CHAPTER 1

Introduction to multivariate distributions-2

1.1

1.2 Distributions

1.2.1 Preliminaries

To start with, the following concepts in matrix algebra will be used extensively throughout our discussions.

1. A matrix $A$ is non-negative definite if $\mathbf{x}^t A \mathbf{x} \geq 0$ for all $\mathbf{x}$ vectors. Further, it is positive definite if the equality holds only if $\mathbf{x} = 0$.

2. Any such positive definite matrix can be written as $A = Q Q^t$ for some non-singular $Q$. Further, we can choose $Q$ to be a positive definite matrix as well.

3. The spectral decomposition of such an $A$ is

$$A = P \Lambda P^t$$

where $P$ is orthogonal and $\Lambda$ a set of positive eigenvalues. $Q$ can be chosen to be $P \Lambda^{1/2} P^t$, and will be denoted $P^{1/2}$. The inverse of $Q$, namely $P \Lambda^{-1/2} P^t$, will be denoted by $P^{-1/2}$. 
As we discuss the multivariate distributions, we will write $X = (X_1, \ldots, X_p)^t$ as a random vector with components as random variables. The distribution of $X$ is defined as a non-negative function $F$ defined on $\mathbb{R}^p$ such that $F(x) = P(X \leq x)$, where the inequality is component-wise. A random vector $X$ is known to have a density $f$ if we have

$$F(x) = \int_{u \leq x} f(u) du$$

We define the expectation $\mu = \mathbb{E}(X)$ as a vector of expectations for each component. In the presence of a density $f$, we can write $\mu = \int u f(u) du$. (For the general case, $\mu = \int u dF(u)$.) The population variance, denoted $\Sigma = V(X)$ is defined as the matrix $\mathbb{E}[(X - \mu)(X - \mu)^t]$. The correlation matrix $\Omega$ will be defined accordingly.

**Marginal and conditional distributions**

Let $X$ be a random vector with $p$ components. Define $X_1$ as a $q$-vector consisting of a proper subset of the components of $X$, for example, say the first $q$ components. Then the marginal distribution of $X_1$ can be obtained by integrating the density $f$ over the remaining $p - q$ components. Namely,

$$f(x_1) = \int f(x) \, dx_2$$

where $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

Writing $X_2$ as the $p - q$ vector consisting of the rest of the variables, the conditional distribution of $X_2$ given $X_1$ is

$$f(x_2|x_1) = \frac{f(x)}{f(x_1)}$$

**Characteristic functions**

Analogous to univariate characteristic functions, the multivariate c.f. is defined as follows: the c.f. for any random $p$-vector $X$ is defined as a complex function $\phi : \mathbb{R}^p \rightarrow \mathbb{C}$ such that,

$$\phi(a) = \mathbb{E}(e^{ia^tX})$$
The c.f. completely specifies the distribution of any random vector and if the c.f. is absolutely integrable, we can derive the p.d.f. of the random vector using the inversion formula

\[ f(x) = \frac{1}{(2\pi)^p} \int e^{-ia^T x} \phi(a) d(a) \]

**Linear combinations:**

As before, we define a new random vector \( Y \) as a linear combination of the random vector \( X \) as, \( Y = AX + b \) where \( A \) is any \( r \times p \) matrix and \( b \) is a \( r \)-vector. Then,

\[ E(Y) = AE(X) + b, \text{ and } Var(Y) = AVar(X)A^T. \]

Moreover, the distribution of a random vector \( X \) is completely determined by the set of all univariate distributions of linear combinations \( a^T X \) where \( a \) ranges through all fixed \( p \)-vectors. This is known as Cramer-Wold theorem and easily follows using c.f.

### 1.2.2 Normal distributions

A generalization of the normal distribution to multi dimensions plays an important role in the analysis. Most of the standard multivariate techniques stem from the assumption that the underlying data comes from a multivariate normal distribution. In reality, a normal density often provides a useful approximation to the true population distribution. Further, this distribution is mathematically nice, and easily tractable. Further, because of the central limit theorem, many multivariate statistics display approximate normality, even in reality.

Now we have reached the state to define the multivariate normal distribution. It is a distribution with expectation \( \mu \) and variance \( \Sigma \), whose density is given by

\[ f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right) \]
It is easy to see that \( f \) is a density, and the corresponding expectation and variance are \( \mu \) and \( \Sigma \). The following results can be proved similarly. We assume that \( X \) has a \( p \)-dimensional normal distribution with mean \( \mu \) and variance \( \Sigma \), written \( X \sim N_p(\mu, \Sigma) \).

**Singular normal distributions**: A variation of the classical normal distribution is the singular normal distribution where the population variance matrix \( \Sigma \) need not be singular. We say that \( X \sim N_p(\mu, \Sigma) \) where \( \Sigma \) is a singular non-negative matrix of rank \( r \) if the density of \( X \) can be written as,

\[
    f(x) = \frac{1}{(2\pi)^{r/2}(\lambda_1 \cdots \lambda_r)^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^t \Sigma^{-1}(x - \mu)\right)
\]

where \( \lambda_1, \ldots, \lambda_r \) are the positive eigenvalues of \( \Sigma \) and \( \Sigma^{-1} \) indicates any generalized inverse of \( \Sigma \).

- Let \( y = a^t X \). then, \( y \) has an univariate normal distribution with mean \( a^t \mu \) and variance \( a^t \Sigma a \). Further, if \( y = a^t X \) is distributed as \( N(a^t \mu, a^t \Sigma a) \) for every \( a \), then \( X \) follows \( N_p(\mu, \Sigma) \).

- Let \( Y = AX + b \), where \( A \) is \( r \times p \), and \( b \) is a fixed \( r \)-vector, then \( Y \sim N_r(A\mu + b, A\Sigma A^t) \).

**Remark**: If \( r \leq p \) and \( A \) has full row-rank, then \( A\Sigma A^t \) is p.d., and hence the result is valid in a straightforward way. However, if \( r > p \), or \( A \) has dependent rows, then the variance matrix \( A\Sigma A^t \) is singular, and the results hold with the singular normal distribution.

- In case of normality, independence is equivalent to uncorrelated-ness.

- c.f. of \( X \) is given by

\[
    \phi(a) = e^{ia^t \mu - \frac{1}{2}a^t \Sigma a}
\]

- The Mahalanobis’ distance \( U = (X - \mu)^t \Sigma^{-1}(X - \mu) \sim \chi^2_p \).

**1.2.3 Other multivariate distributions**

**Multinomial distributions**: As a well known generalization of binomial distribution, this discrete distribution have parameters \( n \) and \( p_1, \ldots, p_k \) such
that $p_i > 0$ and $\sum_{i=1}^{k} p_i \leq 1$. Then the vector $(n_1, \ldots, n_k)$ has probability mass function
\[
p(n_1, \ldots, n_k) = \frac{n!}{n_1! \cdots n_k! n_{k+1}!} n_1^{p_1} \cdots n_k^{p_k} n_{k+1}^{p_{k+1}}
\]
where $n_{k+1} = n - \sum_{i=1}^{k} n_i$ and $p_{k+1} = 1 - \sum_{i=1}^{k} p_i$.

**Dirichlet distribution** : This is a generalization of the Beta distribution. For $(x_1, x_2, \ldots, x_p)$ such that $x_i \geq 0$ and $\sum_{i=1}^{p} x_i \leq 1$, the density is given by
\[
f(x_1, x_2, \ldots, x_k) = \frac{\Gamma(\sum_{i=1}^{k+1} \alpha_i)}{\prod_{i=1}^{k+1} \Gamma(\alpha_i)} \prod_{i=1}^{k+1} x_i^{\alpha_i-1}
\]
where $x_{k+1} = 1 - \sum_{i=1}^{k} x_i$ This is used in Bayesian analysis as a prior for multinomial parameters.

**Spherical distribution** : A family of distribution characterized by the p.d.f. $f$ satisfying the identity
\[f(x) = h(x^tx)\]
for some positive function $h$. A normal distribution is spherical (or symmetric) if it has mean 0 and variance $aI$ for some number $a$. Conversely, any spherical distribution has expectation 0 and variance $aI$ for some $a > 0$, provide those moments exist.

**Wishart distribution** : Let $X$ be a data matrix with mean 0 and variance $\Sigma$. Then, the matrix $X'X$, the sample products matrix has a distribution known as Wishart matrix with parameters $p$, $n$, $\Sigma$. It is a distribution defined on the space of non-negative definite matrices. We do not go in detail discussion though. However, the sample variance matrix can be shown to have a Wishart distribution with parameters $p$, $n-1$ and $\Sigma$ (except for the multiple).

### 1.2.4 Random sample

A random sample $X_1, \ldots, X_n$ are independent random vectors coming from a common multivariate distribution $f(x, \theta)$. The data matrix $X$ is defined as
before, i.e.

\[ X = \begin{bmatrix} X_1^t \\ \vdots \\ X_n^t \end{bmatrix} \]

The sample mean and the sample variance, as before, are defined as,

\[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} X'1 \]

and the sample variance-covariance matrix is

\[ S_x = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})' = \frac{1}{n-1} X'(I - \frac{1}{n} 11')X \]

The generalized variance of the sample is defined as \(|S_x|\), the determinant of the sample covariance matrix. This variance measures the overall spread of the data. In particular, if the data cloud, though spreading over a large area, is particularly thin in any particular orientation, the generalized variance will be small as well. Sometimes, for the scale-invariance property, the generalized variance of the standardized variables, namely \(|R|\) (where \(R\) is the correlation matrix) is taken into consideration. The quantities \(|S|\) and \(|R|\) are connected by the relationship \(|S| = \prod_{i=1}^{p} s_{ii}|R|\).

To facilitate our notations, we introduce vectorization and Kronecker’s product as follows. Define,

\[ vec(X) = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \]

a long \(np \times 1\) vector. To write the mean and variance of this vector, the Kronecker’s product between matrices \(A_{m,n}\) and \(B_{pq}\) is defined as the \(mp \times nq\) matrix

\[ A \otimes B = \begin{pmatrix} a_{11}B & \ldots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \ldots & a_{mn}B \end{pmatrix} \]

Then,

\[ \mathbb{E}(vec(X)) = 1 \otimes \mu \]
and

\[ V(vec(X)) = I \otimes \Sigma \]

The following results can be proved

- \( vec(X) \) has a normal distribution with the mean and variance given above.

- \( \bar{X} \) is normal with mean \( \mu \) and variance \( 1/n \Sigma \).

- \( \mathbb{E}(S_x) = \Sigma \).

- Central limit theorem :
  
  Let the sample size \( n \) grow to infinity, but \( p \) remain fixed, then

  \[ \sqrt{n}(\bar{X} - \mu) \Rightarrow N(0, \Sigma) \]

  where \( \Rightarrow \) means converges in distribution.