Topic 3: Estimation of vector mean and of variance matrices: point estimates

Estimation of the mean vector and covariance matrix of multivariate normal distribution

In the previous topic, we have derived the multivariate normal distribution and its basic properties. In order to make use of them, we must now estimate this distribution from observed data. This topic derives the estimators for the parameters of this distribution and some properties of these estimators.

Maximum Likelihood Estimation

Optional viewing: Maximum Likelihood estimation - An introduction

Likelihood function

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

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pj} & \cdots & X_{pn} \end{pmatrix} = [X_1, X_2, \ldots, X_n] \tag{1.8}$$

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 the previous topic we can write down the *likelihood function*

$$L(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})} \tag{1.9}$$

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

$$\log L(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) \tag{1.10}$$

It is well known that maximising either (1.9) or (1.10) will give the same solution for the MLE.

We start deriving the MLE by trying to maximise (1.10). To this end, first note that by utilising properties of traces from slide "Vectors and matrices", we can transform:

$$\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) = \sum_{i=1}^{n}\operatorname{tr}\left[\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top\right] =$$

$\operatorname{tr}\left[\Sigma^{-1}\left(\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top\right)\right] =$

(by adding $\pm\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$ to each term $(\boldsymbol{x}_i - \boldsymbol{\mu})$ in $\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$)

$\operatorname{tr}\left[\Sigma^{-1}\left(\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top\right)\right]$

$$= \operatorname{tr}\left[\Sigma^{-1}\left(\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top\right)\right] + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top\Sigma^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})$$

Thus

$$\begin{aligned}\log L(\boldsymbol{x};\boldsymbol{\mu},\Sigma) \;=\;& -\tfrac{np}{2}\log(2\pi)-\tfrac{n}{2}\log(|\Sigma|)-\tfrac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(\textstyle\sum_{i=1}^{n}\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}\right)^{\top}\right)\right]-\\ &\tfrac{1}{2}n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})\end{aligned} \tag{1.11}$$

Maximum Likelihood Estimators

The MLE are the ones that maximise (1.11). Looking at (1.11) we realise that (since Σ is non-negative definite) the minimal value for $\frac{1}{2}n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})$ is zero and is attained when $\boldsymbol{\mu}=\bar{\boldsymbol{x}}$.

Finding the MLE for Σ is more challenging. We will not be deriving it in this class, since it doesn't have much bearing on applications, but we will state the following intermediate result that will prove useful later:

Theorem 1.2 (Anderson's lemma)**.** *If $A\in\mathcal{M}_{p,p}$ is symmetric positive definite, then the maximum of the function $(G)=-n\log(|G|)-\operatorname{tr}\left(G^{-1}A\right)$ (defined over the set of symmetric positive definite matrices $G\in\mathcal{M}_{p,p}$) exists, occurs at $G=\frac{1}{n}A$ and has the maximal value of $np\log(n)-n\log(|A|)-np$.*

This is useful because the parts of the log-likelihood function in (1.11) that depend on Σ have the form required by Theorem (1.2) with $G=\Sigma$ and $A=\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\top}$. From there, we can show (again, don't worry about the details) the following:

Theorem 1.3. *Suppose $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n$ is a random sample from $N_p(\boldsymbol{\mu},\Sigma),p<n$. Then $\hat{\boldsymbol{\mu}}=\bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}}=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\top}$ are the* maximum likelihood estimators *of $\boldsymbol{\mu}$ and Σ, respectively.*

Application in correlation matrix estimation

The correlation matrix can be defined in terms of the elements of the covariance matrix Σ. The correlation coefficients $\rho_{ij},i=1,\ldots,p,j=1,\ldots,p$ are defined as $\rho_{ij}=\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$ where $\Sigma=(\sigma_{ij},i=1\ldots,p;j=1,\ldots,p)$ is the covariance matrix. Note that $\rho_{ii}=1,i=1,\ldots,p$. To derive the MLE of $\rho_{ij},i=1,\ldots,p,j=1,\ldots,p$ 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

$$\hat{\rho}_{ij}=\frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}}\sqrt{\hat{\sigma}_{jj}}},i=1,\ldots,p,j=1,\ldots,p. \tag{1.12}$$

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

Inference is not restricted to only finding 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).

Sampling distribution of $\bar{\boldsymbol{X}}$

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 (Basu's Lemma). 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 $\bar{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \in \mathbb{R}^p$. For any $\boldsymbol{l} \in \mathbb{R}^p : \boldsymbol{l}^\top \bar{\boldsymbol{X}}$ is a linear combination of normals and hence is normal (see Property 1 of MVN). Since taking expected value is a linear operation, we have $\mathrm{E}\,\bar{\boldsymbol{X}} = \frac{1}{n} n \boldsymbol{\mu} = \boldsymbol{\mu}$; In analogy with the univariate case we could formally write $\mathrm{Cov}\,\bar{\boldsymbol{X}} = \frac{1}{n^2} n\,\mathrm{Cov}\,\boldsymbol{X}_1 = \frac{1}{n}\Sigma$ and hence $\bar{\boldsymbol{X}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\Sigma)$.

If you are interested, see the next (optional) slide for a detailed proof.

Independence of $\bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}}$

How can we show that $\bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}}$ are independent? If you are interested, see the next (optional) slide, where we prove the following theorem:

Theorem 1.4. *For a sample of size n from $N_p(\boldsymbol{\mu}, \Sigma), p < n$ the sample average $\bar{\boldsymbol{X}} \sim N_p(\mu, \frac{1}{n}\Sigma)$. Moreover, the MLE $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}}$ are independent.*

Sampling distribution of the MLE of Σ

Definition 1.2. A random matrix $\boldsymbol{U} \in \mathcal{M}_{p,p}$ has a **Wishart distribution** with parameters Σ, p, n (denoting this by $\boldsymbol{U} \sim W_p(\Sigma, n)$) if there exist n independent random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ each with $N_p(0, \Sigma)$ distribution such that the distribution of $\sum_{i=1}^{n} \boldsymbol{Y}_i \boldsymbol{Y}_i^\top$ coincides with the distribution of $\boldsymbol{U}$.

Note that we *require* that $p < n$ and that $\boldsymbol{U}$ be non-negative definite.

Having in mind the proof of Theorem 1.4 we can claim that the distribution of the matrix $n\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top$ is the same as that of $\sum_{i=2}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i^\top$ and therefore is Wishart with parameters $\Sigma, p, n-1$. That is, we can denote:

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

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:

1. If $p=1$ and if we denote the "matrix" Σ 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 χ^2 distribution.
2. For an arbitrary fixed matrix $H \in \mathcal{M}_{k,p}, k \le p$ one has:

   $$nH\hat{\boldsymbol{\Sigma}}H^\top \sim W_k(H\Sigma H^\top, n-1).$$

3. Refer to the previous case for the particular value of $k=1$. The matrix $H \in \mathcal{M}_{1,p}$ is just a p-dimensional row vector that we could denote by $\boldsymbol{c}^\top$. Then:
   (i) $n\frac{\boldsymbol{c}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{c}}{\boldsymbol{c}^\top\Sigma\boldsymbol{c}} \sim \chi^2_{n-1}$
   (ii) $n\frac{\boldsymbol{c}^\top\Sigma^{-1}\boldsymbol{c}}{\boldsymbol{c}^\top\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{c}} \sim \chi^2_{n-p}$
4. Let us partition $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top \in \mathcal{M}_{p,p}$ into

   $$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{pmatrix}, \boldsymbol{S}_{11} \in \mathcal{M}_{r,r}, r < p$$

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

   Further, denote

   $$\boldsymbol{S}_{1|2} = \boldsymbol{S}_{11} - \boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21},\ \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

   Then it holds

   $$(n-1)\boldsymbol{S}_{11} \sim W_r(\Sigma_{11}, n-1)$$

   $$(n-1)\boldsymbol{S}_{1|2} \sim W_r(\Sigma_{1|2}, n-p+r-1)$$

Distributions of MLE of mean vector and covariance matrix of multivariate normal distribution: Detailed derivations

Note: This slide is not examinable

Sampling distribution of $\bar{\boldsymbol{X}}$

Here, we prove the claims about the sampling distribution of $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i \in \mathbb{R}^p$ from the previous slide.

Let the random vector $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i \in \mathbb{R}^p$. For any $\boldsymbol{l} \in \mathbb{R}^p : \boldsymbol{l}^\top\bar{\boldsymbol{X}}$ is a linear combination of normals and hence is normal (see Definition 1.1). Since taking expected value is a linear operation, we have $\mathrm{E}\,\bar{\boldsymbol{X}} = \frac{1}{n}n\boldsymbol{\mu} = \boldsymbol{\mu}$; In analogy with the univariate case we could formally write $\mathrm{Cov}\,\bar{\boldsymbol{X}} = \frac{1}{n^2}n\,\mathrm{Cov}\,\boldsymbol{X}_1 = \frac{1}{n}\Sigma$ and hence $\bar{\boldsymbol{X}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\Sigma)$. 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 $A \in \mathcal{M}_{m,n}$ *and* $B \in \mathcal{M}_{p,q}$ is denoted by $A \otimes B$ and is defined (in block matrix notation) as

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix} \tag{1.13}$$

The following basic properties of Kronecker products will be used:

$$(A \otimes B) \otimes C = A \otimes (B \otimes C)$$

$$(A + B) \otimes C = A \otimes C + B \otimes C$$

$$(A \otimes B)^\top = A^\top \otimes B^\top$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$

$$(A \otimes B)(C \otimes D) = AC \otimes BD$$

(whenever the corresponding matrix products and inverses exist)

$$\mathrm{tr}(A \otimes B) = \mathrm{tr}(A)\,\mathrm{tr}(B)$$

$$|A \otimes B| = |A|^p|B|^m$$

(in case $A \in \mathcal{M}_{m,m}, B \in \mathcal{M}_{p,p}$).

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

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

Let us see how we could utilise the above to derive the distribution of $\bar{\boldsymbol{X}}$. Denote by $\mathbf{1}_n$ the vector of n ones. Note that if $\boldsymbol{X}$ is the random data matrix (see (0.11) in slide "Standard facts about multivariate distributions") then $\vec{\boldsymbol{X}} \sim N(\mathbf{1}_n \otimes \boldsymbol{\mu}, I_n \otimes \Sigma)$ and $\bar{\boldsymbol{X}} = \frac{1}{n}(\mathbf{1}_n{}^\top \otimes I_p)\vec{\boldsymbol{X}}$. Hence $\bar{\boldsymbol{X}}$ is multivariate normal with

$$\begin{aligned}
\mathrm{E}\,\bar{\boldsymbol{X}} &= \frac{1}{n}(\mathbf{1}_n{}^\top \otimes I_p)(\mathbf{1}_n \otimes \boldsymbol{\mu}) = \frac{1}{n}(\mathbf{1}_n{}^\top \mathbf{1}_n \otimes \boldsymbol{\mu}) = \frac{1}{n} n\boldsymbol{\mu} = \boldsymbol{\mu}, \\
\mathrm{Cov}\,\bar{\boldsymbol{X}} &= n^{-2}(\mathbf{1}_n{}^\top \otimes I_p)(I_{_n} \otimes \Sigma)(\mathbf{1}_n \otimes I_p) = n^{-2}(\mathbf{1}_n{}^\top \mathbf{1}_n \otimes \Sigma) = n^{-1}\Sigma.
\end{aligned}$$

Independence of $\bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}}$

Here, we prove Theorem 1.4 from the previous slide. Recall the likelihood function

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

We have two summands in the exponent from which one is a function of the observations through $n\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$ only and the other one depends on the observations through $\bar{\boldsymbol{x}}$ only. The idea is now to transform the original data matrix $\boldsymbol{X} \in \mathcal{M}_{p,n}$ into a new matrix $\boldsymbol{Z} \in \mathcal{M}_{p,n}$ of n independent $N(\mathbf{0}, \Sigma)$ vectors in such a way that $\bar{\boldsymbol{X}}$ would only be a function of $\boldsymbol{Z}_1$ whereas $\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$ would only be a function of $\boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n$. If we succeed then clearly $\bar{\boldsymbol{X}}$ and $\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top = n\hat{\boldsymbol{\Sigma}}$ would be independent.

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

$$\mathrm{E}(\vec{\boldsymbol{Z}}) = (A^\top \otimes I_p)(\mathbf{1}_n \otimes \boldsymbol{\mu}) = A^\top \mathbf{1}_n \otimes \boldsymbol{\mu} = \begin{pmatrix} \sqrt{n} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \otimes \boldsymbol{\mu}$$

Further,

$$\operatorname{Cov}(\vec{\boldsymbol{Z}}) = (A^\top \otimes I_p)(I_n \otimes \Sigma)(A \otimes I_p) = A^\top A \otimes I_p \Sigma I_p = I_n \otimes \Sigma$$

which means that the $\boldsymbol{Z}_i, i = 1, \ldots, n$ are *independent*. Note $\boldsymbol{Z}_1 = \sqrt{n}\bar{\boldsymbol{X}}$ holds (because of the choice of the first column of the orthogonal matrix A). Further

$$\sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top = \sum i = 1^n \boldsymbol{X}_i \boldsymbol{X}_i^\top - \frac{1}{n}(\sum_{i=1}^{n} \boldsymbol{X}_i)(\sum_{i=1}^{n} \boldsymbol{X}_i^\top) =$$

$$\boldsymbol{Z} A^\top A \boldsymbol{Z}^\top - \boldsymbol{Z}_1 {\boldsymbol{Z}_1}^\top = \sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i^\top - \boldsymbol{Z}_1 \boldsymbol{Z}_1^\top = \sum_{i=2}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i^\top$$

Hence we proved the following

Theorem 1.4. *For a sample of size n from $N_p(\boldsymbol{\mu}, \Sigma), p < n$ the sample average $\bar{\boldsymbol{X}} \sim N_p(\mu, \frac{1}{n}\Sigma)$. Moreover, the MLE $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}}$ are independent.*