

# Lecture 1: Fundamentals of probability



School of Mathematical Sciences, Xiamen University

## 1. Elementary probabilities

- The *probability space*:  $(\Omega, \mathcal{B}, \mathbb{P})$ . Here,  $\Omega$  is an abstract set called the *sample space*. The set  $\mathcal{B}$  (a  $\sigma$ -algebra) is a collection of subsets of  $\Omega$ , satisfying the following conditions:

- (i)  $\Omega \in \mathcal{B}$ , and if  $A \in \mathcal{B}$ , then  $\Omega \setminus A \in \mathcal{B}$ ,
- (ii) if  $A_1, A_2, \dots \in \mathcal{B}$ , then  $\bigcup_{j=1}^{\infty} A_j \in \mathcal{B}$ .

We call the set  $\mathcal{B}$  the *event space*, and the individual sets in it are referred to as *events*. The *probability measure*  $\mathbb{P}$  is a mapping

$$\mathbb{P} : \mathcal{B} \rightarrow \mathbb{R}, \quad \mathbb{P}(E) = \text{probability of } E, \quad E \in \mathcal{B},$$

that must satisfy the following conditions:

- (i)  $\mathbb{P}(\Omega) = 1$ , and for all  $E \in \mathcal{B}$ ,  $0 \leq \mathbb{P}(E) \leq 1$ ,
- (ii) if  $A_j \in \mathcal{B}$ ,  $j = 1, 2, \dots$  with  $A_j \cap A_k = \emptyset$  whenever  $j \neq k$ , then

$$\mathbb{P}\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mathbb{P}(A_j).$$

- It follows from the definition that

$$\mathbb{P}(\Omega \setminus A) = 1 - \mathbb{P}(A),$$

which implies that  $\mathbb{P}(\emptyset) = 0$ . Moreover, if  $A_1, A_2 \in \mathcal{B}$  and  $A_1 \subset A_2 \subset \Omega$ , then

$$\mathbb{P}(A_1) \leq \mathbb{P}(A_2).$$

- Two events,  $A$  and  $B$ , are *independent*, if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

- The *conditional probability* of  $A$  given  $B$  is the probability that  $A$  happens *provided* that  $B$  happens,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \text{assuming that } \mathbb{P}(B) > 0.$$

**Exercise:** Prove that  $\mathbb{P}(A \mid B) \leq 1$ .

- It follows from the definition of independent events that, if  $A$  and  $B$  are mutually independent, then

$$\mathbb{P}(A \mid B) = \mathbb{P}(A), \quad \mathbb{P}(B \mid A) = \mathbb{P}(B).$$

Vice versa, if one of the above equalities holds, then by the definition of conditional probabilities  $A$  and  $B$  must be independent.

- *Bayes' formula* for elementary events.

Assume that  $\mathbb{P}(A) > 0$  and  $\mathbb{P}(B) > 0$ . From

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad \text{and} \quad \mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)},$$

we obtain

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}.$$

## 2. Probability distributions and densities

- Given a sample space  $\Omega$ , a real valued random variable  $X$  is a mapping

$$X : \Omega \rightarrow \mathbb{R},$$

which assigns to each element of  $\Omega$  a real value  $X(\omega)$ , such that for every open set  $A \subset \mathbb{R}$ ,  $X^{-1}(A) \in \mathcal{B}$ . ( $X$  is a measurable function.)

We call  $x = X(\omega)$ ,  $\omega \in \Omega$ , a *realization* of  $X$ .

- For each  $A \subset \mathbb{R}$ , we define

$$\mu_X(A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}\{\omega \in \Omega : X(\omega) \in A\},$$

and call  $\mu_X$  the *probability distribution* of  $X$ , i.e.,  $\mu_X(A)$  is the probability of the event  $\{\omega \in \Omega : X(\omega) \in A\}$ . The probability distribution  $\mu_X(A)$  measures the size of the subset of  $\Omega$  mapped onto  $A$  by the random variable  $X$ .

- We only consider simple cases, meaning that there exists a function, the *probability density*  $\pi_X$  of  $X$ , such that

$$\mu_X(A) = \int_A \pi_X(x) dx.$$

A function is a probability density if it satisfies the following two conditions:

$$\pi_X(x) \geq 0,$$

$$\int_{\mathbb{R}} \pi_X(x) dx = 1.$$

Any function satisfying the above conditions can be viewed as a probability density of some random variable.

- The *cumulative distribution function* (cdf) of a real-valued random variable is defined as

$$\Phi_X(x) = \int_{-\infty}^x \pi_X(x') dx' = \mathbb{P}\{X \leq x\}.$$

Observe that  $\Phi_X(x)$  is non-decreasing, and it satisfies

$$\lim_{x \rightarrow -\infty} \Phi_X(x) = 0, \quad \lim_{x \rightarrow \infty} \Phi_X(x) = 1.$$

- The definition of random variables can be generalized to cover multidimensional state spaces. Given two real-valued random variables  $X$  and  $Y$ , the joint probability distribution defined over Cartesian products of sets is

$$\mu_{XY}(A \times B) = \mathbb{P}(X^{-1}(A) \cap Y^{-1}(B)) = \mathbb{P}\{X \in A, Y \in B\},$$

the probability of the event that  $X \in A$  and, at the same time,  $Y \in B$ , where  $A, B \subset \mathbb{R}$ .

- Assuming that the probability distribution can be written as an integral of the form

$$\mu_{XY}(A \times B) = \iint_{A \times B} \pi_{XY}(x, y) dx dy,$$

the non-negative function  $\pi_{XY}$  defines the *joint probability density* of the random variables  $X$  and  $Y$ . We may define a two-dimensional random variable,

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix} : \Omega \rightarrow \mathbb{R}^2,$$

and by approximating general two-dimensional sets by unions of rectangles, we may write

$$\mathbb{P}\{Z \in B \subset \mathbb{R}^2\} = \iint_B \pi_{XY}(x, y) dx dy = \int_B \pi_Z(z) dz,$$

where we used the notation  $\pi_{XY}(x, y) = \pi_Z(z)$ , and the integral with respect to  $z$  is the two-dimensional integral,  $dz = dx dy$ .



- More generally, we define a multivariate random variable as a measurable mapping

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} : \Omega \rightarrow \mathbb{R}^n,$$

where each component  $X_i$  is a real-valued random variable. The probability density of  $X$  is the joint probability density

$$\pi_X = \pi_{X_1 X_2 \dots X_n} : \mathbb{R}^n \rightarrow \mathbb{R}_+$$

of its components, satisfying

$$\mathbb{P}\{X \in B\} = \mu_X(B) = \int_B \pi_X(x) dx, \quad B \subset \mathbb{R}^n.$$

- The joint probability density  $\pi_{XY}$  of two multivariate random variables  $X : \Omega \rightarrow \mathbb{R}^n$  and  $Y : \Omega \rightarrow \mathbb{R}^m$  can be defined in the space  $\mathbb{R}^{n+m}$  analogously.

- The random variables  $X \in \mathbb{R}^n$  and  $Y \in \mathbb{R}^m$  are *independent* if

$$\pi_{XY}(x, y) = \pi_X(x)\pi_Y(y),$$

in agreement with the definition of independent events. This formula gives us also a way to calculate the joint probability density of two independent random variables.

- Given two not necessarily independent random variables  $X \in \mathbb{R}^n$  and  $Y \in \mathbb{R}^m$  with joint probability density  $\pi_{XY}(x, y)$ , the *marginal density* of  $X$  is the probability of  $X$  when  $Y$  may take on any value,

$$\pi_X(x) = \int_{\mathbb{R}^m} \pi_{XY}(x, y) dy.$$

In other words, the marginal density of  $X$  is simply the probability density of  $X$  without any thoughts about  $Y$ . The marginal of  $Y$  is defined analogously by the formula

$$\pi_Y(y) = \int_{\mathbb{R}^n} \pi_{XY}(x, y) dx.$$

- Consider the last formula, and assume that  $\pi_Y(y) \neq 0$ . Dividing both sides by the scalar  $\pi_Y(y)$  gives the identity

$$\int_{\mathbb{R}^n} \frac{\pi_{XY}(x, y)}{\pi_Y(y)} dx = 1.$$

Since the integrand is a non-negative function, it defines a probability density for  $X$ , for fixed  $y$ . We define the *conditional probability density* of  $X$  given  $Y$ ,

$$\pi_{X|Y}(x | y) = \frac{\pi_{XY}(x, y)}{\pi_Y(y)}, \quad \pi_Y(y) \neq 0.$$

With some caution, and in a rather cavalier way, one can interpret  $\pi_{X|Y}$  as the probability density of  $X$ , assuming that the random variable  $Y$  takes on the value  $Y = y$ .

- The conditional density of  $Y$  given  $X$  is defined similarly as

$$\pi_{Y|X}(y | x) = \frac{\pi_{XY}(x, y)}{\pi_X(x)}, \quad \pi_X(x) \neq 0.$$

Observe that the symmetric roles of  $X$  and  $Y$  imply that

$$\pi_{XY}(x, y) = \pi_{X|Y}(x | y)\pi_Y(y) = \pi_{Y|X}(y | x)\pi_X(x),$$

leading to the important identity known as *Bayes' formula* for probability densities,

$$\pi_{X|Y}(x | y) = \frac{\pi_{Y|X}(y | x)\pi_X(x)}{\pi_Y(y)}.$$

### 3. Change of variables in probability densities

- Assume that we have two real-valued random variables  $X, Z$  that are related to each other through a functional relation

$$X = \phi(Z),$$

where  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  is a one-to-one mapping. For simplicity, assume that  $\phi$  is strictly increasing and differentiable, so that  $\phi'(z) > 0$ . If the probability density function  $\pi_X$  of  $X$  is given, what is the corresponding density  $\pi_Z$  of  $Z$ ?

First, note that since  $\phi$  is increasing, for any values  $a < b$ , we have

$$a < Z < b \text{ if and only if } a' = \phi(a) < \phi(Z) = X < \phi(b) = b',$$

therefore

$$\mathbb{P}\{a' < X < b'\} = \mathbb{P}\{a < Z < b\}.$$

Equivalently, the probability density of  $Z$  satisfies

$$\int_a^b \pi_Z(z) dz = \int_{a'}^{b'} \pi_X(x) dx.$$

Performing a change of variables in the integral on the right,

$$x = \phi(z), \quad dx = \frac{d\phi}{dz}(z) dz,$$

we obtain

$$\int_a^b \pi_Z(z) dz = \int_a^b \pi_X(\phi(z)) \frac{d\phi}{dz}(z) dz.$$

This holds for all  $a$  and  $b$ , and therefore we arrive at the conclusion that

$$\pi_Z(z) = \pi_X(\phi(z)) \frac{d\phi}{dz}(z).$$

- In the derivation above, we assumed that  $\phi$  was increasing. If it is decreasing, the derivative is negative. In general, since the density needs to be non-negative, we write

$$\pi_Z(z) = \pi_X(\phi(z)) \left| \frac{d\phi}{dz}(z) \right|.$$

- The above reasoning for one-dimensional random variables can be extended to multivariate random variables as follows. Let  $X \in \mathbb{R}^n$  and  $Z \in \mathbb{R}^n$  be two random variables such that

$$X = \phi(Z),$$

where  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is a one-to-one differentiable mapping.

Consider a set  $B \subset \mathbb{R}^n$ , and let  $B' = \phi(B) \subset \mathbb{R}^n$  be its image in the mapping  $\phi$ . Then we may write

$$\int_B \pi_Z(z) dz = \int_{B'} \pi_X(x) dx.$$

- We perform the change of variables  $x = \phi(z)$  in the latter integral, remembering that

$$dx = |\det(D\phi(z))|dz,$$

where  $D\phi(z)$  is the Jacobian of the mapping  $\phi$ ,

$$D\phi(z) = \begin{bmatrix} \frac{\partial \phi_1}{\partial z_1} & \cdots & \frac{\partial \phi_1}{\partial z_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \phi_n}{\partial z_1} & \cdots & \frac{\partial \phi_n}{\partial z_n} \end{bmatrix} \in \mathbb{R}^{n \times n},$$

and its determinant, the Jacobian determinant, expresses the local volume scaling of the mapping  $\phi$ . Occasionally, the Jacobian determinant is written in a suggestive form to make it formally similar to the one-dimensional equivalent,

$$\frac{\partial \phi}{\partial z} = \det(D\phi(z)).$$



- With this notation,

$$\int_B \pi_Z(z) dz = \int_{B'} \pi_X(x) dx = \int_B \pi_X(\phi(z)) \left| \frac{\partial \phi}{\partial z} \right| dz$$

for all  $B \subset \mathbb{R}^n$ , and we arrive at the conclusion that

$$\pi_Z(z) = \pi_X(\phi(z)) \left| \frac{\partial \phi}{\partial z} \right|.$$

This is the change of variables formula for probability densities.

## 4. Expectation

- Given a random variable  $X \in \mathbb{R}$  with probability density  $\pi_X$ , its *expected value*, or *mean*, is defined as

$$\mathbb{E}(X) = \bar{x} = \int_{\mathbb{R}} x\pi_X(x)dx \in \mathbb{R}.$$

- Given a random variable  $X \in \mathbb{R}^n$  with probability density  $\pi_X$ , and a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , we define the *expectation* of  $f(X)$  as

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

**Exercise:** If two random variables  $X$  and  $Y$  are independent then

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

- Linearity:** for any random variables  $X$  and  $Y$ , and any  $\lambda \in \mathbb{R}$ ,

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y), \quad \mathbb{E}(\lambda X) = \lambda\mathbb{E}(X).$$

- Given a random variable  $X \in \mathbb{R}^n$  with probability density  $\pi_X$ , the mean of  $X$  is the vector in  $\mathbb{R}^n$ ,

$$\bar{x} = \int_{\mathbb{R}^n} x \pi_X(x) dx = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_n \end{bmatrix} \in \mathbb{R}^n,$$

or, component-wise,

$$\bar{x}_j = \int_{\mathbb{R}^n} x_j \pi_X(x) dx \in \mathbb{R}, \quad 1 \leq j \leq n.$$

**Exercise:** Prove that the  $j$ th component of the expectation of a multivariate random variable  $X \in \mathbb{R}^n$  can be calculated by using the corresponding marginal density. That is to say,

$$\bar{x}_j = \int_{\mathbb{R}} x_j \pi_{X_j}(x_j) dx_j = \mathbb{E}(X_j), \quad 1 \leq j \leq n.$$

## 4.1 Markov's inequality

- Let  $X$  be a non-negative random variable. For any  $\alpha > 0$ ,

$$\mathbb{P}\{X \geq \alpha\} \leq \frac{\mathbb{E}(X)}{\alpha}.$$

*Proof.* For any  $\alpha > 0$ , define the following function

$$f(X) = \begin{cases} 1, & \text{if } X \geq \alpha, \\ 0, & \text{otherwise.} \end{cases}$$

Then  $f(X) \leq X/\alpha$ , which yields  $\mathbb{E}(f(X)) \leq \mathbb{E}(X)/\alpha$ . It follows from

$$\mathbb{E}(f(X)) = 1 \cdot \mathbb{P}\{X \geq \alpha\} + 0 \cdot \mathbb{P}\{X < \alpha\} = \mathbb{P}\{X \geq \alpha\}$$

that

$$\mathbb{P}\{X \geq \alpha\} \leq \frac{\mathbb{E}(X)}{\alpha}. \quad \square$$

## 4.2 Conditional expectation

- Given two random variables  $X \in \mathbb{R}^n$  and  $Y \in \mathbb{R}^m$ , we define

$$\mathbb{E}(X \mid y) = \int_{\mathbb{R}^n} x \pi_{X|Y}(x \mid y) dx.$$

- Compute the expectation of  $X$  via its conditional expectation:

$$\begin{aligned}\mathbb{E}(X) &= \int_{\mathbb{R}^n} x \pi_X(x) dx = \int_{\mathbb{R}^n} x \left( \int_{\mathbb{R}^m} \pi_{XY}(x, y) dy \right) dx \\ &= \int_{\mathbb{R}^n} x \left( \int_{\mathbb{R}^m} \pi_{X|Y}(x \mid y) \pi_Y(y) dy \right) dx \\ &= \int_{\mathbb{R}^m} \left( \int_{\mathbb{R}^n} x \pi_{X|Y}(x \mid y) dx \right) \pi_Y(y) dy \\ &= \int_{\mathbb{R}^m} \mathbb{E}(X \mid y) \pi_Y(y) dy.\end{aligned}$$

This is the law of total expectation:  $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \mid Y))$ .

## 5. Variance and covariance

- The *variance* of the random variable  $X$  is the expectation of the squared deviation from the expectation,

$$\text{Var}(X) = \mathbb{E}((X - \bar{x})^2) = \sigma_X^2 = \int_{\mathbb{R}} (x - \bar{x})^2 \pi_X(x) dx.$$

The square root  $\sigma_X$  of the variance is the *standard deviation* of  $X$ . Obviously, it holds  $\text{Var}(X) = \mathbb{E}(X^2) - \bar{x}^2 \leq \mathbb{E}(X^2)$ .

- The  $k$ th moment of a probability density function is defined as

$$\mathbb{E}((X - \bar{x})^k) = \int_{\mathbb{R}} (x - \bar{x})^k \pi_X(x) dx.$$

The *skewness* and the *kurtosis* of the probability density are

$$\text{skew}(X) = \frac{\mathbb{E}((X - \bar{x})^3)}{\sigma_X^3}, \quad \text{kurt}(X) = \frac{\mathbb{E}((X - \bar{x})^4)}{\sigma_X^4}.$$

- The covariance of two random variables  $X$  and  $Y$  is defined as

$$\mathbb{C}\text{ov}(X, Y) = \mathbb{E}((X - \bar{x})(Y - \bar{y})).$$

$X$  and  $Y$  are said to be uncorrelated if  $\mathbb{C}\text{ov}(X, Y) = 0$ .

- If the random variables  $X$  and  $Y$  are independent, then

$$\mathbb{C}\text{ov}(X, Y) = 0 \quad \text{and} \quad \mathbb{V}\text{ar}(X + Y) = \mathbb{V}\text{ar}(X) + \mathbb{V}\text{ar}(Y).$$

Also, for any real  $\lambda$ , it holds  $\mathbb{V}\text{ar}(\lambda X) = \lambda^2 \mathbb{V}\text{ar}(X)$ .

- Given a random variable  $X \in \mathbb{R}^n$  with probability density  $\pi_X$ , the *covariance* of  $X$  is an  $n \times n$  matrix with elements

$$\mathbb{C}\text{ov}(X, X)_{ij} = \int_{\mathbb{R}^n} (x_i - \bar{x}_i)(x_j - \bar{x}_j) \pi_X(x) dx \in \mathbb{R}, \quad 1 \leq i, j \leq n.$$

Alternatively, we can define the covariance using vector notation as

$$\mathbb{C}\text{ov}(X, X) = \int_{\mathbb{R}^n} (x - \bar{x})(x - \bar{x})^\top \pi_X(x) dx \in \mathbb{R}^{n \times n}.$$

- The variance of the  $j$ th component  $X_j$  of  $X$  is

$$\text{Var}(X_j) = \int_{\mathbb{R}} (x_j - \bar{x}_j)^2 \pi_{X_j}(x_j) dx_j.$$

The  $j$ th diagonal entry of  $\text{Cov}(X, X)$  is

$$\text{Cov}(X, X)_{jj} = \int_{\mathbb{R}^n} (x_j - \bar{x}_j)^2 \pi_X(x) dx.$$

**Exercise:** Prove that

$$\text{Var}(X_j) = \text{Cov}(X, X)_{jj}, \quad 1 \leq j \leq n.$$

- We also use the notation  $\text{Var}(X)$  to denote  $\text{Cov}(X, X)$ .

**Exercise:** Prove that

$$\text{Var}(X) = \mathbb{E}((X - \bar{x})(X - \bar{x})^\top) = \mathbb{E}(XX^\top) - \bar{x}\bar{x}^\top.$$



**Exercise:** Given a nonzero vector  $v \in \mathbb{R}^n$  and a random variable  $X \in \mathbb{R}^n$ , define the real-valued random variable

$$X_v = v^\top X = \sum_{i=1}^n v_i X_i.$$

Compute the mean and variance of  $X_v$ .

- The covariance of a random variable  $X \in \mathbb{R}^n$  and a random variable  $Y \in \mathbb{R}^m$  is the  $n \times m$  matrix,

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbb{E}((X - \bar{x})(Y - \bar{y})^\top) \\ &= \mathbb{E}(XY^\top) - \bar{x}\bar{y}^\top,\end{aligned}$$

where  $\bar{x}$  and  $\bar{y}$  are the means of  $X$  and  $Y$  respectively.

**Exercise:** Prove that

$$\text{Cov}(X, Y) = (\text{Cov}(Y, X))^\top.$$

## 6. Other properties of expectation, variance, and covariance

- $\mathbb{E}$  is order preserving:

$$\mathbb{E}(X) \leq \mathbb{E}(Y), \quad \text{if } X \leq Y.$$

- Cauchy–Schwarz inequality:

If  $X$  and  $Y$  have finite variances, then  $|\mathbb{E}(XY)| < \infty$  and

$$|\mathbb{E}(XY)| \leq \mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

In particular,

$$|\text{Cov}(X, Y)| \leq \sigma_X \sigma_Y.$$

More generally,

$$|\mathbb{E}(X^\top Y)| \leq \mathbb{E}(|X^\top Y|) \leq \sqrt{\mathbb{E}(\|X\|^2)\mathbb{E}(\|Y\|^2)}.$$

- Jensen's inequality: If  $\psi$  is a convex function, then

$$\psi(\mathbb{E}(X)) \leq \mathbb{E}(\psi(X)).$$

In particular,  $\|\mathbb{E}(X)\| \leq \mathbb{E}(\|X\|)$ .

- Chebyshev's inequality: For any  $\alpha > 0$ ,

$$\mathbb{P}\{|X - \mathbb{E}(X)| \geq \alpha\} \leq \frac{\text{Var}(X)}{\alpha^2}.$$

- Cov is bilinear and shift invariant:

For any constants  $a$  and  $b$  and any  $c$ ,

$$\text{Cov}(aX + bY + c, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z),$$

$$\text{Cov}(Z, aX + bY + c) = a\text{Cov}(Z, X) + b\text{Cov}(Z, Y).$$

In particular,

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm (\text{Cov}(X, Y) + \text{Cov}(Y, X)).$$

- Covariance transformation:

For any matrices  $\mathbf{A}$  and  $\mathbf{B}$  (of appropriate sizes),

$$\mathbb{C}\text{ov}(\mathbf{A}X, \mathbf{B}Y) = \mathbf{A}\mathbb{C}\text{ov}(X, Y)\mathbf{B}^\top.$$

In particular,

$$\mathbb{V}\text{ar}(aX) = a^2\mathbb{V}\text{ar}(X), \quad \mathbb{V}\text{ar}(\mathbf{A}X) = \mathbf{A}\mathbb{V}\text{ar}(X)\mathbf{A}^\top.$$

- Expectation of a quadratic form:

If  $\mathbb{E}(X) = \bar{x}$ , then

$$\mathbb{E}(X^\top \mathbf{A}X) = \bar{x}^\top \mathbf{A}\bar{x} + \text{tr}(\mathbf{A}\mathbb{V}\text{ar}(X)),$$

where  $\text{tr}$  denotes the trace of the matrix.

## 7. Normal distributions

- A random variable  $X \in \mathbb{R}$  is *normally distributed*, or Gaussian, indicated symbolically by

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

if its cumulative distribution is given by

$$\mathbf{P}\{X \leq t\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^t \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) dx.$$

Hence, the Gaussian probability density is

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

We have

$$\mathbb{E}(X) = \mu, \quad \mathbb{V}\text{ar}(X) = \sigma^2.$$

- Gaussian multivariate random variable  $X \in \mathbb{R}^n$ :

$$X \sim \mathcal{N}(\mu, \mathbf{C}),$$

where  $\mu \in \mathbb{R}^n$  and  $\mathbf{C}$  is a symmetric positive definite matrix.  
The probability density is

$$\begin{aligned}\pi_X(x) &= \mathcal{N}(x \mid \mu, \mathbf{C}) \\ &= \left( \frac{1}{(2\pi)^n \det(\mathbf{C})} \right)^{1/2} \exp \left( -\frac{1}{2} (x - \mu)^\top \mathbf{C}^{-1} (x - \mu) \right).\end{aligned}$$

We have

$$\mathbb{E}(X) = \mu, \quad \text{Var}(X) = \mathbf{C}.$$

**Exercise:** Assume  $X \sim \mathcal{N}(\mu, \mathbf{C})$ . Prove that the  $n$  components  $X_j$ ,  $1 \leq j \leq n$ , of  $X$  are mutually independent Gaussian random variables if and only if  $\mathbf{C}$  is a diagonal matrix with positive diagonal entries.

- Affine transformations preserve multivariate Gaussianity:  
If  $X \in \mathbb{R}^n$  with  $X \sim \mathcal{N}(\mu, \mathbf{C})$ ,  $\mathbf{A} \in \mathbb{R}^{m \times n}$  and  $\mathbf{a} \in \mathbb{R}^m$ , then

$$Y = \mathbf{A}X + \mathbf{a} \sim \mathcal{N}(\mathbf{A}\mu + \mathbf{a}, \mathbf{A}\mathbf{C}\mathbf{A}^\top).$$

- Random variables that are jointly Gaussian and uncorrelated are also independent.
- Definition:  $X \sim \mathcal{N}(0, \mathbf{I}_n)$  is called a *standard normal n-variate random variable* (also referred to as *Gaussian white noise*).

**Exercise:** Assume  $X \sim \mathcal{N}(\mu, \mathbf{C})$  and  $\mathbf{C} = \mathbf{R}^\top \mathbf{R}$  is a Cholesky factorization. Prove that the random variable

$$Z = \mathbf{R}^{-\top}(X - \mu)$$

is a standard normal random variable. The above formula defines a *whitening transformation*, or *Mahalanobis transformation*, of the random variable  $X$  into Gaussian white noise.

## 7.1 Conditional distributions of the Gaussian

- Let  $X \sim \mathcal{N}(\mu, \mathbf{C})$ . Any vector  $[X_{k_1} \cdots X_{k_\ell}]^\top$  made of different components of  $X$  is Gaussian.
- Let

$$X = \begin{bmatrix} U \\ V \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_U \\ \mu_V \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{C}_U & \mathbf{C}_{UV} \\ \mathbf{C}_{VU} & \mathbf{C}_V \end{bmatrix}.$$

We have  $\mu_U = \mathbb{E}(U)$ ,  $\mu_V = \mathbb{E}(V)$ ,  $\mathbf{C}_U = \mathbb{V}\text{ar}(U)$ ,  $\mathbf{C}_V = \mathbb{V}\text{ar}(V)$ ,  $\mathbf{C}_{UV} = \mathbb{C}\text{ov}(U, V)$ ,  $\mathbf{C}_{VU} = \mathbb{C}\text{ov}(V, U)$ . Note that

$$U \sim \mathcal{N}(\mu_U, \mathbf{C}_U), \quad V \sim \mathcal{N}(\mu_V, \mathbf{C}_V).$$

The conditional density of  $U$  given  $V$  is  $\mathcal{N}(\mu_{U|V}, \mathbf{C}_{U|V})$ , where

$$\mu_{U|V} = \mathbf{C}_{UV} \mathbf{C}_V^{-1} (V - \mu_V) + \mu_U,$$

$$\mathbf{C}_{U|V} = \mathbf{C}_U - \mathbf{C}_{UV} \mathbf{C}_V^{-1} \mathbf{C}_{VU}. \quad (\text{Schur complement})$$

- To summarize, all marginals and conditionals of a multivariate Gaussian distribution are Gaussian.