities. In other words, an r.v. $\mathrm{x}(\omega)$ attributes a real number x to each point of the sample space. The particular value x assumed by the random variable is referred to as a *realization*. A random variable is defined in such a way that all sets $B \subset \Omega$ of the form

$$B = \{\omega : \mathrm{x}(\omega) \leq \xi\} \tag{1.4}$$

are in $\mathcal{B}$, for any value of $\xi \in R^1$.

1.2.1 Distribution and Density Functions

Each r.v. has a distribution function defined as

$$F_{\mathrm{X}}(x) \equiv P(\{\omega : \mathrm{x}(\omega) \leq x\}), \tag{1.5}$$

which represents the probability that x is less than or equal to x.

It follows, directly from the properties of the probability measure given above, that $F_{\mathrm{X}}(x)$ should be a non-decreasing function of x, with $F_{\mathrm{X}}(-\infty) = 0$ and $F_{\mathrm{X}}(\infty) = 1$. Under reasonable conditions, we can define a function called a probability density function, derived from the distribution function:

$$p_{\mathrm{X}}(x) \equiv \frac{dF_{\mathrm{X}}(x)}{dx}, \tag{1.6}$$

Table 1.1: Properties of probability density functions and distribution functions

.

$F_{\mathrm{X}}(-\infty) = 0$		(a)
$F_{\mathrm{X}}(+\infty) = 1$		(b)
$F_{\mathrm{X}}(x_1) \leq F_{\mathrm{X}}(x_2)$,	for all $x_1 \leq x_2$	(c)
$p_{\mathrm{X}}(x) \geq 0$,	for all x	(d)
$\int_{-\infty}^{\infty} p_{\mathrm{X}}(x)\,dx = 1$		(e)

Consequently, the inverse relation

$$F_{\mathrm{X}}(x) = \int_{-\infty}^{x} p_{\mathrm{X}}(s)\,ds\,, \tag{1.7}$$

provides the distribution function. The probability density function should be non-negative, and its integral over the real line should be unity. Table 1.1 presents a summary of the properties of probability density functions and distribution functions.

A few examples of continuous distribution functions are given below:

(I) Uniform:

$$p_{\mathrm{X}}(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \tag{1.8}$$

$$F_{\mathrm{X}}(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases} \tag{1.9}$$

(ii) Exponential:

$$p_{\mathrm{X}}(x) = \begin{cases} \frac{1}{a} e^{-x/a} & 0 < x \\ 0 & \text{otherwise} \end{cases} \tag{1.10}$$

$$F_{\mathrm{X}}(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-x/a} & x \geq 0 \end{cases} \tag{1.11}$$

(iii) Rayleigh:

$$p_{\mathrm{X}}(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{a^2} e^{-x^2/2a^2} & x \geq 0 \end{cases} \tag{1.12}$$

$$F_{\mathrm{X}}(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-x^2/2a^2} & x \geq 0 \end{cases} \tag{1.13}$$

(iv) Gaussian:

$$p_{\mathrm{X}}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{1.14}$$

$$F_{\mathrm{X}}(x) = \operatorname{erf}\left(\frac{x-\mu}{\sigma}\right) \tag{1.15}$$

where $\operatorname{erf}(x)$ is the error function (Arfken [5], p. 568):

$$\operatorname{erf}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\,dy \tag{1.16}$$

Remark: An r.v. with Gaussian distribution is said to be normally distributed, with mean μ and variance σ^2 (see following section) and is represented symbolically by $\mathrm{x} \sim \mathcal{N}(\mu, \sigma^2)$.

(iv) χ^2 (Chi–Square):

$$p_{\mathrm{X}}(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{1.17}$$

$$F_{\mathrm{X}}(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^x u^{\nu/2-1} e^{-u/2}\, du \tag{1.18}$$

where $\Gamma(\nu)$ is the gamma function (Arfken [5], Chapter 10):

$$\Gamma(\nu) = \int_0^\infty t^{n-1} e^{-t}\, dt \tag{1.19}$$

Remarks: An r.v. that is χ^2 distributed has the form: $\chi^2 = \mathrm{x}_1^2 + \mathrm{x}_2^2 + \cdots + \mathrm{x}_\nu^2$, with the variables x_i, for $i = 1, 2, \cdots, \nu$, being normally distributed with mean zero and unity variance.

1.2.2 Expectations and Moments

The mean of an r.v. x is defined as

$$\mathcal{E}\{\mathrm{x}\} \equiv \int_{-\infty}^{\infty} x\, p_{\mathrm{X}}(x)\, dx\,. \tag{1.20}$$

In this course we will use interchangeably the expressions expected value and expectation as synonyms for mean. There are extensions of this definition for those cases in which the probability density $p_{\mathrm{X}}(x)$ does not exist; however in the context that interests us, the definition above is sufficient. Any measurable function of an r.v. is also an r.v. and its mean is given by:

$$\mathcal{E}\{f(\mathrm{x})\} = \int_{-\infty}^{\infty} f(x)\, p_{\mathrm{X}}(x)\, dx\,. \tag{1.21}$$

In particular, if $f(\mathrm{x}) = a = const.$, $\mathcal{E}\{a\} = a$, due to property (e) in Table 1.1.

If $f(\mathrm{x}) = a_1 g_1(\mathrm{x}) + a_2 g_2(\mathrm{x})$ then

$$\mathcal{E}\{a_1 g_1(\mathrm{x}) + a_2 g_2(\mathrm{x})\} = a_1 \mathcal{E}\{g_1(\mathrm{x})\} + a_2 \mathcal{E}\{g_2(\mathrm{x})\}. \tag{1.22}$$

A function of special interest is $f(\mathrm{x}) = \mathrm{x}^n$, where n is a positive integer. The means

$$\mathcal{E}\{\mathrm{x}^n\} \equiv \int_{-\infty}^{\infty} x^n p_{\mathrm{X}}(x) dx\,, \tag{1.23}$$

define the moments of order n of x. In particular, $\mathcal{E}\{\mathrm{x}^2\}$ is called the mean–square value. The expectations

$$\mathcal{E}\{(\mathrm{x} - \mathcal{E}\{\mathrm{x}\})^n\} \equiv \int_{-\infty}^{\infty} (x - \mathcal{E}\{\mathrm{x}\})^n p_{\mathrm{X}}(x) dx\,, \tag{1.24}$$

define the n–th moments of x about its mean (n–th central moment).

The second moment of x about its mean is called the variance of x, and is given by:

$$\begin{aligned} var(\mathrm{x}) \equiv \mathcal{E}\{(\mathrm{x} - \mathcal{E}\{\mathrm{x}\})^2\} &= \mathcal{E}\{\mathrm{x}^2\} - 2\mathcal{E}\{\mathrm{x}\mathcal{E}\{\mathrm{x}\}\} + (\mathcal{E}\{\mathrm{x}\})^2 \\ &= \mathcal{E}\{\mathrm{x}^2\} - (\mathcal{E}\{\mathrm{x}\})^2 . \end{aligned} \tag{1.25}$$

That is, the variance is the mean–square minus the square of the mean. Finally, the standard deviation is defined as the square–root of the variance:

$$\sigma(\mathrm{x}) \equiv \sqrt{[var(\mathrm{x})]}. \tag{1.26}$$

It is worth mentioning at this point that in many cases, the mean value of an r.v. is used as a guess (or estimate) for the true value of that variable. Other quantities of interest in this sense are the median, the mid–range, and the mode values. The median μ_1 is given by

$$\int_{-\infty}^{\mu_1} p_{\mathrm{X}}(x)\, dx = \int_{\mu_1}^{\infty} p_{\mathrm{X}}(x)\, dx = \frac{1}{2}\,, \tag{1.27}$$

the mid–range μ_∞ is given by

$$\mu_\infty = \frac{\max_x(x) + \min_x(x)}{2} \tag{1.28}$$

and the mode m is given by

$$\left.\frac{dp_{\mathrm{X}}(x)}{dx}\right|_{x=m} = 0 \tag{1.29}$$

The median divides the probability density function in two, each one covering the same area. The mode corresponds to values of the random variable for which the probability density function is maximum, that is, it corresponds to the most likely value. The importance, and more general meaning, of these quantities will become clear as we advance.

1.2.3 Characteristic Function

An r.v. can be represented, alternatively, by its characteristic function which is defined as

$$\phi_{\mathrm{X}}(u) \equiv \mathcal{E}\{\exp(iu\mathrm{x})\}\,, \tag{1.30}$$

where $i = \sqrt{-1}$.

According to the definition (1.20) of mean we see that the characteristic function is nothing more than the Fourier transform of the density function:

$$\phi_{\mathrm{X}}(u) = \int_{-\infty}^{\infty} \exp(iux) p_{\mathrm{X}}(x)\, dx\,, \tag{1.31}$$

from this it follows that the probability density is the inverse Fourier transform of the characteristic function, that is,

$$p_{\mathrm{X}}(x) = (1/2\pi) \int_{-\infty}^{\infty} \exp(-iux) \phi_{\mathrm{X}}(u)\, du\,. \tag{1.32}$$

Let us now take the derivative of the characteristic function (1.31) with respect to u:

$$\begin{aligned} \frac{d\phi_{\mathrm{X}}(u)}{du} &= \frac{d}{du}\int_{-\infty}^{\infty} \exp(iux) p_{\mathrm{X}}(x)\, dx\,, \\ &= \int_{-\infty}^{\infty} \frac{d\exp(iux)}{du} p_{\mathrm{X}}(x)\, dx\,, \\ &= i\mathcal{E}\{\mathrm{x}\exp(iu\mathrm{x})\}\,, \end{aligned} \tag{1.33}$$

where we used the definition of characteristic function to get the last equality. Notice that by choosing calculate the expression above for at $u = 0$ we have

$$\left.\frac{d\phi_{\mathrm{X}}(u)}{du}\right|_{u=0} = i\mathcal{E}\{\mathrm{x}\}\,, \tag{1.34}$$

or better yet,

$$\mathcal{E}\{\mathrm{x}\} = \frac{1}{i}\left.\frac{d\phi_{\mathrm{X}}(u)}{du}\right|_{u=0}\,, \tag{1.35}$$

which give an alternative way of calculating the first moment, if the characteristic function is given. As a matter of fact moments of order n can be calculated analogously, by taking n derivatives of the characteristic function and evaluating the result at $u = 0$. This procedure produces the equation

$$\mathcal{E}\{\mathrm{x}^n\} = \frac{1}{i^n}\left.\frac{d^n\phi_{\mathrm{X}}(u)}{du^n}\right|_{u=0}\,, \tag{1.36}$$

for the n–th moment.

1.3 Jointly Distributed Random Variables

1.3.1 Distribution, Density Function and Characteristic Function

The r.v.'s $\mathrm{x}_1, \cdots, \mathrm{x}_n$ are said to be jointly distributed if they are defined in the same probability space. They can be characterized by the joint distribution function

$$F_{\mathrm{X}_1\cdots\mathrm{X}_n} \equiv P\{\omega : \mathrm{x}_1 \le x_1, \cdots, \mathrm{x}_n \le x_n\} \tag{1.37}$$

where

$$\{\omega : \mathrm{x}_1 \le x_1, \cdots, \mathrm{x}_n \le x_n\} \equiv \{\mathrm{x}_1(\omega) \le x_1\} \cap \cdots \cap \{\mathrm{x}_n(\omega) \le x_n\} \tag{1.38}$$

or alternatively, by their joint density function:

$$F_{\mathrm{X}_1\cdots\mathrm{X}_n}(x_1, \cdots, x_n) \equiv \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} p_{\mathrm{X}_1\cdots\mathrm{X}_n}(x'_1, \cdots, x'_n) dx'_1 \cdots dx'_n\,, \tag{1.39}$$

from which it follows that

$$p_{\mathrm{X}_1\cdots\mathrm{X}_n}(x_1, \cdots, x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F_{\mathrm{X}_1\cdots\mathrm{X}_n}(x_1, \cdots, x_n) \tag{1.40}$$

assuming the existence of the derivatives.

The characteristic function of jointly distributed r.v.'s $\mathrm{x}_1, \cdots, \mathrm{x}_n$ is defined as:

$$\phi_{\mathrm{X}_1\cdots\mathrm{X}_n}(u_1, \cdots, u_n) \equiv \mathcal{E}\{\exp\left(i\sum_{j=1}^{n} u_j \mathrm{x}_j\right)\}\,. \tag{1.41}$$

1.3.2 Expectations and Moments

If f is a function of jointly distributed r.v.'s $\mathrm{x}_1, \cdots, \mathrm{x}_n$, and $\mathrm{y} = f(\mathrm{x}_1, \cdots, \mathrm{x}_n)$, then

$$\mathcal{E}\{\mathrm{y}\} \equiv \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \cdots, x_n) p_{\mathrm{x}_1 \cdots \mathrm{x}_n}(x_1, \cdots, x_n)\, dx_1 \cdots dx_n\,. \tag{1.42}$$

The expected value of x_k is given by

$$\mathcal{E}\{\mathrm{x}_k\} \equiv \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_k p_{\mathrm{x}_1 \cdots \mathrm{x}_n}(x_1, \cdots, x_n)\, dx_1 \cdots dx_n \tag{1.43}$$

and its second-order moment is given by

$$\mathcal{E}\{\mathrm{x}_k^2\} \equiv \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_k^2 p_{\mathrm{x}_1 \cdots \mathrm{x}_n}(x_1, \cdots, x_n)\, dx_1 \cdots dx_n\,. \tag{1.44}$$

Moments of higher order and central moments can be introduced in analogy to the definitions in Section 1.2.2. Joint moments and joint central moments can be defined as:

$$\mathcal{E}\{\mathrm{x}_k^\alpha \mathrm{x}_\ell^\beta\} \tag{1.45}$$

and

$$\mathcal{E}\{[\mathrm{x}_k - \mathcal{E}\{\mathrm{x}_k\}]^\alpha \, [\mathrm{x}_\ell - \mathcal{E}\{\mathrm{x}_\ell\}]^\beta\}\,, \tag{1.46}$$

respectively, where α and β are positive integers.

Notice that the characteristic function, of the jointly distributed r.v.'s, gives a convenient way of computing moments, just as it did in the scalar case. Taking the first derivative of the characteristic function (1.41) with respect to component u_k we have

$$\frac{\partial \phi_{\mathrm{x}_1 \cdots \mathrm{x}_n}(u_1, \cdots, u_n)}{\partial u_k} \;=\; i\mathcal{E}\{\mathrm{x}_k \exp\left(i \sum_{j=1}^{n} u_j \mathrm{x}_j\right)\}\,. \tag{1.47}$$

Evaluating this derivative at $(u_1, \cdots, u_n) = (0, \cdots, 0)$ provides a way to compute the first moment with respect to component x_k, that is,

$$\mathcal{E}\{\mathrm{x}_k\} \;=\; \frac{1}{i} \left. \frac{\partial \phi_{\mathrm{x}_1 \cdots \mathrm{x}_n}(u_1, \cdots, u_n)}{\partial u_k} \right|_{(u_1, \cdots, u_n) = (0, \cdots, 0)} \tag{1.48}$$

Successive n derivatives, with respect to arbitrary n components of $(u_1, \cdots, u_n)$ produce the n–th non–central moment

$$\mathcal{E}\{\mathrm{x}_k \mathrm{x}_l \cdots\} \;=\; \frac{1}{i^n} \left. \frac{\partial \phi_{\mathrm{x}_1 \cdots \mathrm{x}_n}(u_1, \cdots, u_n)}{\partial u_k \partial u_l \cdots} \right|_{(u_1, \cdots, u_n) = (0, \cdots, 0)}\,. \tag{1.49}$$

Of fundamental importance is the concept of covariance between x_k and x_ℓ, defined as:

$$cov(\mathrm{x}_k, \mathrm{x}_\ell) \equiv \mathcal{E}\{[\mathrm{x}_k - \mathcal{E}\{\mathrm{x}_k\}]\,[\mathrm{x}_\ell - \mathcal{E}\{\mathrm{x}_\ell\}]\}\,. \tag{1.50}$$

We have that

$$cov(\mathrm{x}_k, \mathrm{x}_\ell) = \mathcal{E}\{\mathrm{x}_k \mathrm{x}_\ell\} - \mathcal{E}\{\mathrm{x}_k\}\mathcal{E}\{\mathrm{x}_\ell\} \tag{1.51}$$

and also,

$$cov(\mathrm{x}_k, \mathrm{x}_k) = var(\mathrm{x}_k)\,. \tag{1.52}$$

The ratio

$$\rho(\mathrm{x}_k, \mathrm{x}_\ell) \equiv \frac{cov(\mathrm{x}_k, \mathrm{x}_\ell)}{\sigma(\mathrm{x}_k)\sigma(\mathrm{x}_\ell)} \tag{1.53}$$

defines the correlation coefficient between x_k and x_ℓ. Therefore, $\rho(\mathrm{x}_k, \mathrm{x}_k) = 1$.

It is of frequent interest to obtain the probability distribution or density function of a random variable, given its corresponding joint function. That is, consider two r.v.'s x_1 and x_2, jointly distributed, then

$$F_{\mathrm{x}_1}(x_1) = F_{\mathrm{x}_1\mathrm{x}_2}(x_1, \infty) = \int_{-\infty}^{x_1}\int_{-\infty}^{\infty} p_{\mathrm{x}_1\mathrm{x}_2}(s_1, s_2)\, ds_1 ds_2\,, \tag{1.54}$$

and analogously, $F_{\mathrm{x}_2}(x_2) = F_{\mathrm{x}_1\mathrm{x}_2}(\infty, x_2)$, where $F_{\mathrm{x}_1}(x_1)$ and $F_{\mathrm{x}_2}(x_2)$, are referred to as marginal distribution functions. The marginal density function is then given by

$$p_{\mathrm{x}_1}(x_1) = \frac{\partial F_{\mathrm{x}_1\mathrm{x}_2}(x_1, \infty)}{\partial x_1} = \int_{-\infty}^{\infty} p_{\mathrm{x}_1\mathrm{x}_2}(x_1, x_2)\, dx_2\,. \tag{1.55}$$

It is convenient, at this point, to introduce a more compact notation utilizing vectors. Define the vector random variable (or simply the random vector) in n dimensions as:

$$\mathbf{x} = (\mathrm{x}_1 \mathrm{x}_2 \cdots \mathrm{x}_n)^T \tag{1.56}$$

where lower case bold letters refer to vectors, and T refers to the transposition operation. By analogy with the notation we have utilized up to here, we will refer to the value assumed by the random vector $\mathbf{x}$ as $\boldsymbol{x} = (x_1 x_2 \cdots x_n)^T$. In this manner,

$$p_{\mathbf{x}}(\boldsymbol{x}) \equiv p_{\mathrm{x}_1\mathrm{x}_2\cdots\mathrm{x}_n}(x_1, x_2, \cdots, x_n) \tag{1.57}$$

Likewise, the probability distribution can be written as

$$\begin{aligned} F_{\mathbf{x}}(\boldsymbol{x}) &\equiv \int_{-\infty}^{\boldsymbol{x}} p_{\mathbf{x}}(\boldsymbol{x}')d\boldsymbol{x}' \\ &= \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n} p_{\mathrm{x}_1\cdots\mathrm{x}_n}(x_1', \cdots, x_n')dx_1'\cdots dx_n'\,, \end{aligned} \tag{1.58}$$

where we call attention for the notation $d\boldsymbol{x} = d\mathrm{x}_1 \cdots d\mathrm{x}_n$, and similarly the probability density function becomes

$$p_{\mathbf{x}}(\boldsymbol{x}) = \frac{\partial^n F_{\mathbf{x}}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{\partial^n F_{\mathbf{x}}(\boldsymbol{x})}{\partial x_1 \cdots \partial x_n} \tag{1.59}$$

The marginal probability density can be written as

$$p_{\mathrm{x}_k}(x_k) = \frac{\partial F_{\mathrm{x}_k}(x_k)}{\partial x_k} = \int_{-\infty}^{\infty} p_{\mathbf{x}}(\boldsymbol{x}')\, d\boldsymbol{x}'_{-k}\,, \tag{1.60}$$

where $d\boldsymbol{x}_{-k} = dx_1 \cdots dx_{k-1} dx_{k+1} \cdots dx_n$.

According to the definition of mean of a random variable, the mean of a random vector is given by the mean of its components:

$$\mathcal{E}\{\mathbf{x}\} = \begin{bmatrix} \mathcal{E}\{\mathrm{x}_1\} \\ \vdots \\ \mathcal{E}\{\mathrm{x}_n\} \end{bmatrix} = \begin{bmatrix} \int_{-\infty}^{\infty} x_1 p_{\mathbf{x}}(\boldsymbol{x}') d\boldsymbol{x}' \\ \vdots \\ \int_{-\infty}^{\infty} x_n p_{\mathbf{x}}(\boldsymbol{x}') d\boldsymbol{x}' \end{bmatrix} \tag{1.61}$$

Analogously, the mean of a random matrix is the mean of the matrix elements. The matrix formed by the mean of the outer product of the vector $\mathbf{x} - \mathcal{E}\{\mathbf{x}\}$ with itself is the $n \times n$ covariance matrix:

$$\begin{aligned} \mathbf{P}_{\mathbf{x}} &= \mathcal{E}\{(\mathbf{x} - \mathcal{E}\{\mathbf{x}\})(\mathbf{x} - \mathcal{E}\{\mathbf{x}\})^T\} \\ &= \begin{bmatrix} var(x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_n) \\ cov(x_2, x_1) & var(x_2) & \cdots & cov(x_1, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(x_n, x_1) & cov(x_n, x_2) & \cdots & var(x_n) \end{bmatrix}. \end{aligned} \tag{1.62}$$

Notice that $\mathbf{P}_{\mathbf{x}}$ is a symmetric positive semi-definite matrix, that is, $\mathbf{y}\mathbf{P}_{\mathbf{x}}\mathbf{y}^T \geq \mathbf{0}$, for all $\mathbf{y} \in R^n$.

Two scalar r.v.'s x and y are said to be independent if any of the (equivalent) conditions are satisfied:

$$F_{\mathrm{xy}}(x, y) = F_{\mathrm{x}}(x) F_{\mathrm{y}}(y) \tag{1.63a}$$

$$p_{\mathrm{xy}}(x, y) = p_{\mathrm{x}}(x) p_{\mathrm{y}}(y) \tag{1.63b}$$

$$\mathcal{E}\{f(\mathrm{x}) g(\mathrm{y})\} = \mathcal{E}\{f(\mathrm{x})\} \mathcal{E}\{g(\mathrm{y})\} \tag{1.63c}$$

Analogously, two vector r.v.'s $\mathbf{x}$ and $\mathbf{y}$ are said to be jointly independent if

$$p_{\mathbf{xy}}(\boldsymbol{x}, \boldsymbol{y}) = p_{\mathbf{x}}(\boldsymbol{x}) p_{\mathbf{y}}(\boldsymbol{y}) \tag{1.64}$$

We say that two jointly distributed random vectors $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated if

$$cov(\mathbf{x}, \mathbf{y}) = \mathbf{0}\,, \tag{1.65}$$

since the correlation coefficient defined in (1.53) is null. As a matter of fact, two r.v.'s are said to be orthogonal when

$$\mathcal{E}\{\mathbf{x}\mathbf{y}^T\} = \mathbf{0}\,. \tag{1.66}$$

This equality is often referred to as the *orthogonality principle*.

The n r.v.'s $\{\mathrm{x}_1, \cdots, \mathrm{x}_n\}$ are said to be jointly Gaussian, or jointly normal, if their joint probability density function is given by

$$p_{\mathbf{x}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2} |\mathbf{P}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \mu)^T \mathbf{P}^{-1} (\boldsymbol{x} - \mu)\right], \tag{1.67}$$

where the notation $|\mathbf{P}|$ stands for the determinant of $\mathbf{P}$, and $\mathbf{P}^{-1}$ refers to the inverse of the matrix $\mathbf{P}$. The vector $\mathbf{x}$ is said to be normally distributed or Gaussian, with mean $\mu = \mathcal{E}\{\mathbf{x}\}$ and covariance $\mathbf{P}$, and is abbreviated by $\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{P})$. Observe that, in order to simplify the notation, we temporarily eliminated the subscript $\mathbf{x}$ referring to the r.v. in question in $\boldsymbol{\mu}$ and $\mathbf{P}$.

Utilizing the vector notation, the joint characteristic function (1.41) can be written as:

$$\phi_{\mathbf{x}}(\mathbf{u}) = \mathcal{E}\{\ \exp(i\mathbf{u}^T\mathbf{x})\ \}\,. \tag{1.68}$$

In this way, the characteristic function of a normally distributed random vector can be calculated using the expression above and the transformation of variables $\boldsymbol{x} = \mathbf{P}^{1/2}\boldsymbol{y} + \boldsymbol{\mu}$, that is,

$$\begin{aligned} \phi_{\mathbf{x}}(\mathbf{u}) \ &\equiv\ \int_{-\infty}^{\infty} p_{\mathbf{x}}(\boldsymbol{x}) e^{i\mathbf{u}^T\boldsymbol{x}}\, d\boldsymbol{x} = \frac{1}{(2\pi)^{n/2}|\mathbf{P}|^{1/2}} \\ &\times \int_{-\infty}^{\infty} \exp(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}) \exp\left[i\mathbf{u}^T(\mathbf{P}^{1/2}\boldsymbol{y} + \boldsymbol{\mu})\right]\, ||\mathrm{Jac}[\boldsymbol{x}(\boldsymbol{y})]||\, d\boldsymbol{y} \end{aligned} \tag{1.69}$$

where $||\mathrm{Jac}[\boldsymbol{x}(\boldsymbol{y})]||$ is the absolute value of the determinant of the Jacobian matrix, defined as

$$\mathrm{Jac}[\boldsymbol{x}(\boldsymbol{y})] \ \equiv\ \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}^T} \ =\ \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix} \tag{1.70}$$

of the transformation. Using the fact that

$$|\mathrm{Jac}[\boldsymbol{x}(\boldsymbol{y})]| \ =\ |\mathbf{P}^{1/2}| \ =\ |\mathbf{P}|^{1/2} \tag{1.71}$$

we can write

$$\phi_{\mathbf{x}}(\mathbf{u}) \ =\ \exp(i\mathbf{u}^T\boldsymbol{\mu}) \frac{1}{(2\pi)^{n/2}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}(\boldsymbol{y}^T\boldsymbol{y} - 2i\mathbf{u}^T\mathbf{P}^{1/2}\boldsymbol{y})\right]\, d\boldsymbol{y} \tag{1.72}$$

Adding and subtracting $(1/2)\mathbf{u}^T\mathbf{P}\mathbf{u}$ to complete the square in the integrand above, we obtain:

$$\begin{aligned} \phi_{\mathbf{x}}(\mathbf{u}) \ &=\ \exp(i\mathbf{u}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{u}^T\mathbf{P}\mathbf{u}) \\ &\quad \times \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}(\boldsymbol{y}^T\boldsymbol{y} - 2i\mathbf{u}^T\mathbf{P}^{1/2}\boldsymbol{y} - \mathbf{u}^T\mathbf{P}\mathbf{u})\right]\, d\boldsymbol{y} \\ &=\ \exp(i\mathbf{u}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{u}^T\mathbf{P}\mathbf{u}) \\ &\quad \times \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}(\boldsymbol{y} - i\mathbf{P}^{1/2}\mathbf{u})^T(\boldsymbol{y} - i\mathbf{P}^{1/2}\mathbf{u})\right]\, d\boldsymbol{y} \end{aligned} \tag{1.73}$$

and making use of the integral

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}\right)\, d\boldsymbol{y} \ =\ \sqrt{(2\pi)^n} \tag{1.74}$$

we have that

$$\phi_{\mathbf{x}}(\mathbf{u}) = \exp(i\mathbf{u}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{u}^T\mathbf{P}\mathbf{u}), \tag{1.75}$$

is the characteristic function for a Gaussian distribution.

In the calculation of the integral above we defined the vector $\mathbf{y}$ as a function of the random vector $\mathbf{x}$ and transformed the integral in to a simpler integral. This gives an opportunity for us to mention a theorem relating functional transformation of random variable (vectors) and their respective probability distributions. Consider two n–dimensional random vectors $\mathbf{x}$ and $\mathbf{y}$ (not related to the characteristic function calculated above), that are related through a function $\mathbf{f}$ as $\mathbf{y} = \mathbf{f}(\mathbf{x})$, such that the inverse functional relation $\mathbf{x} = \mathbf{f}^{-1}(\mathbf{y})$ exists. In this case, the probability density $p_{\mathbf{y}}(\boldsymbol{y})$ of $\mathbf{y}$ can be obtained given the probability density $p_{\mathbf{x}}(\boldsymbol{x})$ of $\mathbf{x}$ by the transformation:

$$p_{\mathbf{y}}(\boldsymbol{y}) \;=\; p_{\mathbf{x}}[\mathbf{f}^{-1}(\boldsymbol{y})]\,||\mathrm{Jac}([\mathbf{f}^{-1}(\boldsymbol{y})]|| \tag{1.76}$$

where $||\mathrm{Jac}([\mathbf{f}^{-1}(\boldsymbol{y})]||$ is the absolute value of the determinant of the Jacobian of the inverse transformation of $\mathbf{x}$ in to $\mathbf{y}$. A proof of this theorem is given in Jazwinski [84], pp. 34–35.

1.3.3 Conditional Expectations

Motivated by the conditional probability concept presented in Section 1.1.2, we now introduce the concept of conditional probability density. If $\mathbf{x}$ and $\mathbf{y}$ are random vectors, the probability density that the event $\mathbf{x}$ occurs given that the event $\mathbf{y}$ occurred is defined as

$$p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) \equiv \frac{p_{\mathbf{xy}}(\boldsymbol{x},\boldsymbol{y})}{p_{\mathbf{y}}(\boldsymbol{y})}. \tag{1.77}$$

Analogously, reversing the meaning of $\mathbf{x}$ and $\mathbf{y}$,

$$p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}|\boldsymbol{x}) \equiv \frac{p_{\mathbf{xy}}(\boldsymbol{x},\boldsymbol{y})}{p_{\mathbf{x}}(\boldsymbol{x})}, \tag{1.78}$$

and Bayes rule for probability densities immediately follows:

$$p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}|\boldsymbol{x})p_{\mathbf{x}}(\boldsymbol{x})}{p_{\mathbf{y}}(\boldsymbol{y})}. \tag{1.79}$$

Based on the definition (1.77) we can define the conditional expectation (or mean) of an r.v. $\mathbf{x}$ given an r.v. $\mathbf{y}$ as:

$$\mathcal{E}\{\mathbf{x}|\mathbf{y}\} \equiv \int_{-\infty}^{\infty} \boldsymbol{x}\,p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y})\,d\boldsymbol{x}. \tag{1.80}$$

Now remember that the unconditional mean is given by

$$\mathcal{E}\{\mathbf{x}\} = \int_{-\infty}^{\infty} \boldsymbol{x}\,p_{\mathbf{x}}(\boldsymbol{x})\,d\boldsymbol{x}, \tag{1.81}$$

and that the marginal probability density $p_{\mathbf{x}}(\boldsymbol{x})$ can be obtained from the joint probability density $p_{\mathbf{xy}}(\boldsymbol{x},\boldsymbol{y})$ as,

$$p_{\mathbf{x}}(\boldsymbol{x}) = \int_{-\infty}^{\infty} p_{\mathbf{xy}}(\boldsymbol{x},\boldsymbol{y})\, d\boldsymbol{y}\,. \tag{1.82}$$

Considering the definition (1.77) we can write,

$$p_{\mathbf{x}}(\boldsymbol{x}) = \int_{-\infty}^{\infty} p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) p_{\mathbf{y}}(\boldsymbol{y})\, d\boldsymbol{y} \tag{1.83}$$

and substituting this result in (1.81) we have that

$$\begin{aligned} \mathcal{E}\{\mathrm{x}\} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \boldsymbol{x} p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) p_{\mathbf{y}}(\boldsymbol{y})\, d\boldsymbol{y} d\boldsymbol{x} \\ &= \int_{-\infty}^{\infty} \mathcal{E}\{\mathrm{x}|\boldsymbol{y}\} p_{\mathbf{y}}(\boldsymbol{y})\, d\boldsymbol{y} \\ &= \mathcal{E}\{\mathcal{E}\{\mathrm{x}|\mathrm{y}\}\}\,, \end{aligned} \tag{1.84}$$

where we used definition (1.80) of conditional expectation. The expression above is sometimes referred to as the *chain rule* for conditional expectations. Analogously we can obtain:

$$\mathcal{E}\{f(\mathrm{x},\mathrm{y})\} = \mathcal{E}\{\mathcal{E}\{f(\mathrm{x},\mathrm{y})|\mathrm{y}\}\}\,. \tag{1.85}$$

We can also define the conditional covariance matrix as

$$\mathbf{P}_{\mathbf{x}|\mathbf{y}} \equiv \mathcal{E}\{[\mathrm{x} - \mathcal{E}\{\mathrm{x}|\mathrm{y}\}][\mathrm{x} - \mathcal{E}\{\mathrm{x}|\mathrm{y}\}]^T|\mathrm{y}\}\,, \tag{1.86}$$

where we notice that $\mathbf{P}_{\mathbf{x}|\mathbf{y}}$ is a random matrix, contrary to what we encountered when we defined the unconditional covariance matrix (1.62).

We will now prove the following important result for normally distributed r.v.'s: the conditional probability of two normally distributed random vectors $\mathbf{x}$ and $\mathbf{y}$, with dimensions n and m respectively, is also normal and is given by:

$$p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}|\mathbf{P}_{\mathbf{x}|\mathbf{y}}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}})^T \mathbf{P}_{\mathbf{x}|\mathbf{y}}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}})\right]\,, \tag{1.87}$$

where

$$\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} = \boldsymbol{\mu}_{\mathbf{x}} + \mathbf{P}_{\mathbf{xy}}\mathbf{P}_{\mathbf{y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\mathbf{y}})\,, \tag{1.88}$$

and

$$\mathbf{P}_{\mathbf{x}|\mathbf{y}} = \mathbf{P}_{\mathbf{x}} - \mathbf{P}_{\mathbf{xy}}\mathbf{P}_{\mathbf{y}}^{-1}\mathbf{P}_{\mathbf{xy}}^T\,. \tag{1.89}$$

Now consider the following vector $\mathbf{z} = [\mathbf{x}^T\ \mathbf{y}^T]^T$ of dimension $(n+m)$. This vector has mean $\boldsymbol{\mu}_{\mathbf{z}}$ given by

$$\boldsymbol{\mu}_{\mathbf{z}} = \mathcal{E}\{\mathbf{z}\} = \begin{bmatrix} \mathcal{E}\{\mathrm{x}\} \\ \mathcal{E}\{\mathrm{y}\} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\mu}_{\mathbf{y}} \end{bmatrix} \tag{1.90}$$

and covariance $\mathbf{P}_{\mathbf{z}}$ that can be written as

$$\begin{aligned} \mathbf{P}_{\mathbf{z}} &= \mathcal{E}\{(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}})(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}})^T\} \\ &= \begin{bmatrix} \mathcal{E}\{(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T\} & \mathcal{E}\{(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})^T\} \\ \mathcal{E}\{(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T\} & \mathcal{E}\{(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})^T\} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{P}_{\mathbf{x}} & \mathbf{P}_{\mathbf{xy}} \\ \mathbf{P}_{\mathbf{xy}}^T & \mathbf{P}_{\mathbf{y}} \end{bmatrix}\,. \end{aligned} \tag{1.91}$$

Let us make use of the following equality (simple to verify):

$$\begin{bmatrix} \mathbf{I} & -\mathbf{P}_{\mathrm{xy}}\mathbf{P}_{\mathrm{y}}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{P}_{\mathrm{z}} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{P}_{\mathrm{y}}^{-1}\mathbf{P}_{\mathrm{xy}}^{T} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\mathbf{P}_{\mathrm{xy}}\mathbf{P}_{\mathrm{y}}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \times \begin{bmatrix} \mathbf{P}_{\mathrm{x}} - \mathbf{P}_{\mathrm{xy}}\mathbf{P}_{\mathrm{y}}^{-1}\mathbf{P}_{\mathrm{xy}}^{T} & \mathbf{P}_{\mathrm{xy}} \\ \mathbf{0} & \mathbf{P}_{\mathrm{y}} \end{bmatrix} = \begin{bmatrix} \mathbf{P}_{\mathrm{x|y}} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{\mathrm{y}} \end{bmatrix}, \tag{1.92}$$

where $\mathbf{P}_{\mathrm{x|y}}$ is defined as in (1.89), and we are assuming that $\mathbf{P}_{\mathrm{y}}^{-1}$ exists. From this expression, it follows that the determinant of the covariance matrix $\mathbf{P}_{\mathrm{z}}$ is

$$\begin{aligned} |\mathbf{P}_{\mathrm{z}}| &= |\mathbf{P}_{\mathrm{x|y}}|\,|\mathbf{P}_{\mathrm{y}}| \\ &= |\mathbf{P}_{\mathrm{x}} - \mathbf{P}_{\mathrm{xy}}\mathbf{P}_{\mathrm{y}}^{-1}\mathbf{P}_{\mathrm{xy}}^{T}|\,|\mathbf{P}_{\mathrm{y}}| \end{aligned} \tag{1.93}$$

(Householder [83], p. 17). Moreover, we have that

$$\begin{aligned} \mathbf{P}_{\mathrm{z}}^{-1} &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{P}_{\mathrm{y}}^{-1}\mathbf{P}_{\mathrm{xy}}^{T} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{P}_{\mathrm{x|y}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{\mathrm{y}}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{P}_{\mathrm{xy}}\mathbf{P}_{\mathrm{y}}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{P}_{\mathrm{x|y}}^{-1} & -\mathbf{P}_{\mathrm{x|y}}^{-1}\mathbf{P}_{\mathrm{xy}}\mathbf{P}_{\mathrm{y}}^{-1} \\ -\mathbf{P}_{\mathrm{y}}^{-1}\mathbf{P}_{\mathrm{xy}}^{T}\mathbf{P}_{\mathrm{x|y}}^{-1} & \mathbf{P}_{\mathrm{y}}^{-1}\mathbf{P}_{\mathrm{xy}}^{T}\mathbf{P}_{\mathrm{x|y}}^{-1}\mathbf{P}_{\mathrm{xy}}\mathbf{P}_{\mathrm{y}}^{-1} + \mathbf{P}_{\mathrm{y}}^{-1} \end{bmatrix} \end{aligned} \tag{1.94}$$

Therefore, multiplying $\mathbf{P}_{\mathrm{z}}^{-1}$ by $(\boldsymbol{z} - \boldsymbol{\mu}_{\mathrm{z}})^T$ on the left and by $(\boldsymbol{z} - \boldsymbol{\mu}_{\mathrm{z}})$ on the right, we have

$$\begin{aligned} (\boldsymbol{z} - \boldsymbol{\mu}_{\mathrm{z}})^T\mathbf{P}_{\mathrm{z}}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_{\mathrm{z}}) &= (\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x}})^T\mathbf{P}_{\mathrm{x|y}}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x}}) \\ &- (\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x}})^T\mathbf{P}_{\mathrm{x|y}}^{-1}\mathbf{P}_{\mathrm{xy}}\mathbf{P}_{\mathrm{y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}}) \\ &- (\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}})^T\mathbf{P}_{\mathrm{y}}^{-1}\mathbf{P}_{\mathrm{xy}}^{T}\mathbf{P}_{\mathrm{x|y}}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x}}) \\ &+ (\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}})^T\mathbf{P}_{\mathrm{y}}^{-1}\mathbf{P}_{\mathrm{xy}}^{T}\mathbf{P}_{\mathrm{x|y}}^{-1}\mathbf{P}_{\mathrm{xy}}\mathbf{P}_{\mathrm{y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}}) \\ &+ (\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}})^T\mathbf{P}_{\mathrm{y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}}) \end{aligned} \tag{1.95}$$

and using the definition (1.88) we can write

$$\begin{aligned} (\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x|y}})^T\mathbf{P}_{\mathrm{x|y}}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x|y}}) &= [(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x}}) - \mathbf{P}_{\mathrm{xy}}\mathbf{P}_{\mathrm{y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}})]^T\mathbf{P}_{\mathrm{x|y}}^{-1} \\ &\quad \times [(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x}}) - \mathbf{P}_{\mathrm{xy}}\mathbf{P}_{\mathrm{y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}})] \\ &= (\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x}})^T\mathbf{P}_{\mathrm{x|y}}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x}}) \\ &- (\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x}})^T\mathbf{P}_{\mathrm{x|y}}^{-1}\mathbf{P}_{\mathrm{xy}}\mathbf{P}_{\mathrm{y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}}) \\ &- (\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}})^T\mathbf{P}_{\mathrm{y}}^{-1}\mathbf{P}_{\mathrm{xy}}^{T}\mathbf{P}_{\mathrm{x|y}}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x}}) \\ &+ (\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}})^T\mathbf{P}_{\mathrm{y}}^{-1}\mathbf{P}_{\mathrm{xy}}^{T}\mathbf{P}_{\mathrm{x|y}}^{-1}\mathbf{P}_{\mathrm{xy}}\mathbf{P}_{\mathrm{y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}}) \end{aligned} \tag{1.96}$$

so that (1.95) reduces to

$$(\boldsymbol{z} - \boldsymbol{\mu}_{\mathrm{z}})^T\mathbf{P}_{\mathrm{z}}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_{\mathrm{z}}) = (\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x|y}})^T\mathbf{P}_{\mathrm{x|y}}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x|y}}) + (\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}})^T\mathbf{P}_{\mathrm{y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\mathrm{y}}) \tag{1.97}$$

By the definition of conditional probability we have

$$
\begin{aligned}
p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) &= \frac{p_{\mathbf{xy}}(\boldsymbol{x},\boldsymbol{y})}{p_{\mathbf{y}}(\boldsymbol{y})} = \frac{p_{\mathbf{z}}(\boldsymbol{z})}{p_{\mathbf{y}}(\boldsymbol{y})} \\
&= \frac{1}{(2\pi)^{n/2}} \frac{|\mathbf{P_y}|^{1/2}}{|\mathbf{P_z}|^{1/2}} \frac{\exp[-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_\mathbf{z})^T\mathbf{P}_\mathbf{z}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_\mathbf{z})]}{\exp[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_\mathbf{y})^T\mathbf{P}_\mathbf{y}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_\mathbf{y})]} \qquad (1.98)
\end{aligned}
$$

and utilizing (1.93) and (1.97) we obtain

$$
\begin{aligned}
p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) &= \frac{1}{(2\pi)^{n/2}|\mathbf{P_x}-\mathbf{P_{xy}}\mathbf{P_y}^{-1}\mathbf{P}_\mathbf{xy}^T|^{1/2}} \\
&\quad \times \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}})^T[\mathbf{P_x}-\mathbf{P_{xy}}\mathbf{P_y}^{-1}\mathbf{P}_\mathbf{xy}^T]^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}})]
\end{aligned} \qquad (1.99)
$$

which is the desired result. The assumption made above that the inverse of $\mathbf{P_y}$ exists is not necessary. When this inverse does not exist, it is possible to show (Kalman [89]) that the same result is still valid, but in place of the inverse of $\mathbf{P_y}$, we should utilize the pseudo–inverse $\mathbf{P}_\mathbf{y}^+$.

The calculation above involved construction of the joint probability distribution $p_\mathbf{z}(\boldsymbol{z})$ of the random vector $\mathbf{z} = [\mathbf{x}^T\mathbf{y}^T]^T$. Let us assume for the moment that the two random vectors $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated, that is, $\mathbf{P_{xy}} = \mathbf{0}$. Hence, referring back to (1.88) and (1.89) it follows that,

$$
\begin{aligned}
\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} &= \boldsymbol{\mu}_\mathbf{x} \qquad (1.100) \\
\mathbf{P}_{\mathbf{x}|\mathbf{y}} &= \mathbf{P_x} \qquad (1.101)
\end{aligned}
$$

which is intuitively in agreement with the notion of independence. Introducing these results in (1.97) we have

$$
(\boldsymbol{z}-\boldsymbol{\mu}_\mathbf{z})^T\mathbf{P}_\mathbf{z}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_\mathbf{z}) = (\boldsymbol{x}-\boldsymbol{\mu}_\mathbf{x})^T\mathbf{P}_\mathbf{x}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_\mathbf{x}) + (\boldsymbol{y}-\boldsymbol{\mu}_\mathbf{y})^T\mathbf{P}_\mathbf{y}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_\mathbf{y}) \qquad (1.102)
$$

Moreover, it follows from (1.93) that for uncorrelated random vectors $\mathbf{x}$ and $\mathbf{y}$,

$$
|\mathbf{P_z}| = |\mathbf{P_x}|\,|\mathbf{P_y}| \qquad (1.103)
$$

Thus, the joint probability density function $p_\mathbf{z}(\boldsymbol{z})$ can then be written as

$$
\begin{aligned}
p_\mathbf{z}(\boldsymbol{z}) &= \frac{1}{(2\pi)^{(n+m)/2}}\frac{1}{|\mathbf{P_z}|^{1/2}}\exp[-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_\mathbf{z})^T\mathbf{P}_\mathbf{z}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_\mathbf{z})] \\
&= \frac{1}{(2\pi)^{(n+m)/2}}\frac{1}{|\mathbf{P_x}|^{1/2}|\mathbf{P_y}|^{1/2}} \\
&\quad \times\exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_\mathbf{x})^T\mathbf{P}_\mathbf{x}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_\mathbf{x}) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_\mathbf{y})^T\mathbf{P}_\mathbf{y}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_\mathbf{y})] \\
&= \frac{1}{(2\pi)^{n/2}}\frac{1}{|\mathbf{P_x}|^{1/2}}\exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_\mathbf{x})^T\mathbf{P}_\mathbf{x}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_\mathbf{x})] \\
&\quad \times\frac{1}{(2\pi)^{m/2}}\frac{1}{|\mathbf{P_y}|^{1/2}}\exp[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_\mathbf{y})^T\mathbf{P}_\mathbf{y}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_\mathbf{y})] \\
&= p_\mathbf{x}(\boldsymbol{x})p_\mathbf{y}(\boldsymbol{y}) \qquad (1.104)
\end{aligned}
$$

This shows that two normally distributed random vectors that are uncorrelated are also independent. We have seen earlier in this section that independence among random variables implied they are uncorrelated; the contrary was not necessarily true. However, as we have just shown, the contrary is true in the case of normally distributed random variables (vectors).

Exercises

1. Using the definition of Rayleigh probability density function given in (1.12): (a) calculate the mean and standard deviation for a r.v. with that distribution; (b) find the mode of the r.v., that is, is most likely value.

2. (Brown [19], Problem 1.40) A pair of random variables, x e y, have the following joint probability density function:

$$p_{\mathrm{xy}}(x, y) = \begin{cases} 1 & 0 \le y \le 2x \text{ e } 0 \le x \le 1 \\ 0 & \text{em everywhere else} \end{cases}$$

   Find $\mathcal{E}\{\mathrm{x}|\mathrm{y} = .5\}$. [Hint: Use (1.77) to find $p_{\mathrm{x}|\mathrm{y}}(x)$ for $y = 0.5$, and then integrate $x p_{\mathrm{x}|\mathrm{y}}(x)$ to find $\mathcal{E}\{\mathrm{x}|\mathrm{y} = .5\}$.]

3. Consider a zero–mean Gaussian random vector, with probability density and characteristic functions

$$f_{\mathbf{x}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{P}|^{1/2}} \exp\left[-\frac{1}{2}\boldsymbol{x}^T\mathbf{P}^{-1}\boldsymbol{x}\right],$$

$$\phi_{\mathbf{x}}(\mathbf{u}) = \exp\left[-\frac{1}{2}\mathbf{u}^T\mathbf{P}\mathbf{u}\right],$$

   respectively. Show that the following holds for the first four moments of this distribution:

$$\begin{array}{ll} \mathcal{E}\{\mathrm{x}_k\} = 0 & \mathcal{E}\{\mathrm{x}_k\mathrm{x}_l\} = P_{kl} \\ \mathcal{E}\{\mathrm{x}_k\mathrm{x}_l\mathrm{x}_m\} = 0 & \mathcal{E}\{\mathrm{x}_k\mathrm{x}_l\mathrm{x}_m\mathrm{x}_n\} = P_{kl}P_{mn} + P_{km}P_{ln} + P_{kn}P_{lm} \end{array}$$

   where x_i, $i \in \{k, l, m, n\}$, are elements of the random vector x, and P_{ij}, $i, j \in \{k, l, m, n\}$, are elements of $\mathbf{P}$.

4. Show that the linear transformation of a normally distributed vector is also normally distributed. That is, show that for a given normally distributed vector $\mathbf{x}$, with mean $\mu_{\mathbf{x}}$ and covariance $\mathbf{R}_{\mathbf{x}}$, the linear transformation

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$$

   produces a normally distributed vector $\mathbf{y}$ with mean $\mu_{\mathbf{y}} = \mathbf{A}\mu_{\mathbf{x}} + \mathbf{b}$ and covariance $\mathbf{R}_{\mathbf{y}} = \mathbf{A}\mathbf{R}_{\mathbf{x}}\mathbf{A}^T$.

5. The log–normal distribution is defined by

$$p_{\mathrm{x}}(x) = \frac{1}{\sqrt{2\pi}s}\frac{1}{x}\exp\left[-\frac{1}{2s^2}\left(\ln\frac{x}{x_0}\right)^2\right]$$

   Show that:

(a) its mean and variance are

$$\mu = x_0 e^{s^2/2}$$
$$var(x) = x_0^2 e^{s^2}(e^{s^2} - 1)$$

respectively;

(b) introducing the variable

$$x^* = \beta \ln(\frac{x}{\gamma})$$

the probability density function $p_X(x)$ above can be converted to a Gaussian probability density function $p_X^*(x)$ of the form

$$p_X^*(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x^* - x_0^*)^2}{2\sigma^2}\right]$$

where $\sigma = s\beta$, and

$$x_0^* = \beta \ln(\frac{x_0}{\gamma})$$

This justifies the name log–normal distribution for $p_X(x)$, since the logarithm of its variable is normally distributed.

6. According to what we have seen in the previous exercise, let $\mathbf{v}$ an n–dimensional normally distributed r.v., defined as $\mathbf{v} \sim \mathcal{N}(\mu_\mathbf{v}, \mathbf{P})$. The vector with components w_j defined as $w_j = \exp(v_j)$ for $j = 1, \cdots, n$ is said to be distributed log–normally and it is represented by $\mathbf{w} \sim \mathcal{LN}(\mu_\mathbf{w}, \mathbf{R})$ where $\mu_\mathbf{w}$ is its mean and $\mathbf{R}$ its covariance. Show that

$$\mathcal{E}\{\mathrm{w}_j\} = \exp\left[\mathcal{E}\{\mathrm{v}_j\} + \frac{1}{2}P_{jj}\right] ,$$

and

$$R_{jk} = \mathcal{E}\{\mathrm{w}_j\}\mathcal{E}\{\mathrm{w}_k\}(e^{P_{jk}} - 1) .$$

(Hint: Utilize the concept of characteristic function.)

7. *Computer Assignment:* (Based on Tarantola [126]) Consider the experiment of measuring (estimating) the value of a constant quantity corrupted by "noise". To simulate this situation, let us use Matlab, to generate 101 measurements of the random variable y as follows:

Enter: `y = 21 + rand(101,1)`; the intrinsic Matlab function *rand* generates a uniformly distributed r.v. in the interval $(0, 1)$

Enter: `x=20:0.1:23`; to generate an array with 31 points in the neighborhood of 21

Enter: `hist(y,x)`; this will show you a histogram corresponding to this experiment

Now using the Matlab functions *median, mean, max* and *min*, calculate the median, mean and mid–range values for the experiment you have just performed. What did you get? Three distinct values! Can you tell which of these are the closest value to the true value?

8. *Computer Assignment:* Ok, you probably still can't answer the question above. So here is the real assignment:

(a) Construct a Matlab function that repeats the experiment of the previous exercise 20 times, for the given value of the scalar under noise. For each successive experiment, increase the number of samples used by 100, calculating and storing the values their corresponding median, mean, and mid–range. At the end of the 20 experiments, plot the values obtained for the median, mean and mid–range in each experiment. Can you guess now which one of these is the best estimate?

(b) To really confirm your guess, fix the number of samples at 100 and repeat the experiments 200 times, collecting the corresponding median, mean and mid–range values for each experiment. (It is a good idea, if you do it as another Matlab function.) In end of all 200 experiments, plot the histograms for each of these three quantities. Which one has the least scatter? Is this compatible with your guess from of the previous item?

(c) Repeat items (a) and (b) for the same constant, but now being disturbed by a normally distributed random variable with mean zero and unity variance. That is, replace the Matlab function *rand* by the function *randn*. *Caution: when construction the histogram in this item chose a relatively large interval for the outcome counting, e.g.,* `x=18:0.1:24`. *This is necessary because the Gaussian function has very long tails.*