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

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

where $x^T$ is the transpose of $x$.

The length or the $\ell_2$ norm of a vector $x$ is $\|x\| = \|x\|_2 = \sqrt{x^T x}$ and we have $\|x\|^2 = x^T x$. Let $\alpha$ be the angle between $x$ and $y$. Then $x^T y = \|x\|\|y\| \cos \alpha$. If $x^T y = 0$, then the two are called orthogonal.

For a collection of vectors $v_1, \ldots, v_m$, a linear combination of these is any vector of the form $a_1 v_1 + \cdots + a_m v_m$, $a_i \in \mathbb{R}$. The set of all linear combinations of $v_1, \ldots, v_m$ is their span and denoted as $\text{Span}\{v_1, \ldots, v_m\}$. 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 $v_1, \ldots, v_m$ 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 $a_1 v_1 + \cdots + a_m v_m = 0$ 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 $a$ satisfying $Aa = 0$ is $a = 0$.

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. A square matrix is invertible $\iff$ for all distinct vectors $a$ and $b$, we have $Aa \neq Ab$ $\iff$ the only solution to $Ax = 0$ is $x = 0$ $\iff$ its columns are linearly independent $\iff$ 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 $y$, let $\hat{y}$ be the vector in the subspace that is closest to $y$. That is, we find $\hat{y} \in S$ such that $\|y - \hat{y}\|$ is minimized. Then $\hat{y}$ is called the projection of $y$ onto the subspace $S$. 

45
Lemma 66. Let \( \mathbf{y} \) be the projection of a vector \( \mathbf{y} \) onto a subspace \( S \). Then \( \mathbf{y} - \mathbf{\hat{y}} \) is orthogonal to every vector in \( S \).

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

\[
\| \mathbf{y} - \mathbf{\hat{y}} - a \mathbf{v} \|^2 = (\mathbf{y} - \mathbf{\hat{y}} - a \mathbf{v})^T (\mathbf{y} - \mathbf{\hat{y}} - a \mathbf{v}) = \| \mathbf{y} - \mathbf{\hat{y}} \|^2 - 2a \mathbf{v}^T (\mathbf{y} - \mathbf{\hat{y}}) + a^2 \| \mathbf{v} \|^2.
\]

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

\[
\frac{\partial}{\partial a} \| \mathbf{y} - \mathbf{\hat{y}} - a \mathbf{v} \|^2 = -2 \mathbf{v}^T (\mathbf{y} - \mathbf{\hat{y}}) + 2a \| \mathbf{v} \|^2 = 0 \Rightarrow a = \frac{\mathbf{v}^T (\mathbf{y} - \mathbf{\hat{y}})}{\| \mathbf{v} \|^2} \neq 0.
\]

Let

\[
\mathbf{\hat{y}}' = \mathbf{\hat{y}} + \frac{\mathbf{v}^T (\mathbf{y} - \mathbf{\hat{y}})}{\| \mathbf{v} \|^2} \mathbf{v},
\]

and note that \( \mathbf{\hat{y}}' \) is also in \( S \) but it is closer to \( \mathbf{y} \) contradicting the optimality of \( \mathbf{\hat{y}} \). \( \Box \)

### 4.2 Random vectors

A **random vector** is a vector of random variables. Consider the random vectors \( \mathbf{x} \) and \( \mathbf{y} \)

\[
\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.
\]

The **expected value** of \( \mathbf{x} \) is

\[
\mathbb{E} \mathbf{x} = \begin{pmatrix} \mathbb{E} x_1 \\ \vdots \\ \mathbb{E} x_m \end{pmatrix}.
\]

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

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

- \( \mathbb{E} [\mathbf{A} \mathbf{x} + \mathbf{a}] = \mathbf{A} \mathbb{E} \mathbf{x} + \mathbf{a} \)
- \( \text{Cov}(\mathbf{x}, \mathbf{y}) = \mathbb{E} [\mathbf{x} \mathbf{y}^T] - \mathbb{E} \mathbf{x} \mathbb{E} \mathbf{y}^T \)
- \( \mathbb{E} [(\mathbf{A} \mathbf{x})(\mathbf{B} \mathbf{y})^T] = \mathbf{A} \mathbb{E} [\mathbf{x} \mathbf{y}^T] \mathbf{B}^T \)
- \( \text{Cov}(\mathbf{A} \mathbf{x} + \mathbf{a}, \mathbf{B} \mathbf{y} + \mathbf{b}) = \mathbf{A} \text{Cov}(\mathbf{x}, \mathbf{y}) \mathbf{B}^T \)
- \( \text{Cov}(\mathbf{A} \mathbf{x} + \mathbf{a}) = \mathbf{A} \text{Cov}(\mathbf{x}) \mathbf{A}^T \)
- \( \text{Cov}(\mathbf{w} + \mathbf{x}, \mathbf{y} + \mathbf{z}) = \text{Cov}(\mathbf{w}, \mathbf{y}) + \text{Cov}(\mathbf{w}, \mathbf{z}) + \text{Cov}(\mathbf{x}, \mathbf{y}) + \text{Cov}(\mathbf{x}, \mathbf{z}) \)
4.3 Gaussian Random Vectors (Joint Gaussian Distribution)

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

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

**Definition 67.** 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 68.** For example if $\begin{pmatrix} x \\ y \end{pmatrix}$ is a Gaussian vector, then $z = 2x + 3y$ is Gaussian. Furthermore,

$$\mathbb{E}[z] = 2\mathbb{E}[x] + 3\mathbb{E}[y],$$
$$\text{Cov}(z) = \text{Cov}(2x + 3y, 2x + 3y) = 4\text{Cov}(x, x) + 12\text{Cov}(x, y) + 9\text{Cov}(y, y)$$
$$= 4\text{Var}(x) + 12\text{Cov}(x, y) + 9\text{Var}(y),$$

which completely characterizes the distribution of $z$.

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

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

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

4.4 Maximum likelihood for Gaussian Random Vectors

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

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

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

Given a set of $n$ iid samples $D = \{z_1, z_2, \ldots, z_n\}$, where each $z_i$ is a $d$-dimensional vector, how can we estimate $\mu$ 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 $\mu$ and $K$, we write

$$\ell(\mu, K) = \ln p(D; \mu, K) = \sum_{i=1}^{n} \ln p(z_i; \mu, K) = \frac{n}{2} \ln |K^{-1}| - \frac{1}{2} \sum_{i=1}^{n} (z_i - \mu)^T K^{-1}(z_i - \mu),$$

where we have used the fact that $|K^{-1}| = \frac{1}{|K|}$. 
As seen in the appendix (99-Appendix.tex), for a symmetric matrix \( A \), we have \( \frac{d}{dy} (y^T A y) = 2y^T A \frac{dy}{dx} \). Hence,

\[
\frac{\partial \ell}{\partial \mu} = -\frac{1}{2} \sum_{i=1}^{n} 2(z_i - \mu)^T K^{-1} (-I) = \sum_{i=1}^{n} (z_i - \mu)^T K^{-1}.
\]

Setting this equal to zero yields

\[
\hat{\mu}_{ML} = \bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i.
\]

**Exercise 69.** Using the facts

\[
\frac{\partial}{\partial A} x^T A x = x^T x, \quad \frac{\partial}{\partial A} \ln |A| = A^{-T}
\]

prove that

\[
\hat{K}_{ML} = \frac{1}{n} \sum_{i=1}^{n} (z_i - \bar{z})(z_i - \bar{z})^T
\]