Lecture 1: Fundamentals of probability

School of Mathematical Sciences, Xiamen University

1. Elementary probabilities

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

  (i) $\Omega \in \mathcal{B}$, and if $A \in \mathcal{B}$, then $\Omega \setminus A \in \mathcal{B}$,

  (ii) if $A_1, A_2, \ldots \in \mathcal{B}$, then $\cup_{j=1}^{\infty} A_j \in \mathcal{B}$.

  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

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

  that must satisfying the following conditions:

  (i) $\mathbb{P}(\Omega) = 1$, and for all $E \in \mathcal{B}$, $0 \leq \mathbb{P}(E) \leq 1$,

  (ii) if $A_j \in \mathcal{B}$, $j = 1, 2, \ldots$ with $A_j \cap A_k = \emptyset$ whenever $j \neq k$, then

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

- 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,

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

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 Ω, a real valued random variable X is a mapping
$$X : \Omega \to \mathbb{R},$$
which assigns to each element of Ω 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 μ_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 Ω mapped onto A by the random variable X.

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

$$\mu_X(A) = \int_A \pi_X(x)\mathrm{d}x.$$

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

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

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

Conversely, 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

$$\Phi_X(x) = \int_{-\infty}^{x} \pi_X(x')\mathrm{d}x' = \mathbb{P}\{X \le x\}.$$

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

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

- 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

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

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

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix},$$

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

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

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

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

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

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

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

of its components, satisfying

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

- The joint probability density π_{XY} of two multivariate random variables $X : \Omega \to \mathbb{R}^n$ and $Y : \Omega \to \mathbb{R}^m$ 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,

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

  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

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

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

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

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,

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

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

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

Observe that the symmetric roles of X and Y imply that

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

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

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

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

First, note that since ϕ 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

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

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

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

we obtain

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

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

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

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

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

- 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 $\phi : \mathbb{R}^n \to \mathbb{R}^n$ 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 ϕ. Then we may write

$$\int_B \pi_Z(z)\mathrm{d}z = \int_{B'} \phi(X)\mathrm{d}x.$$

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

$$\mathrm{d}x = |\det(D\phi(z))|\mathrm{d}z,$$

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

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

and its determinant, the Jacobian determinant, expresses the local volume scaling of the mapping ϕ. 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,

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

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 π_X, its *expected value*, or *mean*, is defined as

$$\mathbb{E}(X) = \overline{x} = \int_{\mathbb{R}} x \pi_X(x) \mathrm{d}x \in \mathbb{R}.$$

- 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 π_X, and a function $f : \mathbb{R}^n \to \mathbb{R}$, we define the *expectation* of $f(X)$ as

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

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

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

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

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

or, component-wise,

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

Exercise: Prove that the jth 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,

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

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

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

4.2 Conditional expectation

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

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

- Compute the expectation of X via its conditional expectation:

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

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,

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}((X-\overline{x})^2) = \sigma_X^2 = \int_{\mathbb{R}} (x-\overline{x})^2 \pi_X(x)\mathrm{d}x.$$

The square root σ_X of the variance is the *standard deviation* of X. Obviously, it holds $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - \overline{x}^2 \leq \mathbb{E}(X^2)$.

- The kth moment of a probability density function is defined as

$$\mathbb{E}((X-\overline{x})^k) = \int_{\mathbb{R}} (x-\overline{x})^k \pi_X(x)\mathrm{d}x.$$

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

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

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

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

  X and Y are said to be uncorrelated if $\mathbb{C}\mathrm{ov}(X, Y) = 0$.
- If the random variables X and Y are independent, then

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

  Also, for any real λ, it holds $\mathbb{V}\mathrm{ar}(\lambda X) = \lambda^2 \mathbb{V}\mathrm{ar}(X)$.
- Given a random variable $X \in \mathbb{R}^n$ with probability density π_X, the *covariance* of X is an $n \times n$ matrix with elements

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

  Alternatively, we can define the covariance using vector notation as

$$\mathbb{C}\mathrm{ov}(X, X) = \int_{\mathbb{R}^n} (x - \overline{x})(x - \overline{x})^\top \pi_X(x)\mathrm{d}x \in \mathbb{R}^{n \times n}.$$

- The variance of the jth component X_j of X is

$$\mathbb{V}\mathrm{ar}(X_j) = \int_{\mathbb{R}} (x_j - \overline{x}_j)^2 \pi_{X_j}(x_j) \mathrm{d}x_j.$$

The jth diagonal entry of $\mathbb{C}\mathrm{ov}(X, X)$ is

$$\mathbb{C}\mathrm{ov}(X, X)_{jj} = \int_{\mathbb{R}^n} (x_j - \overline{x}_j)^2 \pi_X(x) \mathrm{d}x.$$

Exercise: Prove that

$$\mathbb{V}\mathrm{ar}(X_j) = \mathbb{C}\mathrm{ov}(X, X)_{jj}, \quad 1 \leq j \leq n.$$

- We also use the notation $\mathbb{V}\mathrm{ar}(X)$ to denote $\mathbb{C}\mathrm{ov}(X, X)$.

Exercise: Prove that

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}((X - \overline{x})(X - \overline{x})^\top) = \mathbb{E}(XX^\top) - \overline{x}\,\overline{x}^\top.$$

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,

$$\begin{aligned} \mathbb{C}\mathrm{ov}(X, Y) &= \mathbb{E}((X - \overline{x})(Y - \overline{y})^\top) \\ &= \mathbb{E}(XY^\top) - \overline{x}\,\overline{y}^\top, \end{aligned}$$

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

Exercise: Prove that

$$\mathbb{C}\mathrm{ov}(X, Y) = (\mathbb{C}\mathrm{ov}(Y, X))^\top.$$

6. Other properties of expectation, variance, and covariance

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

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

- 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,

$$|\mathbb{C}\mathrm{ov}(X, Y)| \leq \sigma_X \sigma_Y.$$

  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 ψ 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 $\alpha > 0$,

$$\mathbb{P}\{|X - \mathbb{E}(X)| \geq \alpha\} \leq \frac{\mathbb{V}\mathrm{ar}(X)}{\alpha^2}.$$

- $\mathbb{C}\mathrm{ov}$ is bilinear and shift invariant:

  For any constants a and b and any c ,

$$\mathbb{C}\mathrm{ov}(aX + bY + c, Z) = a\mathbb{C}\mathrm{ov}(X, Z) + b\mathbb{C}\mathrm{ov}(Y, Z),$$

$$\mathbb{C}\mathrm{ov}(Z, aX + bY + c) = a\mathbb{C}\mathrm{ov}(Z, X) + b\mathbb{C}\mathrm{ov}(Z, Y).$$

  In particular,

$$\mathbb{V}\mathrm{ar}(X \pm Y) = \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y) \pm (\mathbb{C}\mathrm{ov}(X, Y) + \mathbb{C}\mathrm{ov}(Y, X)).$$

- Covariance transformation:

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

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

  In particular,

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

- Expectation of a quadratic form:

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

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

  where 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

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

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

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

- 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

$$\mathbb{E}(X) = \mu, \quad \mathbb{V}\mathrm{ar}(X) = \mathbf{C}.$$

Exercise: Assume $X \sim \mathcal{N}(\mu, \mathbf{C})$. Prove that the n components X_j, $1 \le j \le n$, 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 $\begin{bmatrix} X_{k_1} & \cdots & X_{k_\ell} \end{bmatrix}^\top$ 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.