III Lecture
Estimation of the Mean Vector and Covariance Matrix of Multivariate Normal Distribution

3.1. Maximum Likelihood Estimation

3.1.1. Likelihood function

Suppose we have observed n independent realizations of p-dimensional random vectors from $N_p(\mu, \mathbf{\Sigma})$. Suppose for simplicity that $\mathbf{\Sigma}$ is non-singular. The data matrix has the form

$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & .. & x_{1j} & .. & x_{1n} \\ x_{21} & x_{22} & .. & x_{2j} & .. & x_{2n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_{i1} & x_{i2} & .. & x_{ij} & .. & x_{in} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_{p1} & x_{p2} & .. & x_{pj} & .. & x_{pn} \end{pmatrix} = [\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_n}] \tag{3.1}$$

The goal to estimate the unknown mean vector and the covariance matrix of the multivariate normal distribution by the Maximum Likelihood Estimation (MLE) method.

Based on our knowledge from II Lecture we can write down the *Likelihood function*

$$L(\mathbf{x}; \mu, \mathbf{\Sigma}) = (2\pi)^{-\frac{np}{2}} |\mathbf{\Sigma}|^{-\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x_i}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x_i}-\mu)} \tag{3.2}$$

(Note that we have substituted the observations in (3.2) and consider L as a function of the unknown parameters $\mu, \mathbf{\Sigma}$ only. Correspondingly, we get the *log-likelihood function* in the form

$$\log L(\mathbf{x}; \mu, \mathbf{\Sigma}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\mathbf{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x_i}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x_i}-\mu) \tag{3.3}$$

It is well known that maximizing either (3.2) or (3.3) will give the same solution for the MLE.

We start deriving the MLE by trying to maximize (3.3). To this end, first note that by utilising properties of traces from I lecture, we can transform:

$$\sum_{i=1}^{n}(\mathbf{x_i}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x_i}-\mu) = \sum_{i=1}^{n} tr[\mathbf{\Sigma^{-1}}(\mathbf{x_i}-\mu)(\mathbf{x_i}-\mu)'] =$$

$$tr[\mathbf{\Sigma^{-1}}(\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{x_i}-\mu)(\mathbf{x_i}-\mu)')] = \tag{1}$$

(by adding $\pm\mathbf{\bar{x}} = \mathbf{\frac{1}{n}\sum_{i=1}^{n} x_i}$ to each term $(\mathbf{x_i}-\mu)$ in $\sum_{i=1}^{n}(\mathbf{x_i}-\mu)(\mathbf{x_i}-\mu)'$)

$$tr[\mathbf{\Sigma^{-1}}(\sum_{i=1}^{n}(\mathbf{x_i}-\mathbf{\bar{x}})(\mathbf{x_i}-\mathbf{\bar{x}})' + \mathbf{n}(\mathbf{\bar{x}}-\mu)(\mathbf{\bar{x}}-\mu)')] =$$

$$tr[\mathbf{\Sigma^{-1}}(\sum_{i=1}^{n}(\mathbf{x_i}-\bar{\mathbf{x}})(\mathbf{x_i}-\bar{\mathbf{x}})')]+\mathbf{n}(\bar{\mathbf{x}}-\mu)'\mathbf{\Sigma^{-1}}(\bar{\mathbf{x}}-\mu)$$

Thus

$$\log L(\mathbf{x};\mu,\mathbf{\Sigma})=-\frac{np}{2}\log(2\pi)-\frac{n}{2}\log(|\mathbf{\Sigma}|)-\frac{1}{2}tr[\mathbf{\Sigma^{-1}}(\sum_{i=1}^{n}(\mathbf{x_i}-\bar{\mathbf{x}})(\mathbf{x_i}-\bar{\mathbf{x}})')]-\frac{1}{2}\mathbf{n}(\bar{\mathbf{x}}-\mu)'\mathbf{\Sigma^{-1}}(\bar{\mathbf{x}}-\mu) \tag{3.4}$$

3.1.2 Maximum Likelihood Estimators

The MLE are the ones that maximize (3.4). Looking at (3.4) we realize that (since $\mathbf{\Sigma}$ is non-negative definite) the minimal value for $\frac{1}{2}\mathbf{n}(\bar{\mathbf{x}}-\mu)'\mathbf{\Sigma^{-1}}(\bar{\mathbf{x}}-\mu)$ is zero and is attained when $\mu=\bar{\mathbf{x}}$. It remains to find the optimal value for $\mathbf{\Sigma}$. We will use the following

Theorem 3.1 (Anderson's lemma) If $\mathbf{A}\in\mathscr{M}_{\mathbf{p},\mathbf{p}}$ is symmetric positive definite then the maximum of the function $h(G)=-n\log(|G|)-tr(G^{-1}\mathbf{A})$ (defined over the set of symmetric positive definite matrices $G\in\mathscr{M}_{p,p}$ exists, occurs at $G=\frac{1}{n}\mathbf{A}$ and has the maximal value of $np\log(n)-n\log(|\mathbf{A}|)-np$.

Proof (sketch, details at lecture): Indeed, (see properties of traces):

$$tr(G^{-1}\mathbf{A})=tr((G^{-1}\mathbf{A}^{\frac{1}{2}})\mathbf{A}^{\frac{1}{2}})=tr(\mathbf{A}^{\frac{1}{2}}G^{-1}\mathbf{A}^{\frac{1}{2}})$$

Let $\eta_i, i=1,\dots,p$ be the eigenvalues of $\mathbf{A}^{\frac{1}{2}}G^{-1}\mathbf{A}^{\frac{1}{2}}$. Then (since the matrix $\mathbf{A}^{\frac{1}{2}}G^{-1}\mathbf{A}^{\frac{1}{2}}$ is positive definite) $\eta_i>0, i=1,\dots,p$. Also, $tr(\mathbf{A}^{\frac{1}{2}}G^{-1}\mathbf{A}^{\frac{1}{2}})=\sum_{i=1}^{p}\eta_i$ and $|\mathbf{A}^{\frac{1}{2}}G^{-1}\mathbf{A}^{\frac{1}{2}}|=\prod_{i=1}^{p}\eta_i$ holds (Why (?!) Use the spectral decomposition!). Hence

$$-n\log|G|-tr(G^{-1}\mathbf{A})=n.\sum_{i=1}^{p}\log\eta_i-n\log|\mathbf{A}|-\sum_{i=1}^{p}\eta_i \tag{3.5}$$

Considering the expression $n\sum_{i=1}^{p}\log\eta_i-n\log|\mathbf{A}|-\sum_{i=1}^{p}\eta_i$ as a function of the eigenvalues $\eta_i, i=1,\dots,p$ we realize that it has a **maximum** which is attained when all $\eta_i=n, i=1,\dots,p$. Indeed, the first partial derivatives with respect to $\eta_i, i=1,\dots,p$ are equal to $\frac{n}{\eta_i}-1$ and hence the stationary points are $\eta_i^*=n, i=1,\dots,p$. The matrix of second derivatives calculated at $\eta_i^*=n, i=1,\dots,p$ is equal to $-\mathbf{I_p}$ and hence the stationary points give rise to a maximum of the function. Now, we can check directly by substituting the η^* values that the maximal value of the function is $np\log(n)-n\log(|\mathbf{A}|)-np$. But a direct substitution in the formula $h(G)=-n\log(|G|)-tr(G^{-1}\mathbf{A})$ with $G=\frac{1}{n}\mathbf{A}$ also gives rise to $np\log(n)-n\log(|\mathbf{A}|)-np$, i.e. the maximum is attained at $G=\frac{1}{n}\mathbf{A}$.

Using the structure of the log-likelihood function in (3.4) and Theorem 3.1 (applied for the case $\mathbf{A}=\sum_{i=1}^{n}(\mathbf{x_i}-\bar{\mathbf{x}})(\mathbf{x_i}-\bar{\mathbf{x}})'$ (!)) it is now easy to formulate following

Theorem 3.2 Suppose $\mathbf{x_1},\mathbf{x_2},\dots,\mathbf{x_n}$ is a random sample from $N_p(\mu,\mathbf{\Sigma}), p<n$. Then $\hat{\mu}=\bar{\mathbf{x}}$ and $\hat{\mathbf{\Sigma}}=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i}-\bar{\mathbf{x}})(\mathbf{x_i}-\bar{\mathbf{x}})'$ are the *maximum likelihood estimators* of μ and $\mathbf{\Sigma}$, respectively.

Remark 3.1.3. Alternative proofs of Theorem 3.2 are also available that utilize some formal rules for vector and matrix differentiation that have been developed as a standard machinery in multivariate analysis (recall that according to the folklore, in order to find the maximum of the log-likelihood, we need to differentiate it with respect to its arguments, i.e. with respect to the *vector* μ and to the *matrix* $\mathbf{\Sigma}$), set the derivatives equal to zero and solve the corresponding equation system. If time permits, these matrix differentiation rules will also be discussed later in this course.

3.1.4. Application in correlation matrix estimation

The correlation matrix can be defined in terms of the elements of the covariance matrix $\boldsymbol{\Sigma}$. The correlation coefficients $\rho_{ij}, i = 1, \dots, p, j = 1, \dots, p$ are defined as $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$ where $\boldsymbol{\Sigma} = (\sigma_{ij}, i = 1 \dots, p; j = 1, \dots, p)$ is the covariance matrix. Note that $\rho_{ii} = 1, i = 1, \dots, p$. To derive the MLE of $\rho_{ij}, i = 1, \dots, p, j = 1, \dots, p$ we note that these are continuous transformations of the covariances whose maximum likelihood estimators have already been derived. Then we can claim (*according to the transformation invariance properties of MLE*) that

$$\hat{\rho_{ij}} = \frac{\hat{\sigma_{ij}}}{\sqrt{\hat{\sigma_{ii}}}\sqrt{\hat{\sigma_{jj}}}}, i = 1, \dots, p, j = 1, \dots, p \tag{3.6}$$

3.1.5. Sufficiency of $\hat{\mu}$ and $\hat{\boldsymbol{\Sigma}}$

Back from (3.4) we can write the likelihood function as

$$L(\mathbf{x}; \mu, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{np}{2}} |\boldsymbol{\Sigma}|^{\frac{n}{2}}} e^{-\frac{1}{2} tr[\boldsymbol{\Sigma}^{-1}(\sum_{i=1}^{n} (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})' + \mathbf{n}(\bar{\mathbf{x}} - \mu)(\bar{\mathbf{x}} - \mu)')]}$$

which means that $L(\mathbf{x}; \mu, \boldsymbol{\Sigma})$ can be factorised into $L(\mathbf{x}; \mu, \boldsymbol{\Sigma}) = g_1(\mu, \boldsymbol{\Sigma}).g_2(\mu, \boldsymbol{\Sigma}; \hat{\mu}, \hat{\boldsymbol{\Sigma}})$, i.e. the likelihood function depends on the observations **only** through the values of $\hat{\mu} = \bar{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}}$. Hence the pair $\hat{\mu}$ and $\hat{\boldsymbol{\Sigma}}$ are *sufficient statistics* for μ and $\boldsymbol{\Sigma}$ in the case of a sample from $N_p(\mu, \boldsymbol{\Sigma})$. Note that the structure of the multivariate normal density was essentially used here thus underlying the importance of the check on adequacy of multivariate normality assumptions in practice. If testing indicates significant departures from multivariate normality then inferences that are based solely on $\hat{\mu}$ and $\hat{\boldsymbol{\Sigma}}$ may not be very reliable.

3.2. Distributions of MLE of mean vector and covariance matrix of multivariate normal distribution

Inference is not restricted to only find point estimators but also to construct confidence regions, test hypotheses etc. To this end we need the distribution of the estimators (or of suitably chosen functions of them).

3.2.1. Sampling distribution of $\bar{\mathbf{X}}$

In the univariate case ($p = 1$) it is well known that for a sample of n observations from *normal distribution* $N(\mu, \sigma^2)$ the sample mean is normally distributed: $N(\mu, \frac{\sigma^2}{n})$. Moreover, the sample mean and the sample variance are *independent* in the case of sampling from a univariate normal population. This fact was very useful in developing t-statistics for testing the mean vector. Do we have similar statements about the sample mean and sample variance in the multivariate ($p > 1$) case?

Let the random vector $\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i \in R^p$. For any $\mathbf{L} \in \mathbf{R^p} : \mathbf{L}'\bar{\mathbf{X}}$ is a linear combination of normals and hence is normal (see Definition 2.2.1). Since taking expected value is a linear operation, we have $E\bar{\mathbf{X}} = \frac{\mathbf{1}}{\mathbf{n}}\mathbf{n}\mu = \mu$; In analogy with the univariate case we could formally write $Cov\bar{\mathbf{X}} = \frac{1}{n^2}nCovX_1 = \frac{1}{n}\boldsymbol{\Sigma}$ and hence $\bar{\mathbf{X}} \sim \mathbf{N_p}(\mu, \frac{\mathbf{1}}{\mathbf{n}}\boldsymbol{\Sigma})$. But we would like to develop a more appropriate machinery for the multivariate case that would help us to more rigorously prove statements like the last one. It is based on operations with *Kronecker products*.

Kronecker product of two matrices $\mathbf{A} \in \mathscr{M}_{\mathbf{m,n}}$ and $\mathbf{B} \in \mathscr{M}_{\mathbf{p,q}}$ is denoted by $\mathbf{A} \otimes \mathbf{B}$ and is defined (in block matrix notation) as

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2n}\mathbf{B} \\ \dots & \dots & \dots & \dots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{pmatrix} \tag{3.7}$$

The following basic properties of Kronecker products will be used:

$$\begin{aligned} (A \otimes B) \otimes C &= A \otimes (B \otimes C) \\ (A + B) \otimes C &= A \otimes C + B \otimes C \\ (A \otimes B)' &= A' \otimes B' \\ (A \otimes B)^{-1} &= A^{-1} \otimes B^{-1} \\ (A \otimes B)(C \otimes D) &= AC \otimes BD \end{aligned}$$

(whenever the corresponding matrix products and inverses exist)

$$\begin{aligned} tr(A \otimes B) &= (trA)tr(B) \\ |A \otimes B| &= |A|^p |B|^m \end{aligned}$$

(in case $\mathbf{A} \in \mathscr{M}_{\mathbf{m,m}}, \mathbf{B} \in \mathscr{M}_{\mathbf{p,p}}$)

In addition , the $\rightarrow$ operation on a matrix $A \in \mathscr{M}_{m,n}$ will be defined. This operation creates a vector $\vec{A} \in R^{mn}$ which is composed by stacking the n columns of the matrix $A \in \mathscr{M}_{m,n}$ under each other (the second below the first etc). For matrices A, B and C (of suitable dimensions) it holds:

$$\overrightarrow{ABC} = (C' \otimes A)\vec{B}$$

Let us see how we could utilize the above to derive the distribution of $\bar{\mathbf{X}}$. Denote by $\mathbf{1_n}$ the vector of n ones. Note that if $\mathbf{X}$ is the random matrix (see (2.1) in II lecture) then $\vec{\mathbf{X}} \sim N(\mathbf{1_n} \otimes \mu, I_n \otimes \Sigma)$ and $\bar{\mathbf{X}} = \frac{1}{n}(\mathbf{1'_n} \otimes \mathbf{I_p})\vec{\mathbf{X}}$. Hence $\bar{\mathbf{X}}$ is multivariate normal with

$$E\bar{\mathbf{X}} = \frac{1}{n}(\mathbf{1'_n} \otimes \mathbf{I_p})(\mathbf{1_n} \otimes \mu) = \frac{\mathbf{1}}{\mathbf{n}}(\mathbf{1'_n}\mathbf{1_n} \otimes \mu) = \frac{\mathbf{1}}{\mathbf{n}}\mathbf{n}\mu = \mu$$

$$Cov\bar{\mathbf{X}} = \mathbf{n^{-2}}(\mathbf{1'_n} \otimes \mathbf{I_p})(\mathbf{I_n} \otimes \mathbf{\Sigma})(\mathbf{1_n} \otimes \mathbf{I_p}) = \mathbf{n^{-2}}(\mathbf{1'_n}\mathbf{1_n} \otimes \mathbf{\Sigma}) = \mathbf{n^{-1}}\mathbf{\Sigma}$$

How can we show that $\bar{\mathbf{X}}$ and $\hat{\mathbf{\Sigma}}$ are independent? Recall the likelihood function

$$L(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{np}{2}} |\mathbf{\Sigma}|^{\frac{n}{2}}} e^{-\frac{1}{2}tr[\mathbf{\Sigma^{-1}}(\sum_{i=1}^n (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})' + \mathbf{n}(\bar{\mathbf{x}} - \mu)(\bar{\mathbf{x}} - \mu)')]}$$

We have 2 summands in the exponent from which one is a function of the observations through $n\hat{\mathbf{\Sigma}} = \sum_{i=1}^n (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})'$ only and the other one depends on the observations through $\bar{\mathbf{X}}$ only. The idea is now to transform the original data matrix $\mathbf{X} \in \mathscr{M}_{\mathbf{p,n}}$ into a new matrix $\mathbf{Z} \in \mathscr{M}_{\mathbf{p,n}}$ of n independent $N(\mathbf{0}, \mathbf{\Sigma})$ vectors in such a way that $\bar{\mathbf{X}}$ would only be a function of $\mathbf{Z_1}$ whereas $\sum_{i=1}^n (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})'$ would only be a function of $\mathbf{Z_2}, \dots, \mathbf{Z_n}$. If we succeed then clearly $\bar{\mathbf{X}}$ and $\sum_{i=1}^n (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})' = n\hat{\mathbf{\Sigma}}$ would be independent.

Now the claim is that the sought after transformation is given by $\mathbf{Z} = \mathbf{XA}$ with $\mathbf{A} \in \mathscr{M}_{\mathbf{n,n}}$ being an *orthogonal matrix* with a first column equal to $\frac{1}{\sqrt{n}}\mathbf{1}_n$. Note that the first column of $\mathbf{Z}$ would be then $\sqrt{n}\bar{\mathbf{X}}$. (An explicit form of the matrix $\mathbf{A}$ will be discussed at the lecture). Since $\vec{\mathbf{Z}} = \overrightarrow{\mathbf{I_p X A}} = (\mathbf{A}' \otimes \mathbf{I_p})\vec{\mathbf{X}}$, the Jacobian of the transformation ($\vec{\mathbf{X}}$ into $\vec{\mathbf{Z}}$) is $|\mathbf{A}' \otimes \mathbf{I_p}| = |\mathbf{A}|^{\mathbf{p}} = \pm\mathbf{1}$ (note that $\mathbf{A}$ is orthogonal). Therefore, the absolute value of the Jacobian is equal to one. For $\vec{\mathbf{Z}}$ we have:

$$E(\vec{\mathbf{Z}}) = (\mathbf{A}' \otimes \mathbf{I_p})(\mathbf{1_n} \otimes \mu) = \mathbf{A}'\mathbf{1_n} \otimes \mu = \begin{pmatrix} \sqrt{n} \\ 0 \\ \dots \\ 0 \end{pmatrix} \otimes \mu$$

Further,

$$Cov(\vec{\mathbf{Z}}) = (\mathbf{A}' \otimes \mathbf{I_p})(\mathbf{I_n} \otimes \mathbf{\Sigma})(\mathbf{A} \otimes \mathbf{I_p}) = \mathbf{A}'\mathbf{A} \otimes \mathbf{I_p \Sigma I_p} = \mathbf{I_n} \otimes \mathbf{\Sigma}$$

which means that the $\mathbf{Z_i}, \mathbf{i = 1}, \dots, \mathbf{n}$ are *independent*. Note $\mathbf{Z_1} = \sqrt{\mathbf{n}}\bar{\mathbf{X}}$ holds (because of the choice of the first column of the orthogonal matrix $\mathbf{A}$). Further

$$\sum_{i=1}^n (\mathbf{X_i} - \bar{\mathbf{X}})(\mathbf{X_i} - \bar{\mathbf{X}})' = \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i X_i'} - \frac{\mathbf{1}}{\mathbf{n}}\Big(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i}\Big)\Big(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i'}\Big) =$$

$$\mathbf{ZA'AZ'} - \mathbf{Z_1 Z_1'} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{Z_i Z_i'} - \mathbf{Z_1 Z_1'} = \sum_{\mathbf{i=2}}^{\mathbf{n}} \mathbf{Z_i Z_i'}$$

Hence we proved the following

Theorem 3.3 For a sample of size n from $N_p(\mu, \mathbf{\Sigma}), p < n$ the sample average $\bar{\mathbf{X}} \sim N_p(\mu, \frac{\mathbf{1}}{\mathbf{n}}\mathbf{\Sigma})$. Moreover, the MLE $\hat{\mu} = \bar{\mathbf{X}}$ and $\hat{\mathbf{\Sigma}}$ are independent.

3.2.2. Sampling distribution of the MLE of $\mathbf{\Sigma}$

Definition. A random matrix $\mathbf{U} \in \mathscr{M}_{\mathbf{p,p}}$ has a **Wishart distribution** with parameters Σ, p, n (denoting this by $\mathbf{U} \sim \mathbf{W_p}(\mathbf{\Sigma}, \mathbf{n})$) if there exist n independent random vectors $\mathbf{Y_1}, \dots, \mathbf{Y_n}$ each with $N_p(0, \Sigma)$ distribution such that the distribution of $\sum_{i=1}^n \mathbf{Y_i Y_i'}$ coincides with the distribution of $\mathbf{U}$.

Note that we *require* $p < n$ here and that $\mathbf{U}$ is necessarily non-negative definite.

Having in mind the proof of Theorem 3.3 we can claim that the distribution of the matrix $n\hat{\mathbf{\Sigma}} = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{X_i} - \bar{\mathbf{X}})(\mathbf{X_i} - \bar{\mathbf{X}})'$ is the same as that of $\sum_{i=2}^n \mathbf{Z_i Z_i'}$ and therefore is Wishart with parameters $\Sigma, p, n-1$. That is, we can denote:

$$\mathbf{n}\hat{\mathbf{\Sigma}} \sim W_p(\mathbf{\Sigma}, n-1).$$

The density formula for the Wishart distribution is given in several sources but we will not deal with it in this course. Some properties of Wishart distribution will be mentioned though since we will make use of them later in the course:

- If $p = 1$ and if we denote the "matrix" $\mathbf{\Sigma}$ by σ^2 (as usual) then $W_1(\Sigma, n)/\sigma^2 = \chi^2_n$. In particular, when $\sigma^2 = 1$ we see that $W_1(1, n)$ is exactly the χ^2_n random variable. In that sense we can state that the Wishart distribution is a generalisation (with respect to the dimension p) of the chi square distribution.

- For an arbitrary fixed matrix $\mathbf{H} \in \mathscr{M}_{\mathbf{k,p}}, k \le p$ one has:

$$n\mathbf{H}\hat{\mathbf{\Sigma}}\mathbf{H}' \sim \mathbf{W_k}(\mathbf{H\Sigma H}', \mathbf{n-1})$$

  (Why? Show it!).

- Refer to the previous case for the particular value of k=1. The matrix $\mathbf{H} \in \mathscr{M}_{\mathbf{1,p}}$ is just a p-dimensional row vector that we could denote by c'. Then:

  i) $n\frac{\mathbf{c}'\hat{\mathbf{\Sigma}}\mathbf{c}}{\mathbf{c}'\mathbf{\Sigma}\mathbf{c}} \sim \chi^2_{n-1}$.

  ii) $n\frac{\mathbf{c}'\mathbf{\Sigma}^{-1}\mathbf{c}}{\mathbf{c}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{c}} \sim \chi^2_{n-p}$

- Let us partition $\mathbf{S} = \frac{\mathbf{1}}{\mathbf{n-1}}\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{X_i} - \bar{\mathbf{X}})(\mathbf{X_i} - \bar{\mathbf{X}})' \in \mathscr{M}_{\mathbf{p,p}}$ into

$$\mathbf{S} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}, \mathbf{S_{11}} \in \mathscr{M}_{\mathbf{r,r}}, r < p, \mathbf{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \mathbf{\Sigma_{11}} \in \mathscr{M}_{\mathbf{r,r}}, r < p.$$

  Further, denote

$$\mathbf{S_{1|2}} = \mathbf{S_{11}} - \mathbf{S_{12}}\mathbf{S_{22}^{-1}}\mathbf{S_{21}}, \mathbf{\Sigma_{1|2}} = \mathbf{\Sigma_{11}} - \mathbf{\Sigma_{12}}\mathbf{\Sigma_{22}^{-1}}\mathbf{\Sigma_{21}}$$

  Then it holds

$$(n-1)\mathbf{S_{11}} \sim \mathbf{W_r}(\mathbf{\Sigma_{11}}, \mathbf{n-1})$$

$$(n-1)\mathbf{S_{1|2}} \sim \mathbf{W_r}(\mathbf{\Sigma_{1|2}}, \mathbf{n-p+r-1})$$