3. Multivariate Normal Distribution

The MVN distribution is a generalization of the univariate normal distribution which has the density function (p.d.f.)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \qquad -\infty < x < \infty$$

where μ = mean of distribution, σ^2 = variance. In $p-$dimensions the density becomes

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \right\} \tag{3.1}$$

Within the mean vector $\boldsymbol{\mu}$ there are p (independent) parameters and within the symmetric covariance matrix $\boldsymbol{\Sigma}$ there are $\frac{1}{2}p(p+1)$ independent parameters [$\frac{1}{2}p(p+3)$ independent parameters in total]. We use the notation

$$\boldsymbol{x} \sim \boldsymbol{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3.2}$$

to denote a RV $\boldsymbol{x}$ having the $p-$variate MVN distribution with

$$\begin{aligned} \mathbb{E}(\boldsymbol{x}) &= \boldsymbol{\mu} \\ Cov(\boldsymbol{x}) &= \boldsymbol{\Sigma} \end{aligned}$$

Note that MVN distributions are entirely characterized by the first and second moments of the distribution.

3.1 Basic properties

If $\boldsymbol{x}$ $(p \times 1)$is MVN with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

- Any linear combination of $\boldsymbol{x}$ is MVN

  Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{c}$ with $\boldsymbol{A}\,(q \times p)$ and $\boldsymbol{c}\,(q \times 1)$ then

  $$\boldsymbol{y} \sim \boldsymbol{N}_q(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$$

  where $\boldsymbol{\mu}_y = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{c}$ and $\boldsymbol{\Sigma}_y = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$.

- Any subset of variables in $\boldsymbol{x}$ has a MVN distribution.

- If a set of variables is uncorrelated, then they are independently distributed. In particular

  i) if $\sigma_{ij} = 0$ then x_i, x_j are independent.

ii) if $\boldsymbol{x}$ is MVN with covariance matrix $\boldsymbol{\Sigma}$, then $\boldsymbol{Ax}$ and $\boldsymbol{Bx}$ are independent if and only if

$$\begin{aligned} Cov\left(\boldsymbol{Ax},\boldsymbol{Bx}\right) &= \boldsymbol{A\Sigma B}^T \\ &= \boldsymbol{0} \end{aligned} \tag{3.3}$$

- Conditional distributions are MVN.

Result

For the MVN distribution, variable are uncorrelated $\Leftrightarrow$ variable are independent.

Proof

Let $\boldsymbol{x}$ $(p \times 1)$ be partitioned as

$$\boldsymbol{x} = \begin{array}{cc} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} & \begin{array}{l} q \\ p-q \end{array} \end{array}$$

with mean vector

$$\boldsymbol{\mu} = \begin{array}{cc} \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} & \begin{array}{l} q \\ p-q \end{array} \end{array}$$

and covariance matrix

$$\boldsymbol{\Sigma} = \begin{array}{cc} \begin{array}{cc} q & p-q \end{array} & \\ \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} & \begin{array}{l} q \\ p-q \end{array} \end{array}$$

i) Independent $\Rightarrow$ uncorrelated (always holds).

Suppose $\boldsymbol{x}_1, \boldsymbol{x}_2$ are independent. Then $f(\boldsymbol{x}_1, \boldsymbol{x}_2) = h(\boldsymbol{x}_1)\, g(\boldsymbol{x}_2)$ is a factorization of the multivariate p.d.f.and $\boldsymbol{\Sigma}_{12} = Cov(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}\left[(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)^T\right]$ factorizes into the product of $\mathbb{E}\left[(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)\right]$ and $\mathbb{E}\left[(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)^T\right]$ which are both zero since $\mathbb{E}(\boldsymbol{x}_1) = \boldsymbol{\mu}_1$ and $\mathbb{E}(\boldsymbol{x}_2) = \boldsymbol{\mu}_2$. Hence $\boldsymbol{\Sigma}_{12} = 0$.

ii) Uncorrelated $\Rightarrow$ independent (for MVN)

This result depends on factorizing the p.d.f. (3.1) when $\boldsymbol{\Sigma}_{12} = 0$.

In this case $(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ has the partitioned form

$$\begin{aligned} & \left[\boldsymbol{x}_1^T - \boldsymbol{\mu}_1^T, \quad \boldsymbol{x}_2^T - \boldsymbol{\mu}_2^T\right] \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{x}_2 - \boldsymbol{\mu}_2 \end{bmatrix} \\ = & \left[\boldsymbol{x}_1^T - \boldsymbol{\mu}_1^T, \quad \boldsymbol{x}_2^T - \boldsymbol{\mu}_2^T\right] \begin{bmatrix} \boldsymbol{\Sigma}_{11}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{x}_2 - \boldsymbol{\mu}_2 \end{bmatrix} \\ = & (\boldsymbol{x}_1 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_{11}^{-1} (\boldsymbol{x}_1 - \boldsymbol{\mu}_1) + (\boldsymbol{x}_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^{-1} (\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \end{aligned}$$

so that $\exp\{(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$ factorizes into the product of
$\exp\left\{(\boldsymbol{x}_1-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_1)\right\}$ and $\exp\left\{(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)\right\}$.
Therefore the p.d.f. can be written as

$$f(\boldsymbol{x}) = g(\boldsymbol{x}_1)\, h(\boldsymbol{x}_2)$$

proving that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are independent. ■

3.2 Conditional distribution

Let $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{bmatrix} \begin{matrix} q \\ p-q\end{matrix}$ be a partitioned MVN random $p-$vector,

with mean $\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}$$

The conditional distribution of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1 = \boldsymbol{x}_1$ is MVN with

$$\mathbb{E}(\boldsymbol{X}_2|\boldsymbol{X}_1 = \boldsymbol{x}_1) = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_1) \tag{3.4a}$$
$$Cov(\boldsymbol{X}_2|\boldsymbol{X}_1 = \boldsymbol{x}_1) = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \tag{3.4b}$$

Note: the notation $\boldsymbol{X}_1$ to denote the *r.v.* and $\boldsymbol{x}_1$ to denote a specific constant value (realization of $\boldsymbol{X}_1$) will be very useful here.

Proof of 3.4a

Define a transformation from $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ to new variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2' = \boldsymbol{X}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1$. This is achieved by the linear transformation

$$\begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2'\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{bmatrix} \tag{3.5a}$$
$$= \boldsymbol{A}\boldsymbol{X} \qquad \text{say.} \tag{3.5b}$$

This linear relationship shows that $\boldsymbol{X}_1, \boldsymbol{X}_2'$ are jointly MVN (by first property of MVN stated above.)

We now show that $\boldsymbol{X}_2'$ and $\boldsymbol{X}_1$ are *independent* by proving that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2'$ are uncorrelated.
Approach 1:

$$
\begin{aligned}
Cov\left(\boldsymbol{X}_1, \boldsymbol{X}_2'\right) &= Cov\left(\boldsymbol{X}_1, \boldsymbol{X}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1\right) \\
&= Cov(\boldsymbol{X}_1, \boldsymbol{X}_2) - Cov\left(\boldsymbol{X}_1, \boldsymbol{X}_1\right)\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\
&= \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\
&= \boldsymbol{0}
\end{aligned}
$$

Approach 2:

In (3.3), write $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{B} \\ \boldsymbol{C} \end{bmatrix}$ where $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}$ and $\boldsymbol{C} = \begin{bmatrix} -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \boldsymbol{I} \end{bmatrix}$

$$
\begin{aligned}
Cov\left(\boldsymbol{X}_1, \boldsymbol{X}_2'\right) &= Cov\left(\boldsymbol{BX}, \boldsymbol{CX}\right) \\
&= \boldsymbol{B\Sigma C}^T \\
&= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ \boldsymbol{I} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \end{bmatrix} \begin{bmatrix} -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ \boldsymbol{I} \end{bmatrix} \\
&= \boldsymbol{0}
\end{aligned}
$$

Since $\boldsymbol{X}_2'$ and $\boldsymbol{X}_1$ are MVN and uncorrelated they are independent. Thus

$$
\begin{aligned}
\mathbb{E}\left(\boldsymbol{X}_2' | \boldsymbol{X}_1 = \boldsymbol{x}_1\right) &= \mathbb{E}\left(\boldsymbol{X}_2'\right) \\
&= \mathbb{E}\left(\boldsymbol{X}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1\right) \\
&= \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1
\end{aligned}
$$

Now, as $\boldsymbol{X}_2' = \boldsymbol{X}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1$ and $\boldsymbol{X}_1 = \boldsymbol{x}_1$ is given, we have

$$
\begin{aligned}
\mathbb{E}\left(\boldsymbol{X}_2 | \boldsymbol{X}_1 = \boldsymbol{x}_1\right) &= \mathbb{E}\left(\boldsymbol{X}_2' | \boldsymbol{X}_1 = \boldsymbol{x}_1\right) + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{x}_1 \\
&= \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{x}_1 \\
&= \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\left(\boldsymbol{x}_1 - \boldsymbol{\mu}_1\right)
\end{aligned}
$$

as required.

Proof of 3.4b

Because $\boldsymbol{X}_2'$ is independent of $\boldsymbol{X}_1$

$$
Cov\left(\boldsymbol{X}_2' | \boldsymbol{X}_1 = \boldsymbol{x}_1\right) = Cov\left(\boldsymbol{X}_2'\right)
$$

The left hand side is

$$\begin{aligned} LHS &= Cov\left(\boldsymbol{X}_2'|\boldsymbol{X}_1 = \boldsymbol{x}_1\right) \\ &= Cov\left(\boldsymbol{X}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{x}_1|\boldsymbol{X}_1 = \boldsymbol{x}_1\right) \\ &= Cov\left(\boldsymbol{X}_2|\boldsymbol{X}_1 = \boldsymbol{x}_1\right) \end{aligned}$$

The right hand side is

$$\begin{aligned} RHS &= Cov\left(\boldsymbol{X}_2'\right) \\ &= Cov\left(\boldsymbol{X}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1\right) \\ &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \end{aligned}$$

following from the general expansion

$$\begin{aligned} Cov\left(\boldsymbol{X}_2 - \boldsymbol{D}\boldsymbol{X}_1\right) &= Cov\left(\boldsymbol{X}_2, \boldsymbol{X}_2\right) - \boldsymbol{D}Cov\left(\boldsymbol{X}_1, \boldsymbol{X}_2\right) \\ &\quad - Cov\left(\boldsymbol{X}_2, \boldsymbol{X}_1\right)\boldsymbol{D}^T + \boldsymbol{D}Cov\left(\boldsymbol{X}_1, \boldsymbol{X}_1\right)\boldsymbol{D}^T \end{aligned}$$

with $\boldsymbol{D} = \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$. Therefore

$$Cov\left(\boldsymbol{X}_2|\boldsymbol{X}_1 = \boldsymbol{x}_1\right) = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$$

as required.

Example
Let $\boldsymbol{x}$ have a MVN distribution with covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & 0 \\ \rho^2 & 0 & 1 \end{bmatrix}$$

Show that the conditional distribution of (X_1, X_2) given $X_3 = x_3$ is also MVN with mean

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 + \rho^2\left(x_3 - \mu_3\right) \\ \mu_2 \end{bmatrix}$$

and covariance matrix

$$\begin{bmatrix} 1 - \rho^4 & \rho \\ \rho & 1 \end{bmatrix}$$

Solution

Let $\mathbf{Y}_1 = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ and $\mathbf{Y}_2 = (X_3)$ then

$$\begin{aligned} \mathbb{E}\mathbf{Y}_1 &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\ \mathbb{E}\mathbf{Y}_2 &= (\mu_3)\,. \end{aligned}$$

We have $Cov \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}$ where

$$\begin{aligned} \mathbf{\Sigma}_{11} &= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \\ \mathbf{\Sigma}_{12} &= \begin{bmatrix} \rho^2 \\ 0 \end{bmatrix} = \mathbf{\Sigma}_{21}^T \\ \mathbf{\Sigma}_{22} &= [1] \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{E}\left[\mathbf{Y}_1 | \mathbf{Y}_2 = x_3\right] &= \boldsymbol{\mu}_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\left(x_3 - \mu_3\right) \\ &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \rho^2 \\ 0 \end{bmatrix} \left(\boldsymbol{x}_3 - \boldsymbol{\mu}_3\right) \\ &= \begin{bmatrix} \mu_1 + \rho^2\left(x_3 - \mu_3\right) \\ \mu_2 \end{bmatrix} \end{aligned}$$

and .

$$\begin{aligned} Cov\left[\mathbf{Y}_1 | \mathbf{Y}_2 = x_3\right] &= \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} \\ &= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} - \begin{bmatrix} \rho^2 \\ 0 \end{bmatrix} \begin{bmatrix} \rho^2 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 - \rho^4 & \rho \\ \rho & 1 \end{bmatrix} \end{aligned}$$

3.3 Maximum-likelihood estimation

Let $\mathbf{X}^T = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ contain an independent random sample of size n from $N_p\left(\boldsymbol{\mu}, \mathbf{\Sigma}\right)$.

The maximum likelihood estimates (MLE 's) of $\boldsymbol{\mu}, \mathbf{\Sigma}$ are the sample mean and covariance matrix (with divisor n)

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} \tag{3.6a}$$

$$\hat{\mathbf{\Sigma}} = \boldsymbol{S} \tag{3.6b}$$

The likelihood function is a function of the parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ given the data $\boldsymbol{X}$

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{X}) = \prod_{r=1}^{n} f(\boldsymbol{x}_r|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3.7}$$

The RHS is evaluated by substituting the individual data vectors $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ in turn into the p.d.f. of $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and taking the product.

$$\begin{aligned}\prod_{r=1}^{n} f(\boldsymbol{x}_r|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Sigma}|^{-n/2} \\ &\quad \exp\left\{-\frac{1}{2}\sum_{r=1}^{n}(\boldsymbol{x}_r - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_r - \boldsymbol{\mu})\right\}\end{aligned}$$

Maximizing L is equivalent to *minimizing* the "log likelihood" function

$$\begin{aligned} l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -2\log L \\ &= -2\sum_{r=1}^{n}\log f(\boldsymbol{x}_r|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= K + n\log|\boldsymbol{\Sigma}| + \sum_{r=1}^{n}(\boldsymbol{x}_r - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_r - \boldsymbol{\mu}) \end{aligned} \tag{3.8}$$

where K is a constant independent of $\boldsymbol{\mu}, \boldsymbol{\Sigma}$.

Result 3.3

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = n\left\{\log|\boldsymbol{\Sigma}| + tr\left[\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{S} + \boldsymbol{d}\boldsymbol{d}^T\right)\right]\right\} \tag{3.9}$$

up to an additive constant, where $\boldsymbol{d} = \bar{\boldsymbol{x}} - \boldsymbol{\mu}$.

Proof

Noting that $\boldsymbol{x}_r - \boldsymbol{\mu} = (\boldsymbol{x}_r - \bar{\boldsymbol{x}}) + \boldsymbol{d}$ the final term in the likelihood expression (3.8) becomes

$$\begin{aligned} &\sum_{r=1}^{n}(\boldsymbol{x}_r - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_r - \boldsymbol{\mu}) \\ = &\sum_{r=1}^{n}(\boldsymbol{x}_r - \bar{\boldsymbol{x}})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_r - \bar{\boldsymbol{x}}) + n\boldsymbol{d}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{d} \\ = &\; ntr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right) + n\boldsymbol{d}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{d} \\ = &\; ntr\left[\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{S} + \boldsymbol{d}\boldsymbol{d}^T\right)\right] \end{aligned}$$

proving the expression (3.9). Note that the cross-product terms have vanished because $\sum_{r=1}^{n}\boldsymbol{x}_r =$

$n\bar{\boldsymbol{x}}$ and therefore

$$\begin{aligned}\sum_{r=1}^{n} \boldsymbol{d}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_r - \bar{\boldsymbol{x}}) &= \boldsymbol{d}^T \boldsymbol{\Sigma}^{-1} \sum_{r=1}^{n} (\boldsymbol{x}_r - \bar{\boldsymbol{x}}) \\ &= \sum_{r=1}^{n} (\boldsymbol{x}_r - \bar{\boldsymbol{x}})^T \boldsymbol{\Sigma}^{-1} \boldsymbol{d} \\ &= 0\end{aligned}$$

In (3.9) the dependence on $\boldsymbol{\mu}$ is entirely through $\boldsymbol{d}$. Now assume that is positive definite (p.d.), then so is $\boldsymbol{\Sigma}^{-1}$ as

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{V}\boldsymbol{\Lambda}^{-1}\ \boldsymbol{V}^T$$

where $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$ is the eigenanalysis of $\boldsymbol{\Sigma}$. Thus $\forall \boldsymbol{d} \neq \boldsymbol{0}$ we have $\boldsymbol{d}^T \boldsymbol{\Sigma}^{-1}\ \boldsymbol{d} > 0$. Hence $l(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is minimized with respect to $\boldsymbol{\mu}$ for fixed $\boldsymbol{\Sigma}$ when $\boldsymbol{d} = 0$ i.e.

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$$

Final part of proof: to minimize the log-likelihood $l(\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ w.r.t. $\boldsymbol{\Sigma}$ let

$$\begin{aligned}l(\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) &= n\left\{\log|\boldsymbol{\Sigma}| + tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right)\right\} \\ &= \Phi(\boldsymbol{\Sigma}) \qquad (3.10)\end{aligned}$$

We show that

$$\begin{aligned}\Phi(\boldsymbol{\Sigma}) - \Phi(\boldsymbol{S}) &= n\left\{\log|\boldsymbol{\Sigma}| - \log|\boldsymbol{S}| + tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right) - p\right\} \\ &= n\left\{tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right) - \log|\boldsymbol{\Sigma}^{-1}\boldsymbol{S}| - p\right\} \qquad (3.11) \\ &\geq 0\end{aligned}$$

Lemma 1

$\boldsymbol{\Sigma}^{-1}\boldsymbol{S}$ is positive semi-definite (proved elsewhere). Therefore the eigenvalues of $\boldsymbol{\Sigma}^{-1}\boldsymbol{S}$ are positive.

Lemma 2

For any set of positive numbers

$$A \geq \log G + 1$$

where A and G are the arithmetic, geometric means respectively.

Proof

For all x we have $e^x \geq 1 + x$ (simple exercise).Consider a set of n strictly positive numbers $\{y_i\}$

$$\begin{aligned}
y_i &\geq 1 + \log y_i \\
\sum y_i &\geq n + \sum \log y_i \\
A &\geq 1 + \log \left(\prod y_i \right)^{\frac{1}{n}} \\
&= 1 + \log G
\end{aligned}$$

as required.

Recall that for any $(n \times n)$ matrix $\boldsymbol{A}$, we have $tr\left(\boldsymbol{A}\right) = \sum_{i=1}^{n} \lambda_i$ the sum of the eigenvalues, and $|\,\boldsymbol{A}| = \prod \lambda_i$ the product of the eigenvalues. Let λ_i $(i = 1, ..., p)$ be the positive eigenvalues of $\boldsymbol{\Sigma}^{-1}\boldsymbol{S}$ and substitute in (3.11)

$$\begin{aligned}
\log |\boldsymbol{\Sigma}^{-1}\boldsymbol{S}| &= \log \left(\prod \lambda_i \right) \\
&= p \log G
\end{aligned}$$

$$\begin{aligned}
tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right) &= \sum \lambda_i \\
&= pA
\end{aligned}$$

Hence

$$\begin{aligned}
\Phi\left(\boldsymbol{\Sigma}\right) - \Phi\left(\boldsymbol{S}\right) &= np \left\{ A - \log G - 1 \right\} \\
&\geq 0
\end{aligned}$$

This proves that the MLE's are as stated in (3.6).

3.3 Sampling distribution of $\bar{\mathbf{x}}$ and S

The Wishart distribution (Definition)

If $\boldsymbol{M}$ $(p \times p)$ can be written $\boldsymbol{M} = \boldsymbol{X}^T\boldsymbol{X}$ where $\boldsymbol{X}$ $(m \times p)$ is a data matrix from $N_p\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$ then $\boldsymbol{M}$ is said to have a Wishart distribution with scale matrix $\boldsymbol{\Sigma}$ and degrees of freedom m. We write

$$\boldsymbol{M} \sim W_p\left(\boldsymbol{\Sigma}, m\right) \tag{3.12}$$

When $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ the distribution is said to be in standard form.

Note:

The Wishart distribution is the multivariate generalization of the chi-square χ^2 distribution

Additive property of matrices with a Wishart distribution

Let $\boldsymbol{M}_1$, $\boldsymbol{M}_2$ be matrices having the Wishart distribution

$$\begin{aligned}\boldsymbol{M}_1 &\sim W_p(\boldsymbol{\Sigma}, m_1) \\ \boldsymbol{M}_2 &\sim W_p(\boldsymbol{\Sigma}, m_2)\end{aligned}$$

independently, then

$$\boldsymbol{M}_1 + \boldsymbol{M}_2 \sim W_p(\boldsymbol{\Sigma}, m_1 + m_2)$$

This property follows from the definition of the Wishart distribution because data matrices are additive in the sense that if

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}$$

is a combined data matrix consisting of $m_1 + m_2$ rows then

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{X}_1^T\boldsymbol{X}_1 + \boldsymbol{X}_2^T\boldsymbol{X}_2$$

is matrix (known as the "Gram matrix") formed from the combined data matrix $\boldsymbol{X}$.

Case of $p = 1$

When $p = 1$ we know from the definition of χ_r^2 as the distribution of the sum of squares of r independent $N(0, 1)$ variates that

$$\boldsymbol{M} = \sum_{i=1}^{m} x_i^2 \sim \sigma^2\chi_m^2$$

so that

$$W_1(\sigma^2, m) \equiv \sigma^2\chi_m^2$$

Sampling distributions

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$ be a random sample of size n from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

1. The sample mean $\bar{\boldsymbol{x}}$ has the normal distribution

$$\bar{\boldsymbol{x}} \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right)$$

2. The (scaled) sample covariance matrix has the Wishart distribution:

$$(n-1)\boldsymbol{S}_u \sim W_p(\boldsymbol{\Sigma}, n-1)$$

3. The distributions of $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}_u$ are independent.

3.4 Estimators for special circumstances

3.4.1 μ proportional to a given vector

Sometimes $\boldsymbol{\mu}$ is known to be proportional to a given vector, so $\boldsymbol{\mu} = k\boldsymbol{\mu}_0$ with $\boldsymbol{\mu}_0$ being a known vector.

For example if $\boldsymbol{x}$ represents a sample of repeated measurements then $\boldsymbol{\mu} = k\mathbf{1}$where $\mathbf{1} = (1, 1, ..., 1)^T$ is the $p-$vector of $1's$.

We find the MLE of k for this situation. Suppose $\boldsymbol{\Sigma}$ is known and $\boldsymbol{\mu} = k\boldsymbol{\mu}_0$. Let $\boldsymbol{d}_0 = \bar{\boldsymbol{x}} - k\boldsymbol{\mu}_0$. The log likelihood is

$$
\begin{aligned}
l(k) &= -2\log L \\
&= n\left[\log|\boldsymbol{\Sigma}| + \ tr\ \left\{\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{S} + \boldsymbol{d}_0\boldsymbol{d}_0^T\right)\right\}\right] \\
&= n\left[\log|\boldsymbol{\Sigma}| + \ tr\ \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right) + (\bar{\boldsymbol{x}} - k\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - k\boldsymbol{\mu}_0)\right] \\
&= n\left[\bar{\boldsymbol{x}}^T\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}} - 2k\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}} + k^2\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0\right] \\
&\quad + \text{constant terms indept of } k
\end{aligned}
$$

Set $\dfrac{dl}{dk} = 0$ to minimize $l(k)$ w.r.t. k

$$-2\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}} + 2\left(\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0\right)k = 0$$

from which

$$\hat{k} = \frac{\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}}}{\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0} \tag{3.13}$$

Properties

We now show that $\hat{k}$ is an unbiased estimator of k and determine the variance of $\hat{k}$
In (3.13) $\hat{k}$ takes the form $\dfrac{1}{\alpha}\boldsymbol{c}^T\bar{\boldsymbol{x}}$ with $\boldsymbol{c}^T = \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}$ and $\alpha = \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0$ so

$$
\begin{aligned}
\mathbb{E}\left[\hat{k}\right] &= \frac{\boldsymbol{c}^T\mathbb{E}[\bar{\boldsymbol{x}}]}{\alpha} \\
&= \frac{k\boldsymbol{c}^T\boldsymbol{\mu}_0}{\alpha}. \\
&= \frac{k\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0}{\alpha}
\end{aligned}
$$

since $\mathbb{E}[\bar{\boldsymbol{x}}] = k\boldsymbol{\mu}_0$. Hence

$$\mathbb{E}\left[\hat{k}\right] = k \tag{3.14}$$

showing that $\hat{k}$ is an unbiased estimator.

Note that $Var\left[\bar{\boldsymbol{x}}\right] = \frac{1}{n}\boldsymbol{\Sigma}$ and therefore that $Var\left[\boldsymbol{c}^T\bar{\boldsymbol{x}}\right] = \frac{1}{n}\boldsymbol{c}^T\boldsymbol{\Sigma}\boldsymbol{c}$ we have

$$\begin{aligned} Var\left(\hat{k}\right) &= \frac{1}{n\alpha^2}\boldsymbol{c}^T\boldsymbol{\Sigma}\boldsymbol{c} \\ &= \frac{1}{n}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0\left(\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0\right)^{-2} \\ &= \frac{1}{n\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0} \end{aligned} \tag{3.15}$$

3.4.2 Linear restriction on μ

We determine an estimator for $\boldsymbol{\mu}$ to satisfy a linear restriction

$$\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{b}$$

where $\boldsymbol{A}$ $(m \times p)$ and $\boldsymbol{b}$ $(m \times 1)$ are given constants and $\boldsymbol{\Sigma}$ is assumed to be known.

We write the restriction in vector form $\boldsymbol{g}(\boldsymbol{\mu}) = \boldsymbol{0}$ and form the Lagrangean

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\lambda}) = l(\boldsymbol{\mu}) + 2\boldsymbol{\lambda}^T\boldsymbol{g}(\boldsymbol{\mu})$$

where $\boldsymbol{\lambda}^T = (\lambda_1, ..., \lambda_m)$ is a **vector** of Lagrange multipliers (the factor 2 is inserted just for convenience).

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\lambda}) &= l(\boldsymbol{\mu}) + 2\boldsymbol{\lambda}^T(\boldsymbol{A}\boldsymbol{\mu} - \boldsymbol{b}) \\ &= n\left\{(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) + 2\boldsymbol{\lambda}^T(\boldsymbol{A}\boldsymbol{\mu} - \boldsymbol{b})\right\} \\ &\quad \text{ignore constant terms involving } \boldsymbol{\Sigma} \end{aligned}$$

Set $\frac{d}{d\boldsymbol{\mu}}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \boldsymbol{0}$ using results from Example Sheet 2:

$$\begin{aligned} -2\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) + 2\boldsymbol{A}^T\boldsymbol{\lambda} &= \boldsymbol{0} \\ \bar{\boldsymbol{x}} - \boldsymbol{\mu} &= \boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{\lambda} \end{aligned} \tag{3.16}$$

We use the constraint $\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{b}$ to evaluate the Lagrange multipliers $\boldsymbol{\lambda}$. Premultiply by $\boldsymbol{A}$

$$\begin{aligned} \boldsymbol{A}\bar{\boldsymbol{x}} - \boldsymbol{b} &= \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{\lambda} \\ \boldsymbol{\lambda} &= \left(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\right)^{-1}(\boldsymbol{A}\bar{\boldsymbol{x}} - \boldsymbol{b}) \end{aligned}$$

Substitute into (3.16)

$$\boxed{\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} - \boldsymbol{\Sigma}\boldsymbol{A}^T\left(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\right)^{-1}(\boldsymbol{A}\bar{\boldsymbol{x}} - \boldsymbol{b})} \tag{3.17}$$

3.4.3 Covariance matrix Σ proportional to a given matrix

We consider estimating k when $\boldsymbol{\Sigma} = k\boldsymbol{\Sigma}_0$, where $\boldsymbol{\Sigma}_0$ is a given.constant matrix. The likelihood (3.8) takes the form when $\boldsymbol{d} = \boldsymbol{0}$ ($\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$)

$$l(k) = n\left\{\log|k\boldsymbol{\Sigma}_0| + tr\left(\frac{1}{k}\boldsymbol{\Sigma}_0^{-1}\mathbf{S}\right)\right\}$$

plus constant terms (not involving k).

$$\begin{aligned} l(k) &= \left\{p\log k + \frac{1}{k}tr\left(\boldsymbol{\Sigma}_0^{-1}\mathbf{S}\right)\right\} \\ &\quad + \text{constant terms} \\ \frac{dl}{dk} &= 0 \Rightarrow \frac{p}{k} - \frac{1}{k^2}tr\left(\boldsymbol{\Sigma}_0^{-1}\mathbf{S}\right) = 0 \end{aligned}$$

Hence

$$\boxed{\hat{k} = \frac{tr\left(\boldsymbol{\Sigma}_0^{-1}\mathbf{S}\right)}{p}} \tag{3.18}$$