II Lecture
The Multivariate Normal Distribution

2.1. Standard facts about multivariate distributions

2.1.1. Random samples in multivariate analysis

In order to study the sampling variability of statistics like $\bar{x}$ and S_n that we introduced in I lecture , with the ultimate goal of making inferences, one needs to make some assumptions about the random variables whose values constitute the data set $X \in \mathscr{M}_{p,n}$ in (1.1). Suppose the data has not been observed yet but we *intend* to collect n sets of measurements on p variables. Since the actual observations can not be predicted before the measurements are made, we treat them as random variables. Each set of p measurements can be considered as a realization of p-dimensional *random vector* and we have n independent realizations of such random vectors $\mathbf{X_i}, i = 1, 2, \ldots, n$, so we have the *random matrix* $\mathbf{X} \in \mathscr{M}_{\mathbf{p,n}}$:

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & .. & X_{1j} & .. & X_{1n} \\ X_{21} & X_{22} & .. & X_{2j} & .. & X_{2n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ X_{i1} & X_{i2} & .. & X_{ij} & .. & X_{in} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ X_{p1} & X_{p2} & .. & X_{pj} & .. & X_{pn} \end{pmatrix} = [\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}] \tag{2.1}$$

The vectors $\mathbf{X_i}, i = 1, 2, \ldots, n$ are considered as independent observations of a p-dimensional random vector. We start discussing the distribution of such a vector.

2.1.2. Joint, marginal, conditional distributions

Let a random vector $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_p \end{pmatrix} \in R^p$, $p \geq 2$ has p different components each of which is a random variable with a cumulative distribution function *(cdf)* $F_{X_i}(x_i), i = 1, 2, ..,$p. Each of the functions $F_{X_i}(.)$ is called a *marginal distribution*. The *joint cdf* of the random vector $\mathbf{X}$ is

$$F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_p \leq x_p) = F_{\mathbf{X}}(x_1, x_2, .., x_p)$$

In case of a *discrete* vector of observations $\mathbf{X}$ the *probability mass function* is defined as

$$P_{\mathbf{X}}(\mathbf{x}) = P(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$$

If a *density* $f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(x_1, x_2, .., x_p)$ exists such that

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_1} .. \int_{-\infty}^{x_p} f_{\mathbf{X}}(\mathbf{t}) dt_1 \dots dt_p \tag{2.2}$$

then $\mathbf{X}$ is a *continuous* random vector with a joint density function of p arguments $f_{\mathbf{X}}(\mathbf{x})$. From (2.2) we see that in this case $f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^p F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \partial x_2 .. \partial x_p}$ holds. In case $\mathbf{X}$ has p independent components then

$$F_{\mathbf{X}}(\mathbf{x}) = F_{X_1}(x_1) F_{X_2}(x_2) \dots F_{X_p}(x_p) \tag{2.3}$$

holds and, equivalently, also

$$P_{\mathbf{X}}(\mathbf{x}) = P_{X_1}(x_1) P_{X_2}(x_2) \dots P_{X_p}(x_p), f_{\mathbf{X}}(\mathbf{x}) = f_{X_1}(x_1) f_{X_2}(x_2) f_{X_p}(x_p) \tag{2.4}$$

holds.

The *marginal cdf of the first* $k < p$ components of the vector $\mathbf{X}$ is defined in a natural way as follows:

$$P(X_1 \le x_1, X_2 \le x_2, .., X_k \le x_k) = P(X_1 \le x_1, X_2 \le x_2, \dots, X_k \le x_k, X_{k+1} \le \infty, ..., X_p \le \infty)$$

$$= F_{\mathbf{X}}(x_1, x_2, .., x_k, \infty, \infty, .., \infty) \tag{2.5}$$

The *marginal density* of the first k components can be obtained by partial differentiation in (2.5) and we arrive at

$$\int_{-\infty}^{\infty} .. \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, x_2, .., x_p) dx_{k+1} \dots dx_p$$

For **any** other subset of $k < p$ components of the vector $\mathbf{X}$, their marginal cdf and density can be obtained along the same lines.

The *conditional density* $\mathbf{X}$ **when** $X_{r+1} = x_{r+1}, \dots, X_p = x_p$ is defined by

$$f_{(X_1,...,X_r|X_{r+1},...,X_p)}(x_1, \dots, x_r | x_{r+1}, \dots, x_p) = \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{X_{r+1},...,X_p}(x_{r+1}, \dots, x_p)} \tag{2.6}$$

The above conditional density is interpreted as the joint density of $X_1, \dots, X_r$ when $X_{r+1} = x_{r+1}, \dots, X_p = x_p$ and is only defined when $f_{X_{r+1},...,X_p}(x_{r+1}, \dots, x_p) \neq 0$.

We note that in case of mutual independence the p components, all conditional distributions do **not** depend on the conditions and the factorizations

$$F_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{p} F_{X_i}(x_i), f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{p} f_{X_i}(x_i)$$

hold.

2.1.3. Moments

Given the density $f_{\mathbf{X}}(\mathbf{x})$ of the random vector $\mathbf{X}$ the joint moments of order $s_1, s_2 \ldots, s_p$ are defined, in analogy to the univariate case, as

$$E(X_1^{s_1} \ldots X_p^{s_p}) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} x_1^{s_1} \ldots x_p^{s_p} f_{\mathbf{X}}(x_1, \ldots, x_p) dx_1 \ldots dx_p \tag{2.7}$$

Note that if some of the s_i in (2.7) are equal to zero then in effect we are calculating the joint moment of a subset of the p random variables.

2.1.4. Density transformation formula

Assume , the p existing random variables $X_1, X_2 \ldots, X_p$ with given density $f_{\mathbf{X}}(\mathbf{x})$ have been transformed by a smooth (i.e. differentiable) one-to-one transformation into p new random variables $Y_1, Y_2 \ldots, Y_p$, i.e. a new random vector $\mathbf{Y} \in \mathbf{R^p}$ has been created by calculating

$$Y_i = y_i(X_1, X_2 \ldots, X_p), i = 1, 2, \ldots, p \tag{2.8}$$

The question is how to calculate the density $g_{\mathbf{Y}}(\mathbf{y})$ of $\mathbf{Y}$ by knowing the transformation functions $y_i(X_1, X_2 \ldots, X_p), i = 1, 2, \ldots, p$ and the density $f_{\mathbf{X}}(\mathbf{x})$ of the original random vector. Naturally, since the transformation (2.8) is assumed to be one-to-one, its inverse transformation $X_i = x_i(Y_1, Y_2 \ldots, Y_p), i = 1, 2, \ldots, p$ also exists and then the following density transformation formula applies:

$$g_{\mathbf{Y}}(y_1, \ldots, y_p) = f_{\mathbf{X}}[x_1(y_1, \ldots, y_p), \ldots, x_p(y_1, \ldots, y_p)]|J(y_1, \ldots, y_p)| \tag{2.9}$$

where $J(y_1, \ldots, y_p)$ is the *Jacobian* of the transformation:

$$J(y_1 \ldots, y_p) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_p} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial x_p}{\partial y_1} & \frac{\partial x_p}{\partial y_2} & \cdots & \frac{\partial x_p}{\partial y_p} \end{vmatrix} \tag{2.10}$$

Note that in (2.9) the *absolute value* of the Jacobian is substituted.

2.1.5. Characteristic and moment generating functions

The characteristic function *(cf)* $\varphi(\mathbf{t})$ of the random vector $\mathbf{X} \in \mathbf{R^p}$ is a function of a p-dimensional argument. For any real vector $\mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ . \\ . \\ . \\ t_p \end{pmatrix} \in R^p$ the above *characteristic function* is defined as $\varphi_{\mathbf{X}}(\mathbf{t}) = E(e^{i\mathbf{t}'\mathbf{X}})$ where $i = \sqrt{-1}$. Note that the cf always exists since $|\varphi_{\mathbf{X}}(\mathbf{t})| \leq E(|e^{i\mathbf{t}'\mathbf{X}}|) = 1 < \infty$. Maybe more simple (since it does not involve complex numbers) is the notion of *moment generating function (mgf)*. It is defined as $M_{\mathbf{X}}(\mathbf{t}) = E(e^{\mathbf{t}'\mathbf{X}})$. Note however that in some cases the mgf may not exist for values of $\mathbf{t}$ further away from the zero vector.

Characteristic functions are in one-to-one correspondence with distributions and this is the reason to use them as a machinery to operate with in cases where direct operation

with the distribution is not very convenient. In fact, when the density exists, under mild conditions the following simple inversion formula holds for a one-dimensional random variable:

$$f_{\mathbf{X}}(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itx}\varphi(t)dt.$$

This formula can also be generalised for random vectors.

One important property of *cf* is the following:

If the *cf* $\varphi_{\mathbf{X}}(\mathbf{t})$ of the random vector $\mathbf{X} \in \mathbf{R^p}$ is given and $\mathbf{Y} = \mathbf{AX} + \mathbf{b}, \mathbf{b} \in \mathbf{R^q}, \mathbf{A} \in \mathscr{M}_{\mathbf{q,p}}$ is a linear transformation of $\mathbf{X} \in \mathbf{R^p}$ into a new random vector $\mathbf{Y} \in \mathbf{R^q}$ then it holds for all $\mathbf{s} \in \mathbf{R^q}$ that

$$\varphi_{\mathbf{Y}}(\mathbf{s}) = e^{i\mathbf{s}'\mathbf{b}}\varphi_{\mathbf{X}}(\mathbf{A}'\mathbf{s})$$

Proof: at lectures.

2.2. Multivariate Normal Distribution

2.2.1 Definition

The multivariate normal density is a generalization of the univariate normal for $p \geq 2$ dimensions. Looking at the term $(\frac{x-\mu}{\sigma})^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$ in the exponent of the well known formula

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-[(x-\mu)/\sigma]^2/2}, \quad -\infty < x < \infty \tag{2.11}$$

for the univariate density function, a natural way to generalize this term in higher dimensions is to *replace* it by $(\mathbf{x}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu)$. Here $\mu = E\mathbf{X} \in \mathbf{R^p}$ is the expected value of the random vector $\mathbf{X} \in \mathbf{R^p}$ and the matrix

$$\mathbf{\Sigma} = E[(\mathbf{X}-\mu)(\mathbf{X}-\mu)'] = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \dots & \dots & \dots & \dots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix} \in \mathscr{M}_{\mathbf{p,p}}$$

is the *covariance matrix*. Note that on the diagonals of $\mathbf{\Sigma}$ we get the *variances* of each of the p random variables whereas $\sigma_{ij} = E[(X_i - E(X_i))(X_j - E(X_j))], i \neq j$ are the *covariances* between the ith and jth random variable. Sometimes, we will also denote σ_{ii} by σ_i^2.

Of course, the above replacement would only make sense if $\mathbf{\Sigma}$ was positive definite. In general, however, we can only claim that $\mathbf{\Sigma}$ is (as any covariance matrix) non-negative definite (try to prove this claim e.g. using Exercise 1 from Lecture 1 or some other argument). If $\mathbf{\Sigma}$ was positive definite then the density of the random vector $\mathbf{X}$ can be written as

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-(\mathbf{x}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu)/\mathbf{2}}, -\infty < x_i < \infty, i = 1, 2 \dots, p \tag{2.12}$$

It can be directly checked that the random vector $\mathbf{X} \in \mathbf{R^p}$ has $E\mathbf{X} = \mu$ and

$$E[(\mathbf{X}-\mu)(\mathbf{X}-\mu)'] = \mathbf{\Sigma}.$$

Since the density is uniquely defined by the *mean vector and the covariance matrix* we will denote it by $N_p(\mu, \mathbf{\Sigma})$.

In these notes, however, we will introduce the multivariate normal distribution not through its density formula but through more general reasoning that also allows to cover the case of singular $\mathbf{\Sigma}$. We will utilize the famous **Cramer-Wold argument** according to which the distribution of a p-dimensional random vector $\mathbf{X}$ is completely characterised by the one dimensional distributions of **all** linear transformations $\mathbf{T'X}, \mathbf{T} \in \mathbf{R^p}$. Indeed, if we consider $E[e^{(it\mathbf{T'X})}]$ (which is assumed to be known for every $t \in R^1, \mathbf{T} \in \mathbf{R^p}$) then we see that by substituting $t = 1$ we can get $E[e^{(i\mathbf{T'X})}]$ which is the cf of the vector $\mathbf{X}$ (and the latter uniquely specifies the distribution of $\mathbf{X}$). Hence the following definition will be adopted here:

Definition 2.2.1. The random vector $\mathbf{X} \in \mathbf{R^p}$ has a multivariate normal distribution if and only if (iff) any linear transformation $\mathbf{T'X}, \mathbf{T} \in \mathbf{R^p}$ has a univariate normal distribution.

Theorem 2.2.1. Let for a random vector $\mathbf{X} \in \mathbf{R^p}$ with a normal distribution according to Definition 2.2.1 we have $E(\mathbf{X}) = \mu$ and $D(\mathbf{X}) = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)'] = \mathbf{\Sigma}$. Then:

i) for any fixed $\mathbf{T} \in \mathbf{R^p}$, $\mathbf{T'X} \sim \mathbf{N}(\mathbf{T'}\mu, \mathbf{T'\Sigma T})$ i.e. $\mathbf{T'X}$ has an one dimensional normal distribution with expected value $\mathbf{T'}\mu$ and variance $\mathbf{T'\Sigma T}$.

ii) The *cf* of $\mathbf{X} \in \mathbf{R^p}$ is

$$\varphi_{\mathbf{X}}(\mathbf{T}) = e^{(i\mathbf{T'}\mu - \frac{1}{2}\mathbf{T'\Sigma T})} \tag{2.13}$$

Proof: Part i) is obvious. For part ii) we observe first the well known fact that the cf of the standard univariate normal random variable Z is $e^{-t^2/2}$. Since any $U \sim N_1(\mu_1, \sigma_1^2)$ has a distribution that coincides with the distribution of $\mu_1 + \sigma_1 Z$ we have:

$$\varphi_U(t) = e^{it\mu_1}\varphi_{\sigma_1 Z}(t) = e^{it\mu_1}E(e^{it\sigma_1 Z}) = e^{it\mu_1}\varphi_Z(t\sigma_1) = e^{(it\mu_1 - \frac{1}{2}t^2\sigma_1^2)}$$

But then, for the *univariate random variable* $\mathbf{T'X} \sim \mathbf{N}_1(\mathbf{T'}\mu, \mathbf{T'\Sigma T})$ we would have as a characteristic function $\varphi_{\mathbf{T'X}}(t) = e^{it\mathbf{T'}\mu - \frac{1}{2}\mathbf{t^2 T'\Sigma T}}$. Substituting $t = 1$ in the latter formula we find that

$$\varphi_{\mathbf{X}}(\mathbf{T}) = e^{(i\mathbf{T'}\mu - \frac{1}{2}\mathbf{T'\Sigma T})}$$

As an upshot, we see that given the expected value vector μ and the covariance matrix Σ we can use the *cf* formula (2.13) rather than the density formula (2.12) to define the p dimensional multivariate normal distribution. The advantage of the former in comparison to the latter is that in (2.13) only Σ is used, i.e. this definition makes also sense in cases of singular (i.e. non-invertible) Σ. We still want to know that in case of non-singular Σ the more general definition would give raise to the density (2.12). This is the content of the next theorem.

Theorem 2.2.2. Assume the matrix Σ in (2.13) is nonsingular. Then the density of the random vector $\mathbf{X} \in \mathbf{R^p}$ with cf as in (2.13) is given by (2.12).

Proof: Consider the vector $\mathbf{Y} \in \mathbf{R^p}$ such that $\mathbf{Y} = \mathbf{\Sigma^{-\frac{1}{2}}}(\mathbf{X} - \mu)$ (compare 4.2.5 in I Lecture). Since obviously $E(\mathbf{Y}) = \mathbf{0}$ and $D(\mathbf{Y}) = E(\mathbf{YY'}) = \Sigma^{-\frac{1}{2}}E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)']\Sigma^{-\frac{1}{2}} = I_p$ holds we can substitute to get the cf of $\mathbf{Y} \in \mathbf{R^p}$: $\varphi_{\mathbf{Y}}(\mathbf{T}) = e^{-\frac{1}{2}\sum_1^p t_i^2}$. But the latter can be seen directly to be the characteristic function of the vector of p independent standard normal variables. Hence, from the relation $\mathbf{Y} = \mathbf{\Sigma^{-\frac{1}{2}}}(\mathbf{X} - \mu)$ we can also conclude that $\mathbf{X} = \mu + \mathbf{\Sigma^{\frac{1}{2}}Y}$ where the density $f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{p/2}}e^{-\frac{1}{2}\sum_{i=1}^p y_i^2}$. With other

words, $\mathbf{X}$ is a *linear transformation* of $\mathbf{Y}$ where the density of $\mathbf{Y}$ is *known*. We can therefore apply the density transformation approach (Section 2.1.4. of this lecture) to obtain: $f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{Y}}(\Sigma^{-\frac{1}{2}}(x-\mu))|J(x_1,\ldots,x_p)|$. It is easy to see (because of the linearity of the transformation) that $|J(x_1,\ldots,x_p)| = |\Sigma^{-\frac{1}{2}}| = |\Sigma^{\frac{1}{2}}|^{-1}$. Taking into account that $\sum_{i=1}^{p} y_i^2 = \mathbf{y}'\mathbf{y} = (\mathbf{x}-\mu)'\mathbf{\Sigma^{-\frac{1}{2}}}\mathbf{\Sigma^{-\frac{1}{2}}}(\mathbf{x}-\mu) = (\mathbf{x}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu)$ we finally arrive at the density formula (2.12) for $f_{\mathbf{X}}(\mathbf{x})$.

2.2.2. Properties of multivariate normal

The following *properties* of multivariate normal can be easily derived using the machinery developed so far:

Property 1. If $\mathbf{\Sigma} = \mathbf{D}(\mathbf{X}) = \mathbf{\Lambda}$ is a diagonal matrix then the p components of $\mathbf{X}$ are independent.

(Indeed , in this case $\varphi_{\mathbf{X}}(\mathbf{T}) = e^{i\sum_{j=1}^{p} t_j\mu_j - \frac{1}{2}t_j^2\sigma_j^2}$ which can be seen to be the *cf* of the vector of p independent components each distributed according to $N(\mu_j, \sigma_j^2), j = 1,\ldots,p$).

The above property can be paraphrased as "for a multivariate normal, if its components are uncorrelated they are also independent". On the other hand, it is well known that *always, i.e. not only for normal* from the fact that certain components are independent we can conclude that they are also uncorrelated. Therefore , for the **multivariate normal distribution** we can conclude that its components are **independent if and only if they are uncorrelated!**

Property 2. If $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$ and $\mathbf{C} \in \mathscr{M}_{\mathbf{q,p}}$ is an arbitrary matrix of real numbers then

$$\mathbf{Y} = \mathbf{CX} \sim \mathbf{N_q}(\mathbf{C}\mu, \mathbf{C\Sigma C}')$$

To prove this property note that (see 2.1.5) for any $\mathbf{s} \in \mathbf{R^q}$ we have:

$$\varphi_{\mathbf{Y}}(\mathbf{s}) = \varphi_{\mathbf{X}}(\mathbf{C}'\mathbf{s}) = e^{i\mathbf{s}'\mathbf{C}\mu - \frac{1}{2}\mathbf{s}'\mathbf{C\Sigma C}'\mathbf{s}}$$

which means that $\mathbf{Y} = \mathbf{CX} \sim \mathbf{N_q}(\mathbf{C}\mu, \mathbf{C\Sigma C}')$.

Note also that if it happens that the rank of $\mathbf{C}$ is full and if $rk(\mathbf{\Sigma}) = p$ then the rank of $\mathbf{C\Sigma C}'$ is also full, i.e. the distribution of $\mathbf{Y}$ would not be degenerate in this case.

Property 3. (This is a finer version of property 1). Assume the vector $\mathbf{X} \in \mathbf{R^p}$ is divided into subvectors $\mathbf{X} = \begin{pmatrix} X_{(1)} \\ X_{(2)} \end{pmatrix}$ and according to this subdivision the vector means are $\mu = \begin{pmatrix} \mu_{(1)} \\ \mu_{(2)} \end{pmatrix}$ and the covariance matrix $\mathbf{\Sigma}$ has been subdivided into $\mathbf{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Then the vectors $\mathbf{X_{(1)}}$ and $\mathbf{X_{(2)}}$ are independent iff $\mathbf{\Sigma_{12}} = \mathbf{0}$.

Proof: (Exercise (see lecture)).

Property 4. Let the vector $\mathbf{X} \in \mathbf{R^p}$ is divided into subvectors $\mathbf{X} = \begin{pmatrix} X_{(1)} \\ X_{(2)} \end{pmatrix}$, $X_{(1)} \in R^r, r < p, X_{(2)} \in R^{p-r}$ and according to this subdivision the vector means are $\mu = \begin{pmatrix} \mu_{(1)} \\ \mu_{(2)} \end{pmatrix}$ and the covariance matrix $\mathbf{\Sigma}$ has been subdivided into $\mathbf{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Assume for simplicity that the rank of $\mathbf{\Sigma_{22}}$ is full. Then the conditional density of $\mathbf{X_{(1)}}$ given that

$\mathbf{X_{(2)}} = \mathbf{x_{(2)}}$ is

$$N_r(\mu_{(1)} + \mathbf{\Sigma_{12}\Sigma_{22}^{-1}}(\mathbf{x_{(2)}} - \mu_{(2)}), \mathbf{\Sigma_{11}} - \mathbf{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}})$$

Proof: Perhaps the easiest way to proceed is the following. Note that the expression $\mu_{(1)} + \mathbf{\Sigma_{12}\Sigma_{22}^{-1}}(\mathbf{x_{(2)}} - \mu_{(2)})$ (for which we want to show that it equals the conditional mean), is a function of $\mathbf{x_{(2)}}$. Denote is as $g(\mathbf{x_{(2)}})$ for short. Let us construct the random vectors $\mathbf{Z} = \mathbf{X_{(1)}} - \mathbf{g(X_{(2)})}$ and $\mathbf{Y} = \mathbf{X_{(2)}} - \mu_{(2)}$. Obviously $E\mathbf{Z} = \mathbf{0}$ and $E\mathbf{Y} = \mathbf{0}$ holds. The vector $\begin{pmatrix} \mathbf{Z} \\ \mathbf{Y} \end{pmatrix}$ is a *linear transformation of a normal vector* ($\begin{pmatrix} \mathbf{Z} \\ \mathbf{Y} \end{pmatrix} = \mathbf{A}(\mathbf{X} - \mu)$, $\mathbf{A} = \begin{pmatrix} \mathbf{I_r} & \mathbf{-\Sigma_{12}\Sigma_{22}^{-1}} \\ \mathbf{0} & \mathbf{I_{p-r}} \end{pmatrix}$) and hence, its distribution is normal (Property 2). Calculating the covariance matrix of the vector $\begin{pmatrix} \mathbf{Z} \\ \mathbf{Y} \end{pmatrix}$ we find that

$$\mathbf{Cov}\begin{pmatrix} \mathbf{Z} \\ \mathbf{Y} \end{pmatrix} = \mathbf{A\Sigma A'} =$$

after a simple exercise in block multiplication of matrices=

$$\begin{pmatrix} \mathbf{\Sigma_{11}} - \mathbf{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma_{22}} \end{pmatrix}$$

Hence the two vectors $\mathbf{Z}$ and $\mathbf{Y}$ are uncorrelated normal vectors and therefore are independent (Property 3). But $\mathbf{Y}$ is a linear transformation of $\mathbf{X_{(2)}}$ and this means that $\mathbf{Z}$ and $\mathbf{X_{(2)}}$ are independent. Hence the conditional density of $\mathbf{Z}$ given $\mathbf{X_{(2)}} = \mathbf{x_{(2)}}$ will **not depend on** $\mathbf{x_{(2)}}$ and coincides with the unconditional density of $\mathbf{Z}$. This means, it is normal with zero mean vector and its covariance matrix is

$$Cov(\mathbf{Z}) = \mathbf{\Sigma_{11}} - \mathbf{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} = \mathbf{\Sigma_{1|2}}$$

Hence we can state that $\mathbf{X_{(1)}} - \mathbf{g(x_{(2)})} \sim \mathbf{N(O, \Sigma_{1|2})}$ and correspondingly, the conditional distribution of $\mathbf{X_{(1)}}$ given that $\mathbf{X_{(2)}} = \mathbf{x_{(2)}}$ is

$$N_r(\mu_{(1)} + \mathbf{\Sigma_{12}\Sigma_{22}^{-1}}(\mathbf{x_{(2)}} - \mu_{(2)}), \mathbf{\Sigma_{11}} - \mathbf{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}) \tag{2.14}$$

Exercise 2. As an immediate consequence of Property 4 we see that if $p = 2, r = 1$ then for a two-dimensional normal vector $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix})$ its conditional density $f(x_1|x_2)$ is $N(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2})$

As an exercise, try to derive the above result by direct calculations starting from the joint density $f(x_1, x_2)$, going over to the marginal $f(x_2)$ by integration and finally getting $f(x_1|x_2) = \frac{f(x_1,x_2)}{f(x_2)}$.

Property 5 If $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$ and $\mathbf{\Sigma}$ is nonsingular then $(\mathbf{x} - \mu)'\mathbf{\Sigma^{-1}}(\mathbf{x} - \mu) \sim \chi_p^2$ where χ_p^2 denotes the chi-square distribution with p degrees of freedom.

Proof: It suffices to use the fact that (see also Theorem 2.2.2) the vector $\mathbf{Y} \in \mathbf{R^p}$: $\mathbf{Y} = \mathbf{\Sigma^{-\frac{1}{2}}}(\mathbf{X} - \mu) \sim N(\mathbf{0}, \mathbf{I_p})$ i.e. it has p independent standard normal components. Then

$$(\mathbf{x} - \mu)'\mathbf{\Sigma^{-1}}(\mathbf{x} - \mu) = \mathbf{Y'Y} = \sum_{i=1}^{p} Y_i^2 \sim \chi_p^2$$

according to the definition of χ^2_p as a distribution of the sum of squares of p independent standard normals.

Finally,one more interpretation of the result in Property 4 will be given. Assume we want, as is a typical situation in statistics, to predict a random variable Y that is correlated with some p random variables (predictors) $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ X_p \end{pmatrix}$. Trying to find the *best predictor* of Y we would like to minimize the expected value $E(Y - g(\mathbf{X}))^2$ over all possible choices of the function g such that $Eg(\mathbf{X})^2 < \infty$. A little careful work and use of basic properties of conditional expectations leads us (see lecture) to the conclusion that the optimal solution to the above minimization problem is $g^*(\mathbf{X}) = E(Y|\mathbf{X})$. This optimal solution is also called the *regression function*. Thus given a particular realization $\mathbf{x}$ of the random vector $\mathbf{X}$ the regression function is just the conditional expected value of Y given $\mathbf{X} = \mathbf{x}$.

In general, the conditional expected value may be a complicated nonlinear function of the predictors. However, if we assume *in addition* that the joint $(p+1)$-dimensional distribution of Y and $\mathbf{X}$ is **normal** then by applying Property 4 we see that given the realization $\mathbf{x}$ of $\mathbf{X}$, the best prediction of the Y value is given by $b + \sigma_0' \mathbf{C^{-1}} \mathbf{x}$ where $b = E(Y) - \sigma_0' \mathbf{C^{-1}} \mathbf{E(X)}$, $\mathbf{C}$ is the covariance matrix of the vector $\mathbf{X}$, σ_0 is the vector of Covariances of Y with $X_i, i = 1, \ldots, p$.

Indeed, we know that when the joint $(p+1)$-dimensional distribution of Y and $\mathbf{X}$ is **normal** the regression function is given by

$$E(Y) + \sigma_0' \mathbf{C^{-1}} (\mathbf{x} - \mathbf{E(X)}).$$

By introducing the notation $b = E(Y) - \sigma_0' \mathbf{C^{-1}} \mathbf{E(X)}$ we can write this as $b + \sigma_0' \mathbf{C^{-1}} \mathbf{x}$.

That is, **in case of normality, the optimal predictor of Y in the least squares sense turns out to be a very simple linear function of the predictors**. The vector $\mathbf{C^{-1}} \sigma_0 \in R^p$ is the *vector of the regression coefficients*. Substituting the optimal values we get the minimal value of the sum of squares which is equal to $V(Y) - \sigma_0' \mathbf{C^{-1}} \sigma_0$.

2.2.3. Tests for Multivariate Normality

We have seen that the assumption of multivariate normality may bring essential simplifications in analyzing data. But applying inference methods based on the multivariate normality assumption in cases where it is grossly violated may introduce serious defects in the quality of the analysis. It is therefore important to be able to check the multivariate normality assumption. Based on the properties of normal distributions discussed in this lecture, we know that all linear combinations of normal variables are normal and the contours of the multivariate normal density are ellipsoids. Therefore we can (to some extent) check the multivariate normality hypothesis by:

i) checking if the marginal distributions of each component appear to be normal (by using Q-Q plots , for example);

ii) checking if the scatterplots of pairs of observations give the elliptical appearance expected from normal populations;

iii) are there any outlying observations that should be checked for accuracy.

All this can be done by applying univariate techniques and by drawing scatterplots which are well developed in SAS. To some extent, however, there is a price to be paid for concentrating on univariate and bivariate examinations of normality . There is a need to construct a "good" overall test of multivariate normality. One of the simple and tractable ways to verify the multivariate normality assumption is by using tests based on **Mardia's multivariate skewness and kurtosis measures**. For any general multivariate distribution we define these respectively as

$$\beta_{1,p} = E[(\mathbf{Y}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{X}-\mu)]^{\mathbf{3}} \tag{2.15}$$

provided that $\mathbf{X}$ is independent of $\mathbf{Y}$ but has the same distribution and

$$\beta_{2,p} = E[(\mathbf{X}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{X}-\mu)]^{\mathbf{2}} \tag{2.16}$$

(if the expectations in (2.15) and (2.16) exist). For the $N_p(\mu,\Sigma)$ distribution: $\beta_{1,p}=0$ and $\beta_{2,p}=p(p+2)$.

(Note that when $p=1$, the quantity $\beta_{1,1}$ is the square of the skewness coefficient $\frac{E(X-\mu)^3}{\sigma^3}$ whereas $\beta_{2,1}$ coincides with the kurtosis coefficient $\frac{E(X-\mu)^4}{\sigma^4}$.)

For a sample of size n *consistent estimates* of $\beta_{1,p}$ and $\beta_{2,p}$ can be obtained as

$$\hat{\beta}_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}^3$$

$$\hat{\beta}_{2,p} = \frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n} d_i^4$$

where $g_{ij} = (\mathbf{x_i}-\bar{\mathbf{x}})'\mathbf{S_n^{-1}}(\mathbf{x_j}-\bar{\mathbf{x}})$ and $d_i = \sqrt{g_{ii}}$ is the *Mahalanobis distance*.

Both quantities $\hat{\beta}_{1,p}$ and $\hat{\beta}_{2,p}$ are nonnegative and for multivariate data, one would expect them to be around zero and $p(p+2)$, respectively. If there is a departure from spherical symmetry (that is, zero correlation and equal variance), $\hat{\beta}_{2,p}$ will be large. Both quantities can be utilized to detect departures from multivariate normality. Mardia has shown that asymptotically, $k_1 = n\hat{\beta}_{1,p}/6 \sim \chi^2_{p(p+1)(p+2)/6}$, and $k_2 = [\hat{\beta}_{2,p} - p(p+2)]/[8p(p+2)/n]^{\frac{1}{2}}$ is standard normal. Thus we can use k_1 and k_2 to test the null hypothesis of multivariate normality. If both hypotheses are accepted, the multivariate normality assumption is in reasonable agreement with the data. It also has been observed that Mardia's multivariate kurtosis can be used as a measure to detect outliers from the data that are supposedly distributed as multivariate normal. For given data set, the multivariate kurtosis can be computed using the **CALIS** procedure is SAS. The quantity k_2 is called *Normalized Multivariate Kurtosis* there, whereas $\hat{\beta}_{2,p} - p(p+2)$ bears the name *Mardia's Multivariate Kurtosis*.