Lecture Notes: Random Vectors

From Lecture by Hagit

May 1, 2025

Announcements and Course Logistics

Here are some administrative points discussed during the lecture:

- **Lecture Notes and Examples:** I understand that more examples are needed to fully grasp the material, especially since this might be the first time many of you are encountering multidimensional random variables in depth. While I cannot dedicate more class time to examples without falling behind on the syllabus, I will:
  - Provide these written notes, which aim to be clearer and more structured than the live lecture.
  - Add supplementary written examples to the notes I post, covering the concepts we discussed. I will try to post these notes shortly after the lecture.
  - I saw the solution to Exercise 1 was posted on Moodle. Please review it. I noticed some errors in Exercise 2 submissions that might have been avoided by fully understanding the concepts (like atoms) from Exercise 1. Ensure you understand the provided solutions.
- **Tutorials and Exercises:**
  - There will be **no tutorial session next week**.
  - I acknowledge the feedback that the current tutorial structure might not be sufficient given the amount of material and that there can be a gap between understanding the lecture and solving the problems.
  - Please make use of the provided exercise solutions. Solving exercises, even those not covered in the tutorial, is crucial.
- **Office Hours and Extra Help:**
  - My regular office hours are available. Please utilize them if you have questions. Currently, they are not being fully utilized.
  - If the scheduled office hours do not work for you, please coordinate among yourselves. If a group of students can agree on a suitable time, I am willing to schedule an additional session (1-2 hours) to work through problems and examples together in a classroom setting. Please let me know if you organize such a group. (Note: I need to confirm this availability with the teaching administration).

- **AI Tool Recommendation:** For those interested in exploring concepts further or getting alternative explanations, I recommend looking into the AI tool "Perplexity AI". Unlike some other language models, it tends to be better grounded in academic literature and research papers, which can be very helpful for technical questions in statistics and mathematics. It often provides references for its answers. There's a free tier, though it might limit the number of complex queries. (I used it to get a detailed explanation of Jensen's inequality, for example). Try it out for exploring concepts, but remember it's a tool, not a replacement for understanding.

- **Course Structure Remaining:** We are nearing the end of the current major chapter on random variables/vectors. Following this, we have approximately four smaller chapters remaining, covering topics such as:

  - Transformations of Random Variables/Vectors
  - The Multivariate Normal Distribution
  - Modes of Convergence for Random Variables
  - Limit Theorems (like the Central Limit Theorem)
  - Possibly topics related to dependence structures and linearity in subspaces.

  These remaining chapters are generally shorter than the one we are currently finishing.

- **Quiz Material:** I will ensure the relevant lecture notes covering material up to the upcoming quiz are made available. (Regarding the specific email about quiz links/details, I have received it and will address it).

Please feel free to reach out with questions about these administrative points or the course material.

1 Introduction to Random Vectors

So far in our study of probability, we've primarily focused on **scalar random variables** – variables that map outcomes from a sample space Ω to a single real number in $\mathbb{R}$. However, many real-world phenomena involve multiple random quantities that are often related. For instance, consider the height and weight of a randomly selected person, the temperature and pressure at a certain location, or the position (X, Y, Z) of a particle in space.

To model such situations, we introduce the concept of a **random vector**.

Definition 1.1 (Random Vector)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A function $X : \Omega \to \mathbb{R}^n$ (where $n \geq 1$) is called an n-dimensional **random vector** if for every $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, the set $\{\omega \in \Omega \mid X(\omega) \leq x\}$ is an event in $\mathcal{F}$. The inequality $X(\omega) \leq x$ is interpreted component-wise, meaning $X_i(\omega) \leq x_i$ for all $i = 1, \ldots, n$, where $X(\omega) = (X_1(\omega), \ldots, X_n(\omega))$.*

Essentially, a random vector is a vector whose components $X_1, \ldots, X_n$ are themselves scalar random variables defined on the same probability space.

Remark 1.2 (Notation Convention)**.** *Unless otherwise specified, we will treat random vectors as **column vectors**:*

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

To represent a row vector, we will use the transpose notation, $X^T = (X_1, X_2, \ldots, X_n)$. This convention is standard in many areas, particularly when dealing with linear algebra operations involving vectors and matrices.

2 Joint Cumulative Distribution Function (CDF)

Just as the cumulative distribution function (CDF) characterizes the distribution of a scalar random variable, we can define a joint CDF to characterize the distribution of a random vector.

Definition 2.1 (Joint CDF)**.** *The **joint cumulative distribution function (CDF)** of an n-dimensional random vector $X = (X_1, \ldots, X_n)^T$ is the function $F_X : \mathbb{R}^n \to [0, 1]$ defined by:*

$$F_X(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

We often write this more compactly as $F_X(x) = \mathbb{P}(X \leq x)$, where $x = (x_1, \ldots, x_n)^T$ and the inequality is understood component-wise.

The joint CDF $F_X(x)$ gives the probability that the random vector X falls into the n-dimensional semi-infinite interval $(-\infty, x_1] \times (-\infty, x_2] \times \cdots \times (-\infty, x_n]$.

2.1 Properties of the Joint CDF

The joint CDF shares some properties with the scalar CDF, but with added complexity due to the multiple dimensions. A function $F : \mathbb{R}^n \to \mathbb{R}$ is a valid joint CDF if and only if it satisfies the following properties:

Property 2.2 (Normalization Limits)**.**

(a) *For any fixed* $i \in \{1, \ldots, n\}$,

$$\lim_{x_i \to -\infty} F_X(x_1, \ldots, x_n) = 0$$

This limit must hold even if the other components x_j *(for* $j \neq i$*) are held constant.*

(b)

$$\lim_{x_1 \to \infty, \ldots, x_n \to \infty} F_X(x_1, \ldots, x_n) = 1$$

Here, all components must tend to infinity simultaneously (in any manner).

Remark 2.3 (On the Limits)**.** *The condition for the limit to 0 as* $x_i \to -\infty$ *is stronger than requiring the limit to be 0 only when all components go to* $-\infty$ *simultaneously. We need this component-wise condition to ensure that the probability measure derived from* F_X *is well-behaved. For the limit to 1, however, it suffices that all components go to* ∞*. This addresses a question raised in the lecture: the requirement for the lower limit is indeed element-wise (stronger), while the upper limit allows for joint convergence (weaker).*

Property 2.4 (Right-Continuity)**.** $F_X(x)$ *is continuous from the right in each argument. That is, for any* $x \in \mathbb{R}^n$ *and any* $i \in \{1, \ldots, n\}$,

$$\lim_{\epsilon \downarrow 0^+} F_X(x + \epsilon e_i) = F_X(x)$$

where e_i *is the i-th standard basis vector (a vector with 1 in the i-th position and 0s elsewhere).*

Property 2.5 (Non-negativity / Rectangle Inequality)**.** *For any hyperrectangle* $(a, b] = (a_1, b_1] \times \cdots \times (a_n, b_n]$ *where* $a_i < b_i$ *for all i, the probability mass assigned to this rectangle by* F_X *must be non-negative. This probability is calculated using an inclusion-exclusion-like formula involving the values of* F_X *at the* 2^n *vertices of the hyperrectangle.*

Formally, let Δ_{a_i,b_i} *be the difference operator acting on the i-th argument of a function* $H : \mathbb{R}^n \to \mathbb{R}$*, defined as:*

$$(\Delta_{a_i,b_i} H)(x) = H(x_1, \ldots, x_{i-1}, b_i, x_{i+1}, \ldots, x_n) - H(x_1, \ldots, x_{i-1}, a_i, x_{i+1}, \ldots, x_n)$$

Then, for any $a, b \in \mathbb{R}^n$ *with* $a_i < b_i$ *for all i, we must have:*

$$\mathbb{P}(a < X \leq b) = (\Delta_{a_n,b_n} \circ \cdots \circ \Delta_{a_1,b_1} F_X)(x) \geq 0$$

(The point x at which the final result is evaluated doesn't matter once all differences are taken). This condition generalizes the monotonicity property $F(b) - F(a) \geq 0$ *from the 1D case.*

Example 2.6 (Non-negativity in 2D)**.** *Let* $n = 2$*. We want to find the probability* $\mathbb{P}(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2)$ *for* $a_1 < b_1$ *and* $a_2 < b_2$*. This corresponds to the probability mass in the rectangle* $(a_1, b_1] \times (a_2, b_2]$*. Applying the difference operators:*

$$\begin{aligned}
\mathbb{P}(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2) &= (\Delta_{a_2,b_2} \circ \Delta_{a_1,b_1} F_X)(x) \\
&= \Delta_{a_2,b_2}[F_X(b_1, x_2) - F_X(a_1, x_2)] \\
&= [F_X(b_1, b_2) - F_X(a_1, b_2)] - [F_X(b_1, a_2) - F_X(a_1, a_2)] \\
&= F_X(b_1, b_2) - F_X(a_1, b_2) - F_X(b_1, a_2) + F_X(a_1, a_2)
\end{aligned}$$

The non-negativity property requires that this quantity must be ≥ 0 *for all* $a_1 < b_1, a_2 < b_2$*. This is the inclusion-exclusion principle for calculating the probability of the rectangle using the CDF values at its corners. Visualizing this:* $F_X(b_1, b_2)$ *is the probability in* $(-\infty, b_1] \times (-\infty, b_2]$*. We subtract the probabilities in the regions* $(-\infty, a_1] \times (-\infty, b_2]$ *and* $(-\infty, b_1] \times (-\infty, a_2]$*. Since we subtracted the region* $(-\infty, a_1] \times (-\infty, a_2]$ *twice, we add it back once.*

Any function F_X satisfying Properties 2.2, 2.4, and 2.5 is a valid joint CDF, meaning we can construct a probability space and a random vector X having F_X as its CDF.

3 Types of Random Vectors

Similar to scalar random variables, random vectors can be classified based on the nature of their distribution.

3.1 Absolutely Continuous Random Vectors

This is a common case where the probability mass is spread smoothly over $\mathbb{R}^n$.

Definition 3.1 (Joint PDF)**.** *A random vector X is **absolutely continuous** if its CDF $F_X(x)$ can be represented as the integral of a non-negative function $f_X : \mathbb{R}^n \to [0, \infty)$, called the **joint probability density function (PDF)**:*

$$F_X(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_X(u_1, \ldots, u_n) du_n \ldots du_1$$

The PDF must satisfy $f_X(x) \geq 0$ for all x and $\int_{\mathbb{R}^n} f_X(x)dx = 1$. If F_X is sufficiently differentiable, the PDF can be obtained by differentiation:

$$f_X(x_1, \ldots, x_n) = \frac{\partial^n F_X(x_1, \ldots, x_n)}{\partial x_1 \ldots \partial x_n}$$

(wherever the derivatives exist).

For an absolutely continuous random vector, the probability of X falling into a region $\Gamma \subseteq \mathbb{R}^n$ is found by integrating the PDF over that region:

$$\mathbb{P}(X \in \Gamma) = \int_\Gamma f_X(x)dx$$

Note that for continuous vectors, the probability of X taking any single specific value x_0 is zero, i.e., $\mathbb{P}(X = x_0) = 0$. Similarly, the probability of X lying on any lower-dimensional surface (like a line in $\mathbb{R}^2$ or a plane in $\mathbb{R}^3$) is also zero, as these sets have zero volume (Lebesgue measure) in $\mathbb{R}^n$.

Example 3.2 (3D Uniform Distribution)**.** *Consider a particle whose random location $X = (X_1, X_2, X_3)^T$ is uniformly distributed within a 3D box (a hyperrectangle) defined by $H = (a_1, b_1] \times (a_2, b_2] \times (a_3, b_3]$. The volume of the box is $V = (b_1 - a_1)(b_2 - a_2)(b_3 - a_3)$. The joint PDF is constant within the box and zero outside:*

$$f_X(x_1, x_2, x_3) = \begin{cases} \frac{1}{V} & \text{if } x \in H \\ 0 & \text{otherwise} \end{cases}$$

Let's take a specific case: the unit cube, where $a_i = 0$ and $b_i = 1$ for $i = 1, 2, 3$. Then $V = 1$, and $f_X(x) = 1$ for $x \in (0, 1]^3$ and 0 otherwise. Suppose we want to find the probability that the particle

is in the bottom third of the cube, i.e., $0 < X_3 \leq 1/3$. *The region is* $\Gamma = (0,1] \times (0,1] \times (0,1/3]$.

$$\begin{aligned}
\mathbb{P}(X \in \Gamma) &= \int_0^1 \int_0^1 \int_0^{1/3} f_X(x_1, x_2, x_3) dx_3 dx_2 dx_1 \\
&= \int_0^1 \int_0^1 \int_0^{1/3} 1 \, dx_3 dx_2 dx_1 \\
&= \int_0^1 \int_0^1 [x_3]_0^{1/3} dx_2 dx_1 \\
&= \int_0^1 \int_0^1 \frac{1}{3} dx_2 dx_1 \\
&= \frac{1}{3} \int_0^1 [x_2]_0^1 dx_1 = \frac{1}{3} \int_0^1 1 dx_1 = \frac{1}{3} [x_1]_0^1 = \frac{1}{3}
\end{aligned}$$

This makes intuitive sense: since the distribution is uniform, the probability of being in a sub-volume is proportional to the sub-volume's size relative to the total volume. The bottom third has volume $1 \times 1 \times (1/3) = 1/3$, *which is* $1/3$ *of the total volume 1. (This type of uniform distribution is used in robotics for sampling configurations or planning paths).*

3.2 Discrete Random Vectors

Here, the probability mass is concentrated on a countable set of points.

Definition 3.3 (Joint PMF)**.** *A random vector* X *is **discrete** if it takes values in a countable set* $S = \{x^{(1)}, x^{(2)}, \dots\} \subset \mathbb{R}^n$. *Its distribution is characterized by the **joint probability mass function (PMF)*** $p_X : \mathbb{R}^n \to [0,1]$ *defined by:*

$$p_X(x) = \mathbb{P}(X = x)$$

The PMF is non-zero only for $x \in S$, *and must satisfy* $p_X(x) \geq 0$ *and* $\sum_{x \in S} p_X(x) = 1$.

For a discrete random vector, the probability of X falling into a region $\Gamma \subseteq \mathbb{R}^n$ is found by summing the PMF values for all points in the support S that are also in Γ:

$$\mathbb{P}(X \in \Gamma) = \sum_{x^{(i)} \in S \cap \Gamma} p_X(x^{(i)})$$

Example 3.4 (Multinomial Distribution)**.** *The multinomial distribution is a multivariate generalization of the binomial distribution. Consider an experiment with* k *possible outcomes, with probabilities* $p_1, p_2, \dots, p_k$, *where* $p_j \geq 0$ *and* $\sum_{j=1}^k p_j = 1$. *We perform* n *independent trials of this experiment. Let the random vector* $X = (X_1, X_2, \dots, X_k)^T$ *count the number of times each outcome* j *occurred, so* X_j *is the count for outcome* j. *Clearly, each* X_j *must be an integer between 0 and* n, *and the total number of trials is fixed:* $\sum_{j=1}^k X_j = n$. *The joint PMF for* X *is given by the multinomial formula:*

$$p_X(x_1, \dots, x_k) = \mathbb{P}(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

This formula is valid for non-negative integers $x_1, \dots, x_k$ *such that* $\sum_{j=1}^k x_j = n$. *Otherwise,* $p_X(x) = 0$. *The parameters of the distribution are the number of trials* n *and the vector of probabilities* $p = (p_1, \dots, p_k)^T$. *The vector* p *lies in the standard* $(k-1)$*-simplex,* $S_{k-1} = \{p \in \mathbb{R}^k \mid p_j \geq$

0 for all $j, \sum_{j=1}^{k} p_j = 1\}$. It's called the $(k-1)$-simplex because although there are k components, the sum constraint reduces the degrees of freedom by one. Note that if $k = 2$, we have $X_1 + X_2 = n$ and $p_1 + p_2 = 1$. Let $X_1 = x$, then $X_2 = n - x$. Let $p_1 = p$, then $p_2 = 1 - p$. The formula becomes

$$\frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

which is exactly the binomial PMF for $X_1 \sim Binomial(n, p)$.

3.3 Mixed and Degenerate Distributions

It's possible for distributions to be neither purely continuous nor purely discrete. Furthermore, the geometric complexity of $\mathbb{R}^n$ allows for distributions concentrated on lower-dimensional subsets.

Example 3.5 (Distribution Concentrated on a Line)**.** *Let ξ be a standard 1D continuous random variable with CDF F_ξ and PDF f_ξ. Consider the 2D random vector $X = (X_1, X_2)^T$ defined by $X_1 = \xi$ and $X_2 = \xi$. The joint CDF is:*

$$\begin{aligned} F_X(x_1, x_2) &= \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2) \\ &= \mathbb{P}(\xi \leq x_1, \xi \leq x_2) \\ &= \mathbb{P}(\xi \leq \min(x_1, x_2)) \\ &= F_\xi(\min(x_1, x_2)) \end{aligned}$$

This distribution is interesting. Can we find a joint PDF $f_X(x_1, x_2)$? If we try to differentiate $\frac{\partial^2 F_X}{\partial x_1 \partial x_2}$, we run into problems because of the non-differentiability of the minimum function along the line $x_1 = x_2$. Intuitively, all the probability mass of X lies entirely on the line $x_1 = x_2$ in the $\mathbb{R}^2$ plane. Since a line has zero area (zero 2D Lebesgue measure), the probability density must be infinite on the line and zero elsewhere, meaning a standard PDF does not exist. This distribution is neither absolutely continuous (no 2D PDF) nor discrete (since ξ is continuous, X can take uncountably many values on the line). It's concentrated on a lower-dimensional manifold. We can calculate probabilities like $\mathbb{P}(X_1 = X_2)$. Since X_1 and X_2 are always equal to ξ, this event is $\mathbb{P}(\xi = \xi)$, which is always true, so the probability is 1. If we were to formally try and compute $\mathbb{P}(X_1 = X_2)$ using a hypothetical PDF f_X by integrating over the region where $x_1 = x_2$:

$$\mathbb{P}(X_1 = X_2) = \iint_{x_1 = x_2} f_X(x_1, x_2) dx_1 dx_2$$

This integral would be 0, because we are integrating over a set (a line) of measure zero in $\mathbb{R}^2$. This apparent contradiction confirms that a standard 2D PDF cannot describe this distribution.

4 Marginal Distributions

Often, we have a joint distribution for a vector $X = (X_1, \ldots, X_n)^T$, but we are interested in the distribution of only a subset of its components. This leads to the concept of marginal distributions.

Let $X = (X_1, \ldots, X_n)^T$ be a random vector. Let $J = \{j_1, \ldots, j_k\} \subset \{1, \ldots, n\}$ be a subset of indices, and let $X_J = (X_{j_1}, \ldots, X_{j_k})^T$ be the sub-vector containing only the components indexed by J. We want to find the distribution of X_J.

Definition 4.1 (Marginal CDF)**.** *The **marginal CDF** of the sub-vector X_J is obtained from the joint CDF $F_X(x_1, \ldots, x_n)$ by letting the arguments x_i corresponding to indices $i \notin J$ go to infinity:*

$$F_{X_J}(x_{j_1}, \ldots, x_{j_k}) = \lim_{x_i \to \infty \text{ for all } i \notin J} F_X(x_1, \ldots, x_n)$$

Example 4.2 (Marginal CDF from 3D)**.** *Let $X = (X_1, X_2, X_3)^T$ with joint CDF $F_X(x_1, x_2, x_3)$. Suppose we are interested in the marginal distribution of $X_J = (X_1, X_2)^T$ (so $J = \{1, 2\}$). The marginal CDF is:*

$$F_{X_1,X_2}(x_1, x_2) = \lim_{x_3 \to \infty} F_X(x_1, x_2, x_3)$$

Intuitively, letting $x_3 \to \infty$ removes the constraint $X_3 \leq x_3$, leaving only $\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2)$.

If the joint distribution is described by a PDF or PMF, the marginal distribution can be found more directly:

Definition 4.3 (Marginal PDF/PMF)**.**

- ***Continuous Case:** If X has joint PDF $f_X(x_1, \ldots, x_n)$, the **marginal PDF** of the sub-vector $X_J = (X_{j_1}, \ldots, x_{j_k})^T$ is obtained by integrating f_X over all possible values of the variables X_i for $i \notin J$:*

  $$f_{X_J}(x_{j_1}, \ldots, x_{j_k}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x_1, \ldots, x_n) \prod_{i \notin J} dx_i$$

  (This involves $n - k$ integrations).

- ***Discrete Case:** If X has joint PMF $p_X(x_1, \ldots, x_n)$, the **marginal PMF** of the sub-vector $X_J = (X_{j_1}, \ldots, x_{j_k})^T$ is obtained by summing p_X over all possible values of the variables X_i for $i \notin J$:*

  $$p_{X_J}(x_{j_1}, \ldots, x_{j_k}) = \sum_{x_i \text{ for } i \notin J} p_X(x_1, \ldots, x_n)$$

We say we "integrate out" or "sum out" the unwanted variables.

5 Expectation of Random Vectors

The concept of expectation extends naturally to random vectors.

Definition 5.1 (Expectation of a Random Vector)**.** *The **expectation** (or expected value or mean vector) of a random vector $X = (X_1, \ldots, X_n)^T$ is the vector of the expectations of its components:*

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}$$

provided each component expectation $\mathbb{E}[X_i]$ exists and is finite.

We can also compute the expectation of a function of a random vector. Let $H : \mathbb{R}^n \to \mathbb{R}^m$ be a function, so $H(X)$ is an m-dimensional random vector.

Definition 5.2 (Expectation of a Function of a Random Vector)**.** *Let $H(X) = (H_1(X), \ldots, H_m(X))^T$. The expectation is:*

$$\mathbb{E}[H(X)] = \begin{pmatrix} \mathbb{E}[H_1(X)] \\ \vdots \\ \mathbb{E}[H_m(X)] \end{pmatrix}$$

where each component $\mathbb{E}[H_j(X)]$ is calculated using the joint distribution of X:

- ***Continuous Case:** If X has PDF $f_X(x)$,*

$$\mathbb{E}[H_j(X)] = \int_{\mathbb{R}^n} H_j(x) f_X(x) dx$$

- ***Discrete Case:** If X has PMF $p_X(x)$ with support S,*

$$\mathbb{E}[H_j(X)] = \sum_{x \in S} H_j(x) p_X(x)$$

assuming the integrals or sums converge absolutely.

5.1 Properties of Expectation

Many properties of expectation from the scalar case carry over, especially linearity.

Property 5.3 (Linearity of Expectation)**.** *Let X and Y be random vectors (of possibly different dimensions, say n and p), and let A and B be constant matrices of appropriate dimensions (e.g., A is $m \times n$, B is $m \times p$). Then*

$$\mathbb{E}[AX + BY] = A\mathbb{E}[X] + B\mathbb{E}[Y]$$

provided $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ exist.

6 Covariance Matrix

To describe the relationships *between* the components of a random vector, or between two different random vectors, we use the covariance.

Definition 6.1 (Covariance Matrix)**.** *Let X be an n-dimensional random vector and Y be an m-dimensional random vector, both defined on the same probability space, with finite expectations $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. The **covariance matrix** between X and Y is the $n \times m$ matrix denoted $Cov(X, Y)$ whose (i, j)-th entry is $Cov(X_i, Y_j)$:*

$$Cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^T]$$

Expanding this, the (i, j)-th element is $E[(X_i - \mu_{X,i})(Y_j - \mu_{Y,j})]$. A useful computational formula is:

$$Cov(X, Y) = \mathbb{E}[XY^T] - \mathbb{E}[X]\mathbb{E}[Y]^T$$

Here, XY^T is an $n \times m$ matrix (outer product), and $\mathbb{E}[XY^T]$ is the matrix of expectations of its entries. Similarly, $\mathbb{E}[X]\mathbb{E}[Y]^T$ is the outer product of the mean vectors.

A particularly important case is the covariance of a random vector with itself.

Definition 6.2 (Covariance Matrix of a Vector)**.** *The **covariance matrix** (or variance-covariance matrix) of an n-dimensional random vector X with mean $\mu_X = \mathbb{E}[X]$ is the $n \times n$ matrix:*

$$\Sigma_X = Cov(X) = Cov(X, X) = \mathbb{E}[(X - \mu_X)(X - \mu_X)^T]$$

The computational formula is:

$$Cov(X) = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T$$

Remark 6.3 (Structure of Cov(X))**.**

- *The (i, i)-th diagonal entry is $\mathbb{E}[(X_i - \mu_{X,i})^2] = Var(X_i)$, the variance of the i-th component.*
- *The (i, j)-th off-diagonal entry (for $i \neq j$) is $\mathbb{E}[(X_i - \mu_{X,i})(X_j - \mu_{X,j})] = Cov(X_i, X_j)$, the covariance between the i-th and j-th components.*
- *The matrix is symmetric: $Cov(X_i, X_j) = Cov(X_j, X_i)$, so $\Sigma_X = \Sigma_X^T$.*
- *The matrix is positive semi-definite. That is, for any constant vector $a \in \mathbb{R}^n$, $a^T \Sigma_X a \geq 0$. This is because $a^T \Sigma_X a = a^T \mathbb{E}[(X - \mu_X)(X - \mu_X)^T]a = \mathbb{E}[a^T(X - \mu_X)(X - \mu_X)^T a] = \mathbb{E}[(a^T(X - \mu_X))^2] = Var(a^T X) \geq 0$.*

Property 6.4 (Trace of the Covariance Matrix)**.** *The **trace** of a square matrix is the sum of its diagonal elements, $Tr(A) = \sum_i A_{ii}$. For the covariance matrix $Cov(X)$, the trace is the sum of the variances of the components:*

$$Tr(Cov(X)) = \sum_{i=1}^{n} Var(X_i)$$

We can also relate this to the expected squared Euclidean distance of X from its mean μ_X:

$$\begin{aligned}
Tr(Cov(X)) &= Tr(\mathbb{E}[(X - \mu_X)(X - \mu_X)^T]) \\
&= \mathbb{E}[Tr((X - \mu_X)(X - \mu_X)^T)] \quad \textit{(Linearity of Trace and Expectation)} \\
&= \mathbb{E}[Tr((X - \mu_X)^T(X - \mu_X))] \quad \textit{(Using } Tr(AB) = Tr(BA)\textit{)} \\
&= \mathbb{E}[(X - \mu_X)^T(X - \mu_X)] \quad \textit{(Since } v^T v \textit{ is a scalar, } Tr(v^T v) = v^T v\textit{)} \\
&= \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu_{X,i})^2\right] \\
&= \mathbb{E}[\|X - \mu_X\|_2^2]
\end{aligned}$$

The key steps are swapping trace and expectation (which is valid due to linearity) and using the property $Tr(AB) = Tr(BA)$. Since $(X - \mu_X)^T(X - \mu_X)$ is the dot product of the vector $X - \mu_X$ with itself, it's a 1×1 matrix (a scalar) equal to the squared Euclidean norm $\|X - \mu_X\|_2^2$. The trace of a scalar is just the scalar itself. This derivation clarifies the relationship noted at the end of the lecture.

Remark 6.5 (Moment Generating Functions)**.** *The concept of moment generating functions (MGFs) also extends to random vectors, providing a way to characterize the distribution and derive moments. However, we will not cover vector MGFs in detail due to time constraints. The definition exists and is analogous to the scalar case, but its properties and applications in the multivariate setting are more involved.*