Chapter 4

Multivariate Random Variables

In this chapter, we will review some topics related to random vectors, which will be of use in the following chapters.

4.1 Review of Linear Algebra

For two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the **inner product** $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ of $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i.$$

where $\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$.

The **length** or the ℓ_2 norm of a vector $\boldsymbol{x}$ is $\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^T \boldsymbol{x}}$ and we have $\|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^T \boldsymbol{x}$. Let α be the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$. Then $\boldsymbol{x}^T \boldsymbol{y} = \|\boldsymbol{x}\| \|\boldsymbol{y}\| \cos \alpha$. If $\boldsymbol{x}^T \boldsymbol{y} = 0$, then the two are called **orthogonal**.

For a collection of vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$, a **linear combination** of these is any vector of the form $a_1 \boldsymbol{v}_1 + \cdots + a_m \boldsymbol{v}_m, a_i \in \mathbb{R}$. The set of all linear combinations of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ is their **span** and denoted as $\text{Span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$. This is a **subspace** (think line, plane, or the whole space). For a matrix A, the span of the columns of A is the **column space** of A.

The vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ are **linearly independent** if there is no vector among them that can be written as a linear combination of the others, and linearly dependent otherwise. The vectors are linearly independent if and only if the only values for $a_1, \ldots, a_m$ satisfying $a_1 \boldsymbol{v}_1 + \cdots + a_m \boldsymbol{v}_m = 0$ are $a_1, \ldots, a_m = 0$. In particular, the columns of a matrix A are linearly independent if and only if the only vector $\boldsymbol{a}$ satisfying $A\boldsymbol{a} = 0$ is $\boldsymbol{a} = 0$.

The **inverse** of a square matrix A is a matrix A^{-1} such that $AA^{-1} = A^{-1}A = I$, where I is the **identity matrix**, which has 1s on the diagonal and 0s elsewhere. A matrix that has an inverse is called **invertible**. For a square matrix A, the following are equivalent:

- It is invertible.
- For all distinct vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, we have $A\boldsymbol{a} \neq A\boldsymbol{b}$.
- The only solution to $A\boldsymbol{x} = 0$ is $\boldsymbol{x} = 0$.
- Its columns are linearly independent.
- Its determinant $|A|$ is nonzero. (We also have $|A^{-1}| = \frac{1}{|A|}$.)

Given a subspace S (e.g., a plane or the column space of a matrix) and a vector $\boldsymbol{y}$, let $\hat{\boldsymbol{y}}$ be the vector in the subspace that is closest to $\boldsymbol{y}$. That is, we find $\hat{\boldsymbol{y}} \in S$ such that $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|$ is minimized. Then $\hat{\boldsymbol{y}}$ is called the **projection** of $\boldsymbol{y}$ onto the subspace S.

Lemma 4.1. *Let $\hat{\boldsymbol{y}}$ be the projection of a vector $\boldsymbol{y}$ onto a subspace S. Then $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is orthogonal to every vector in S.*

Proof. Suppose that this is not the case. Then there is a nonzero vector $\boldsymbol{v} \in S$ such that $(\boldsymbol{y}-\hat{\boldsymbol{y}})^T\boldsymbol{v} \neq 0$. We will show that this contradicts the minimality of $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|$. For any $a \in \mathbb{R}$,

$$\begin{aligned}\|\boldsymbol{y} - \hat{\boldsymbol{y}} - a\boldsymbol{v}\|_2^2 &= (\boldsymbol{y} - \hat{\boldsymbol{y}} - a\boldsymbol{v})^T(\boldsymbol{y} - \hat{\boldsymbol{y}} - a\boldsymbol{v}) \\ &= \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 - 2a\boldsymbol{v}^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) + a^2\|\boldsymbol{v}\|_2^2.\end{aligned}$$

This is a convex function in a. So setting the derivative to 0 gives the value of a that minimizes the error:

$$\frac{\partial}{\partial a}\|\boldsymbol{y} - \hat{\boldsymbol{y}} - a\boldsymbol{v}\|_2^2 = -2\boldsymbol{v}^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) + 2a\|\boldsymbol{v}\|_2^2 = 0 \Rightarrow a = \frac{\boldsymbol{v}^T(\boldsymbol{y} - \hat{\boldsymbol{y}})}{\|\boldsymbol{v}\|_2^2} \neq 0.$$

Let

$$\hat{\boldsymbol{y}}' = \hat{\boldsymbol{y}} + \frac{\boldsymbol{v}^T(\boldsymbol{y} - \hat{\boldsymbol{y}})}{\boldsymbol{v}^T\boldsymbol{v}}\boldsymbol{v},$$

and note that $\hat{\boldsymbol{y}}'$ is also in S but it is closer to $\boldsymbol{y}$ contradicting the optimality of $\hat{\boldsymbol{y}}$. □

4.2 Random vectors

A **random vector** is a vector of random variables. Consider the random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

The **expected value** of $\boldsymbol{x}$ is

$$\mathbb{E}\,\boldsymbol{x} = \begin{pmatrix} \mathbb{E}\,x_1 \\ \vdots \\ \mathbb{E}\,x_m \end{pmatrix}.$$

The **correlation matrix** of $\boldsymbol{x}$ and $\boldsymbol{y}$ is the $m \times n$ matrix $\mathbb{E}[\boldsymbol{x}\boldsymbol{y}^T]$, whose i,jth element is $\mathbb{E}[x_i y_j]$. The **cross-covariance matrix** of $\boldsymbol{x}$ and $\boldsymbol{y}$ is $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})$ is the matrix $\mathbb{E}[(\boldsymbol{x} - \mathbb{E}\,\boldsymbol{x})^T(\boldsymbol{y} - \mathbb{E}\,\boldsymbol{y})^T]$, whose i, jth element is $\operatorname{Cov}(x_i, y_j)$. The covariance of a vector $\boldsymbol{x}$ is $\operatorname{Cov}(\boldsymbol{x}) = \operatorname{Cov}(\boldsymbol{x}, \boldsymbol{x})$. The **conditional expectation** $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}]$ of $\boldsymbol{x}$ given $\boldsymbol{y}$ is a vector whose ith element is $\mathbb{E}[x_i|\boldsymbol{y}]$.

For matrices A, B, deterministic vectors $\boldsymbol{a}, \boldsymbol{b}$, and random vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{z}$, we have [1]

- $\mathbb{E}[A\boldsymbol{x} + \boldsymbol{a}] = A\,\mathbb{E}\,\boldsymbol{x} + \boldsymbol{a}$
- $\operatorname{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}[\boldsymbol{x}\boldsymbol{y}^T] - \mathbb{E}\,\boldsymbol{x}\,\mathbb{E}\,\boldsymbol{y}^T$
- $\mathbb{E}[(A\boldsymbol{x})(B\boldsymbol{y})^T] = A\,\mathbb{E}[\boldsymbol{x}\boldsymbol{y}^T]B^T$
- $\operatorname{Cov}(A\boldsymbol{x} + \boldsymbol{a}, B\boldsymbol{y} + \boldsymbol{b}) = A\operatorname{Cov}(\boldsymbol{x}, \boldsymbol{y})B^T$
- $\operatorname{Cov}(A\boldsymbol{x} + \boldsymbol{a}) = A\operatorname{Cov}(\boldsymbol{x})A^T$
- $\operatorname{Cov}(\boldsymbol{w} + \boldsymbol{x}, \boldsymbol{y} + \boldsymbol{z}) = \operatorname{Cov}(\boldsymbol{w}, \boldsymbol{y}) + \operatorname{Cov}(\boldsymbol{w}, \boldsymbol{z}) + \operatorname{Cov}(\boldsymbol{x}, \boldsymbol{y}) + \operatorname{Cov}(\boldsymbol{x}, \boldsymbol{z})$

4.3 Gaussian Random Vectors (Joint Gaussian Distribution)

Recall that a random variable x is Gaussian (normal) with mean μ and variance $\sigma^2 > 0$ if the pdf of x is given by

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

Definition 4.2. A collection of random variables is **jointly Gaussian** if any linear combination of these variables is Gaussian. A **Gaussian random vector**, also known as a multivariate normal vector, is a vector whose elements are jointly Gaussian. A collection of random vectors are jointly Gaussian if the vector obtained by concatenating them is jointly Gaussian.

Example 4.3. For example if $\begin{pmatrix} x \\ y \end{pmatrix}$ is a Gaussian vector, then $z = 2x+3y$ is Gaussian. Furthermore,

$$\begin{aligned} \mathbb{E}[z] &= 2\,\mathbb{E}[x] + 3\,\mathbb{E}[y], \\ \operatorname{Var}(z) &= \operatorname{Cov}(2x + 3y, 2x + 3y) = 4\operatorname{Cov}(x, x) + 12\operatorname{Cov}(x, y) + 9\operatorname{Cov}(y, y) \\ &= 4\operatorname{Var}(x) + 12\operatorname{Cov}(x, y) + 9\operatorname{Var}(y), \end{aligned}$$

which completely characterizes the distribution of z. △

For an m dimensional Gaussian vector $\boldsymbol{x}$, the elements of $\boldsymbol{x}$ are **independent** if and only if the covariance matrix is diagonal.

For an m-dimensional Gaussian random vector $\boldsymbol{x}$, assuming that the covariance matrix $K = \operatorname{Cov}(\boldsymbol{x})$ is invertible, we have

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{m/2}|K|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T K^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right).$$

4.4 Maximum likelihood for Gaussian Random Vectors

Let $\boldsymbol{z}$ be a Gaussian random vector of dimension d with mean $\boldsymbol{\mu}$ and covariance matrix K. If K is invertible, the pdf of $\boldsymbol{z}$ can be written as

$$p(\boldsymbol{z}|\boldsymbol{\mu}, K) = \frac{1}{\sqrt{(2\pi)^d|K|}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T K^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right),$$
$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{z}], \quad K = \mathbb{E}[(\boldsymbol{z}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})^T],$$

where $|K|$ is the determinant of K.

Given a set of n iid samples $\mathcal{D} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$, where each $\boldsymbol{z}_i$ is a d-dimensional vector, how can we estimate $\boldsymbol{\mu}$ and K using maximum likelihood? Estimating these quantities allows us to find the distribution. In particular, if we can view z_d as the output variable and $z_1, \ldots, z_{d-1}$ as input variables, then we can estimate z_d based on $z_1, \ldots, z_{d-1}$ as $\mathbb{E}[z_d|z_1, \ldots, z_{d-1}]$.

To estimate $\boldsymbol{\mu}$ and K, we write

$$\begin{aligned}\ell(\boldsymbol{\mu}, K) = \ln p(\mathcal{D}; \boldsymbol{\mu}, K) &= \sum_{i=1}^n \ln p(\boldsymbol{z}_i; \boldsymbol{\mu}, K) \\ &\doteq \frac{n}{2}\ln|K^{-1}| - \frac{1}{2}\sum_{i=1}^n (\boldsymbol{z}_i-\boldsymbol{\mu})^T K^{-1}(\boldsymbol{z}_i-\boldsymbol{\mu}),\end{aligned}$$

where we have used the fact that $|K^{-1}| = \frac{1}{|K|}$.

As seen in the appendix (last chapter), for a symmetric matrix A, we have $\frac{d}{d\boldsymbol{v}}(\boldsymbol{y}^T A\boldsymbol{y}) = 2\boldsymbol{y}^T A\frac{d\boldsymbol{y}}{d\boldsymbol{v}}$. Hence,

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}} = -\frac{1}{2}\sum_{i=1}^n 2(\boldsymbol{z}_i-\boldsymbol{\mu})^T K^{-1}(-I) = \sum_{i=1}^n (\boldsymbol{z}_i-\boldsymbol{\mu})^T K^{-1}.$$

Setting this equal to zero yields

$$\hat{\boldsymbol{\mu}}_{ML} = \bar{\boldsymbol{z}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i.$$

Exercise 4.4. Using the facts

$$\frac{\partial}{\partial A}\boldsymbol{x}^T A\boldsymbol{x} = \boldsymbol{x}\boldsymbol{x}^T, \quad \frac{\partial}{\partial A}\ln|A| = A^{-T}$$

prove that

$$\hat{K}_{ML} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{z}_i-\bar{\boldsymbol{z}})(\boldsymbol{z}_i-\bar{\boldsymbol{z}})^T$$

△

Bibliography

[1] B. Hajek, *Random Processes for Engineers*. 2014.