Chapter 3: Population mean & covariance and Multivariate normal distribution

Zhixiang ZHANG

University of Macau

zhixzhang@um.edu.mo

January 16, 2025

Table of Contents

Random samples from a distribution

- In order to study the sampling variability of statistics like $\bar{\mathbf{x}}$ and $\mathbf{S}_n$ with the ultimate aim of making inferences, we need to make assumptions about the variables whose observed values constitute the data set $\mathbf{X}$.
- Suppose, then, that the data have not yet been observed, but we intend to collect n sets of measurements on p variables. Consequently, we treat them as random variables.
- A random sample can now be defined. If the row vectors $\mathbf{x}_1^\top, \mathbf{x}_2^\top, \ldots, \mathbf{x}_n^\top$ in $\mathbf{X} \in \mathbb{R}^{n \times p}$ represent independent observations drawn from a common joint distribution with density function $f(\mathbf{x}) = f(x_1, x_2, \ldots, x_p)$, then $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are said to form a *random sample* from $f(\mathbf{x})$.

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} \qquad X_i \in \mathbb{R}^p \sim f(\vec{x}) \quad (F(x))$$

Expectation of random matrices(vectors)

- A random matrix/vector is a matrix/vector whose elements are random variables.
- Let $\mathbf{X} = \{X_{ij}\}$ be an $n \times p$ random matrix. Then the expected value of $\mathbf{X}$, denoted by $E(\mathbf{X})$, is the $n \times p$ matrix of numbers (if they exist)

$$E(\mathbf{X}) = \begin{bmatrix} E(X_{11}) & E(X_{12}) & \cdots & E(X_{1p}) \\ E(X_{21}) & E(X_{22}) & \cdots & E(X_{2p}) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_{n1}) & E(X_{n2}) & \cdots & E(X_{np}) \end{bmatrix}.$$

- Let $\mathbf{X}$ and $\mathbf{Y}$ be random matrices of the same dimension, and let $\mathbf{A}$ and $\mathbf{B}$ be conformable matrices of constants. Then

$$E(\mathbf{X} + \mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y})$$

$E[\text{row}_i A \times X \times \text{column}_i B]$

$= E[a_1^T X b_1]$

A is fixed matrix

$$E(\mathbf{AXB}) = \mathbf{A}E(\mathbf{X})\mathbf{B}$$

$E(AXX^TA^T) = AE(XX^T)A^T \neq AE(X)E(X^T)A^T$ because X and X^T is correlated

$E(z^2) = (E(z))^2 \Leftrightarrow Var(z) = 0$

Population mean and variance

The marginal mean and variance of a random vector $\mathbf{X}^\top = [X_1, X_2, \cdots, X_p]$:

$$\mu_i = \begin{cases} \int_{-\infty}^{\infty} x_i f_i(x_i)\, dx_i & \text{if } X_i \text{ is a continuous random variable with probability density function } f_i(x_i) \\ \sum_{\text{all } x_i} x_i p_i(x_i) & \text{if } X_i \text{ is a discrete random variable with probability function } p_i(x_i) \end{cases}$$

$$\sigma_i^2 = \begin{cases} \int_{-\infty}^{\infty} (x_i - \mu_i)^2 f_i(x_i)\, dx_i & \text{if } X_i \text{ is a continuous random variable with probability density function } f_i(x_i) \\ \sum_{\text{all } x_i} (x_i - \mu_i)^2 p_i(x_i) & \text{if } X_i \text{ is a discrete random variable with probability function } p_i(x_i) \end{cases}$$

It will be convenient later to denote the marginal variances by σ_{ii}.

Population covariance

The behavior of any pair of random variables, such as X_i and X_k s described by their joint probability function, and a measure of the linear association between them is provided by the covariance:

$$\sigma_{ik} = E(X_i - \mu_i)(X_k - \mu_k)$$

$$= \begin{cases} \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_i - \mu_i)(x_k - \mu_k) f_{ik}(x_i, x_k)\,dx_i\,dx_k & \text{if } X_i, X_k \text{ are continuous random variables with the joint density function } f_{ik}(x_i, x_k) \\ \displaystyle\sum_{\text{all } x_i}\sum_{\text{all } x_k} (x_i - \mu_i)(x_k - \mu_k) p_{ik}(x_i, x_k) & \text{if } X_i, X_k \text{ are discrete random variable with joint probability function } p_{ik}(x_i, x_k) \end{cases} \tag{2-26}$$

Independence

- More generally, the collective behavior of the p random variables $X_1, X_2, \ldots, X_p$ is described by a joint probability density function $f(x_1, x_2, \ldots, x_p) = f(\mathbf{x})$. ($f(\mathbf{x})$ will often be the multivariate normal density function in this course)
- If

$$P\left[X_i \leq x_i \text{ and } X_k \leq x_k\right] = P\left[X_i \leq x_i\right] P\left[X_k \leq x_k\right]$$

for all pairs of values x_i, x_k, then X_i and X_k are said to be *independent*. When X_i and X_k are continuous random variables with joint density $f_{ik}\left(x_i, x_k\right)$ and marginal densities $f_i\left(x_i\right)$ and $f_k\left(x_k\right)$, the independence condition becomes

$$f_{ik}\left(x_i, x_k\right) = f_i\left(x_i\right) f_k\left(x_k\right)$$

for all pairs (x_i, x_k).

Independence

two variables

$X_i \in R^{4\times1}$

$i=1$ and 2.

$$\begin{array}{cc} X_1 & X_2 \\ \end{array}$$
$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{pmatrix}$$

- The p continuous random variables $X_1, X_2, \ldots, X_p$ are *mutually* independent if their joint density can be factored as

$$f_{12\cdots p}(x_1, x_2, \ldots, x_p) = f_1(x_1) f_2(x_2) \cdots f_p(x_p)$$

for all p-tuples $(x_1, x_2, \ldots, x_p)$.

$Cov(X,Y) = E(X-E(X))(Y-E(Y))$

- Independence of X_i and X_k implies that $\text{Cov}(X_i, X_k) = 0$. However, the converse of is not true in general.

$X_1 = Z \sim N(0,1)$ $Cov(X_1, X_2) = 0 \not\Rightarrow X_1 \perp\!\!\!\perp X_2$

$X_2 = Z^2$

only linear independence

more comprehensive independence

if $X \perp\!\!\!\perp Y \Leftrightarrow E(f_1(X) - E(f_1(X)))(f_2(Y) - E(f_2(Y))) = 0$

$E(f_1(X)f_2(Y)) = E(f_1(X))E(f_2(Y))$

Population mean vec and population covariance mat

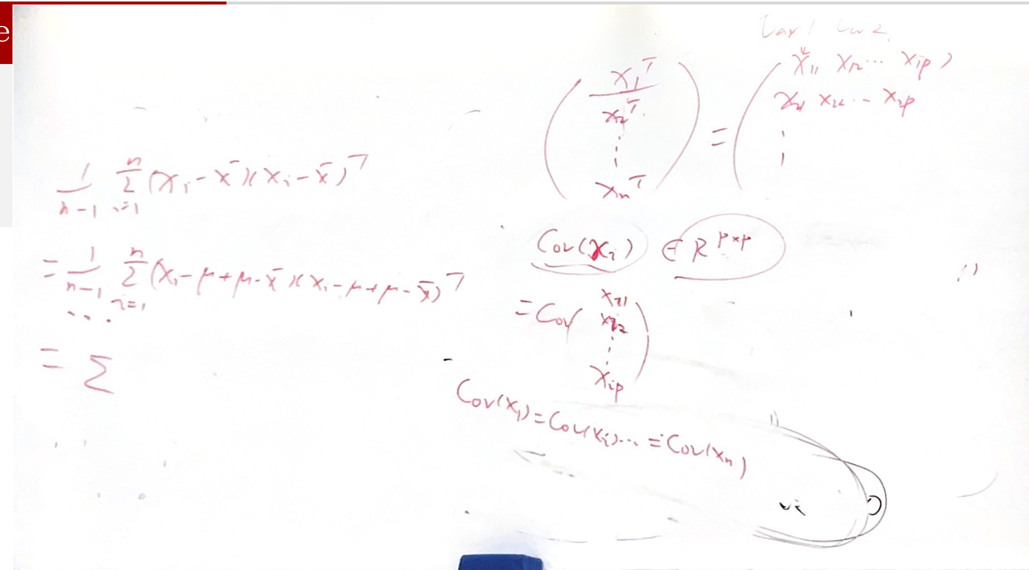
We shall refer to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ below as the *population mean* (vector) and *population covariance* (matrix), respectively:

$$\boldsymbol{\mu} = E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{\Sigma} &= E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top \\ &= E\left(\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{bmatrix} [X_1 - \mu_1, X_2 - \mu_2, \ldots, X_p - \mu_p] \right) \\ &= (\sigma_{ij})_{1 \le i,j \le p}. \end{aligned}$$

Here and below in this Chapter $\mathbf{X}$ refers to a random vector.

Unbiasdedness



If $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are random samples drawn from a distribution with mean $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, sample mean $\bar{\mathbf{x}}$ and sample covariance

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)^T$$

are unbiased estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively.

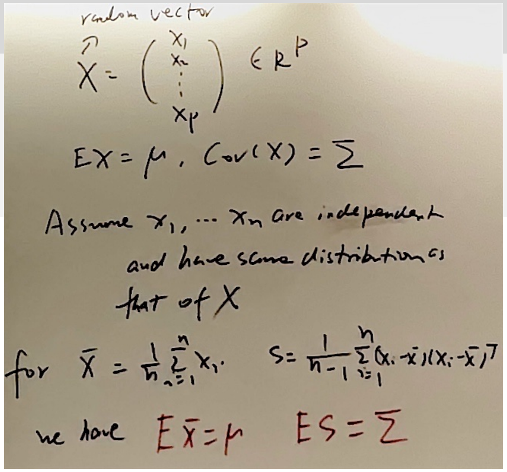
Correlation matrix

The *population correlation matrix*

$$\boldsymbol{\rho} = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{bmatrix} = \mathbf{D}_\sigma^{-1/2} \boldsymbol{\Sigma} \mathbf{D}_\sigma^{-1/2},$$

where $\mathbf{D}_\sigma = \text{diag}\left(\sigma_{11}, \ldots, \sigma_{pp}\right)$.
Correspondingly, the sample correlation matrix can be represented by $\mathbf{D}_s^{-1/2} \mathbf{S} \mathbf{D}_s^{-1/2}$ where $\mathbf{D}_s$ is the diagonal matrix of $\mathbf{S}$.

Linear combiniation of random vector



- The linear combination $\mathbf{c}^\top \mathbf{X} = c_1 X_1 + \cdots + c_p X_p$ has

$$\begin{aligned} \text{mean } &= E\left(\mathbf{c}^\top \mathbf{X}\right) = \mathbf{c}^\top \boldsymbol{\mu} \\ \text{variance } &= \text{Var}\left(\mathbf{c}^\top \mathbf{X}\right) = \mathbf{c}^\top \boldsymbol{\Sigma} \mathbf{c} \end{aligned}$$

  where $\boldsymbol{\mu} = E(\mathbf{X})$ and $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X})$.

- Given $\mathbf{x}_1, \cdots \mathbf{x}_n$ and the sample mean $\bar{\mathbf{x}}$ and sample covariance matrices $\mathbf{S}$. What are the sample mean and sample variance of $c^\top \mathbf{x}_1, \cdots, c^\top \mathbf{x}_n$?

• $X \in \mathbb{R}^p$ $c \in \mathbb{R}^p$ fixed find $E c^T X$ $Var(c^T X)$ Assume $E(X) = \mu$ and $Cov(X) = \Sigma$

r.v.

$E c^T X = c^T E(X) = c^T \mu$

c is not random

$Var(c^T X) = E(c^T X - E c^T X)(c^T X - E c^T X) = E(c^T X - E c^T X)(X^T c - E X^T c) = c^T \{E(X - EX)(X - EX)^T\} c = c^T \Sigma c$

$= \|c\|_2^2 \Sigma$

· by above, $E(c^T X_1) = c^T E(X_1) = c^T \bar{X}$

$Var(c^T X_1) = c^T Var(X_1) c = c^T S c = \|c\|_2^2 S$

for $i = 1$, similar for $i > 1$

$C^T X = \sum_{i=1}^n C_i X_i$

several combination

$$CX = \begin{pmatrix} C_1^T X \\ \vdots \\ C_i^T X \end{pmatrix}$$

$E CX = C E X = C\mu$

$Cov(CX) = C \Sigma C^T$

result

Linear combination of random vector

In general, consider the q linear combinations of the p random variables $X_1, \ldots, X_p$:

$$\mathbf{W} = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_q \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{q1} & c_{q2} & \cdots & c_{qp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \mathbf{CX}.$$

We have

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{W}} &= E(\mathbf{W}) = E(\mathbf{CX}) = \mathbf{C}\boldsymbol{\mu}_{\mathbf{X}}, \\ \boldsymbol{\Sigma}_{\mathbf{W}} &= \mathrm{Cov}(\mathbf{W}) = \mathrm{Cov}(\mathbf{CX}) = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{C}^\top, \end{aligned} \tag{1}$$

where $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$ are the mean vector and variance-covariance matrix of $\mathbf{X}$.

Example

Let $\mathbf{X} = [X_1, X_2]^\top$ be a random vector with mean vector $\boldsymbol{\mu}_{\mathbf{X}} = [\mu_1, \mu_2]^\top$ and variance-covariance matrix

$$\boldsymbol{\Sigma}_{\mathbf{X}} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

Find the covariance matrix for $\mathbf{Z} = [Z_1, Z_2]^\top$ where

$$Z_1 = X_1 - X_2, \quad Z_2 = X_1 + X_2.$$

Solution: We can write

$$\mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} := \mathbf{C}\mathbf{X}.$$

Therefore, the covariance matrix is given by

$$\begin{aligned} \mathrm{Cov}(\mathbf{Z}) = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{C}^\top &= \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{11} - 2\sigma_{12} + \sigma_{22} & \sigma_{11} - \sigma_{22} \\ \sigma_{11} - \sigma_{22} & \sigma_{11} + 2\sigma_{12} + \sigma_{22} \end{bmatrix} \end{aligned}$$

Cross-covariance

where $E(E(Y)^T) = E(Y)^T$

$E[XY^T - XE(Y)^T - E(X)Y^T + E(X)E(Y)^T]$

$= E(XY^T) - E(X)E(Y)^T$

- Cross-covariance of random vectors $\mathbf{X}$ and $\mathbf{Y}$ equals

$$\begin{aligned}\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) &= \mathrm{E}\left[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])^{\mathrm{T}}\right] \\ &= \mathrm{E}\left[\mathbf{X}\mathbf{Y}^{\mathrm{T}}\right] - \mathrm{E}[\mathbf{X}]\mathrm{E}[\mathbf{Y}]^{\mathrm{T}}.\end{aligned}$$

- Let $\mathbf{A}$, $\mathbf{B}$ be conformable matrices of constants, then

$$\mathrm{Cov}(\mathbf{AX}, \mathbf{BY}) = \mathbf{A}\,\mathrm{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}^\top.$$

- $\mathrm{Cov}(\mathbf{CX}) = \mathbf{C}\boldsymbol{\Sigma}_X\mathbf{C}^\top$ is a special case of the above.

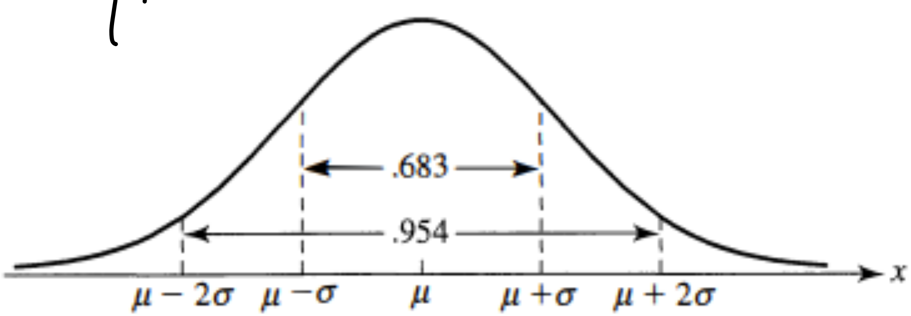
Table of Contents

(Univariate) normal distribution

Univariate Gaussian (normal) density:

$$
\begin{aligned}
f_X(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/\left(2\sigma^2\right)} \\
&= \frac{1}{(2\pi)^{\frac{1}{2}}\left(\sigma^2\right)^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)\frac{1}{\sigma^2}(x-\mu)}
\end{aligned}
$$

$P(X=a)=0$ in $f_X(x)$ is continuous or $X \sim N(\mu, \sigma^2)$



but here $P(\mu-\sigma \le \mu \le \mu+\sigma) = 0.683$

Figure 1: A normal density with mean μ and variance σ^2 and selected areas under the curve

which means the cumulative P is higher around μ

Multivariate normal density

Definition 1

A random vector $\mathbf{X}$ has a Multivariate Normal (MVN) distribution if it has a joint probability density function (pdf) of the form

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\underbrace{\boldsymbol{\Sigma}}_{\text{determinant}}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

mean and covariance

We denote this p-dimensional normal density by $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which is analogous to the normal density in the univariate case.

Special cases: diagonal $\boldsymbol{\Sigma}$

For p-variate case, if $\boldsymbol{\Sigma}$ is diagonal, then

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{\sigma_{11}} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{\sigma_{pp}} \end{bmatrix} \text{ and } |\boldsymbol{\Sigma}| = (\sigma_{11})(\sigma_{22})\cdots(\sigma_{pp})$$

$$\begin{aligned} \text{so} \quad f_{\mathbf{x}}(\mathbf{x}) =& \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{(\sigma_{11})\cdots(\sigma_{pp})}} \\ & \times \exp\left\{-\frac{1}{2}\frac{(x_1-\mu_1)^2}{\sigma_{11}} - \cdots - \frac{1}{2}\frac{(x_p-\mu_p)^2}{\sigma_{pp}}\right\} \\ =& f_{x_1}(x_1)\cdot f_{x_2}(x_2)\cdot\cdots\cdot f_{x_p}(x_p). \end{aligned}$$

they are independent to each other

Bivariate case

- For bivariate case,

$$\begin{pmatrix} \sigma_{11}^2 & \rho_{12}\sigma_{11}\sigma_{22} \\ \rho_{12}\sigma_{11}\sigma_{22} & \sigma_{22}^2 \end{pmatrix}$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} \\ \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{22} \end{pmatrix} \right].$$

- If

$$\rho_{12} = 0, f_{\mathbf{x}}\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = f_{x_1}(x_1) \cdot f_{x_2}(x_2).$$

③ $\sim N(\mu, \sigma^2)$

Shape of the density

Two bivariate distributions are shown below.

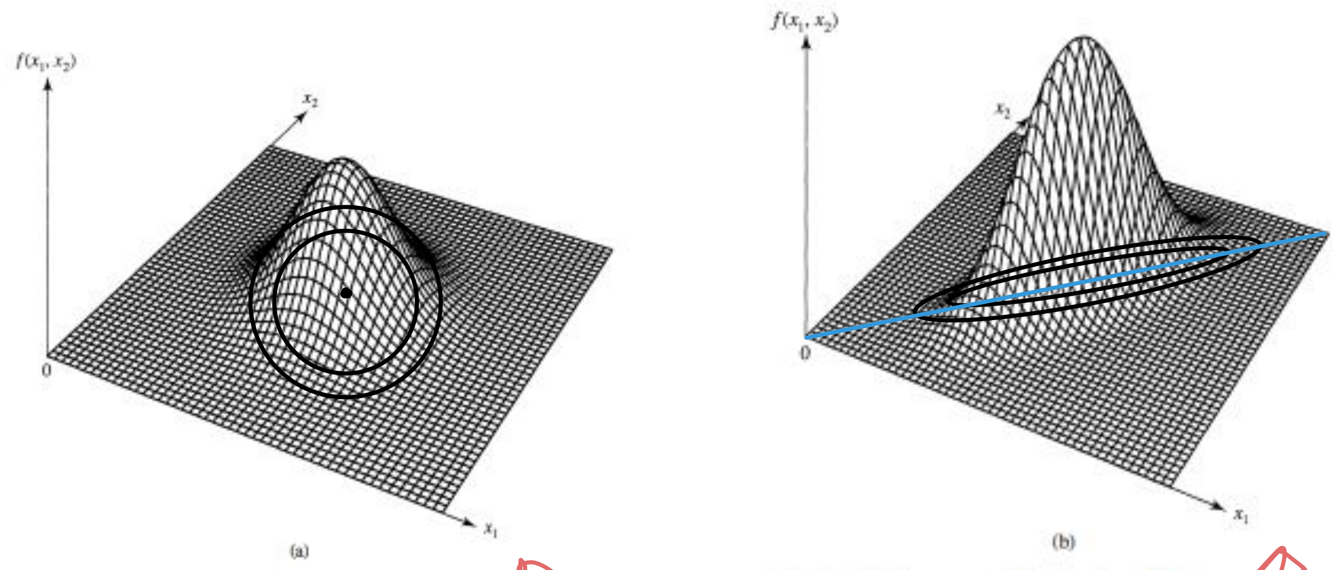


Figure 2: Two bivariate normal distributions. (a) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = 0$. (b) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = .75$.

peak height

$|\Sigma_1|^{\frac{1}{2}} \neq |\Sigma_2|^{\frac{1}{2}}$ $\quad \therefore f_2(x)_{max} > f_1(x)_{max}$

$\Sigma_1 = \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix}$ $\quad \Sigma_2 = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$ $\quad \therefore |\Sigma_2| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 < |\Sigma_1|$

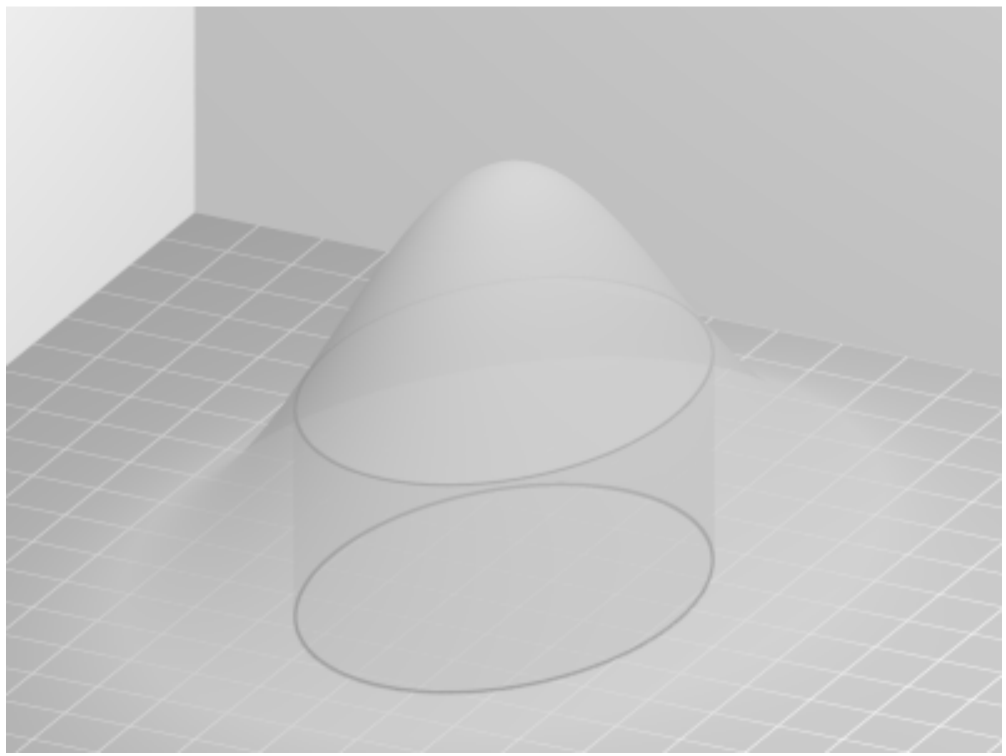
Constant density contour

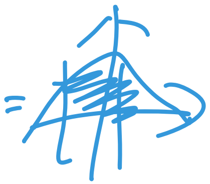
$$f(x) = \frac{1}{\sqrt{(2\pi)^p}\,|\Sigma|^{\frac{1}{2}}} e^{-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}} = d$$

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = -2\underbrace{\log\left(d\cdot\sqrt{(2\pi)^{\frac{p}{2}}}|\Sigma|^{\frac{1}{2}}\right)}_{\text{constant}}$$

constant

Values of $\mathbf{x}$ yielding constant height for density are ellipsoids.
Constant density contour $= \{\mathbf{x} : (\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2\}$



$p(x_1 \in [\ ]) =$ 

in univariate case
area represent P

in multivariate case
volume represent P

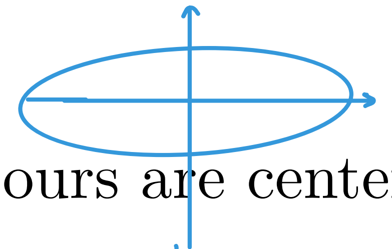
$P((x_1, x_2) \in O) =$ 

Figure 3: Constant density contour for bivariate normal

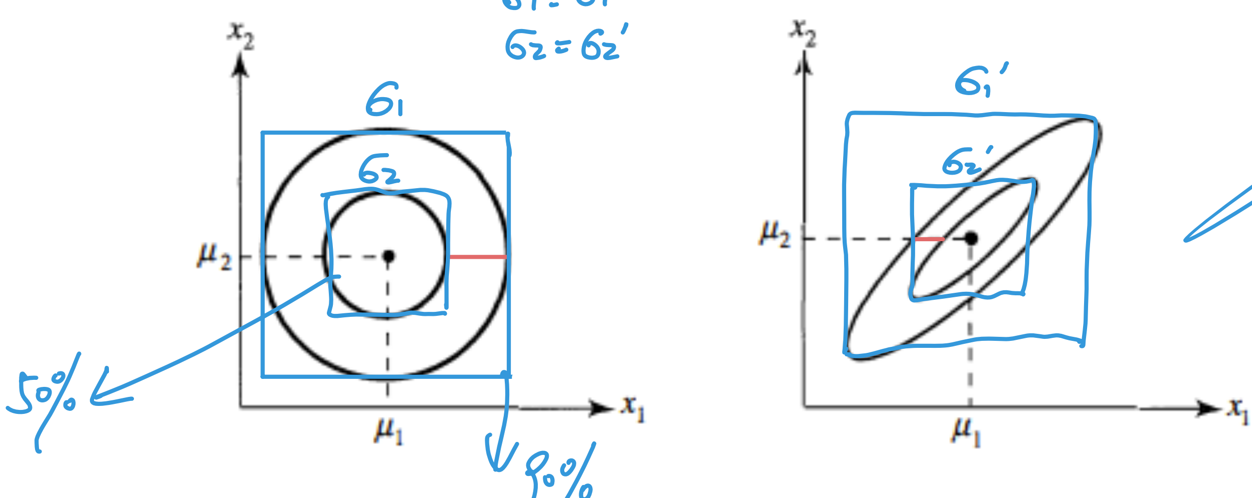
Constant density contour

$$\Sigma = \begin{pmatrix} \sigma_{11} & 0.9\sigma_{11} \\ 0.9\sigma_{11} & \sigma_{11} \end{pmatrix}$$

$\sigma_{11} >> \sigma_{22}$ and $\sigma_{12} = 0$



The constant density contours are centered at $\boldsymbol{\mu}$.



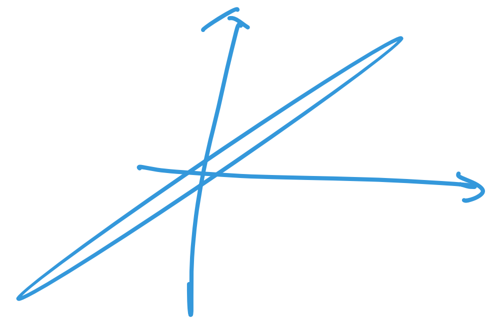


Figure 4: The 50% and 90% contours for the bivariate normal distributions in Figure 2

Some equivalent definitions of MVN

Besides specifying the density to define the multivariate normal distribution, we have some equivalent definitions:

Definition 2

$z_i \sim N(0,1)$

linear combination of Z will be multivariate normal

A random vector $\mathbf{X}$ has a multivariate normal distribution if there exist $\boldsymbol{\mu} \in \mathbb{R}^k$, $\boldsymbol{A} \in \mathbb{R}^{k \times \ell}$ such that $\mathbf{X} = \boldsymbol{A}\mathbf{Z} + \boldsymbol{\mu}$ where $\mathbf{Z}$ is a random vector whose components are independent $\mathcal{N}(0,1)$ random variables.

Definition 3

If for every a, $a^T C X$ is normal then CX is normal

because $a^T C X = (C^T a)^T X$ is normal by df. 3 and $X \sim MVN$

A random vector $\mathbf{X}$ has a multivariate normal distribution if for every real vector $\mathbf{a}$, the random variable $\mathbf{a}^\top \mathbf{X}$ is normal.

non-random

If $X \sim MVN$, then CX is MVN,

we can find A, μ, s.t. $X = AZ + \mu$ $\quad Z = \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix}$ z_i iid $N(0,1)$

$CX = CAZ + C\mu$ by def. 2 it is MVN

Some important properties for multivariate normal vectors

1. Linear combinations of the components of **X** are normally distributed.
2. All subsets of the components of **X** have a (multivariate) normal distribution. linear combination of original vectors
3. Zero covariance implies that the corresponding components are independently distributed.
4. The conditional distributions of the components are (multivariate) normal.

Property 1: Linear combinations of a normal vector is normal

If $X \sim N_p(\mu, \Sigma)$

then $CX+d \sim N_k(C\mu+d, C\Sigma C^T)$

If $\mathbf{X}$ is distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the q linear combinations

if $C \in \mathbb{R}^{k \times p}$

$$\underset{(q\times p)}{\mathbf{A}}\underset{(p\times 1)}{\mathbf{X}} = \begin{bmatrix} a_{11}X_1 + \cdots + a_{1p}X_p \\ a_{21}X_1 + \cdots + a_{2p}X_p \\ \vdots \\ a_{q1}X_1 + \cdots + a_{qp}X_p \end{bmatrix}$$

are distributed as $N_q\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top\right)$. Also, $\underset{(p\times 1)}{\mathbf{X}} + \underset{(p\times 1)}{\mathbf{d}}$, where $\mathbf{d}$ is a vector of constants, is distributed as $N_p(\boldsymbol{\mu} + \mathbf{d}, \boldsymbol{\Sigma})$.

Example

For $\mathbf{X} = (X_1, X_2, X_3)^\top$ distributed as $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, find the distribution of $(X_1 - X_2, X_2 - X_3)^\top$.

Example

For $\mathbf{X} = (X_1, X_2, X_3)^\top$ distributed as $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, find the distribution of $(X_1 - X_2, X_2 - X_3)^\top$.
We write

$$\begin{bmatrix} X_1 - X_2 \\ X_2 - X_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \mathbf{AX}$$

Example

Then the distribution of $\mathbf{AX}$ is multivariate normal with mean

$$\mathbf{A}\boldsymbol{\mu} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_1 - \mu_2 \\ \mu_2 - \mu_3 \end{bmatrix}$$

and covariance matrix

$$\begin{aligned} \mathbf{A\Sigma A}^\top &= \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{11} - 2\sigma_{12} + \sigma_{22} & \sigma_{12} + \sigma_{23} - \sigma_{22} - \sigma_{13} \\ \sigma_{12} + \sigma_{23} - \sigma_{22} - \sigma_{13} & \sigma_{22} - 2\sigma_{23} + \sigma_{33} \end{bmatrix} \end{aligned}$$

Property 2: Subsets of components of normal are normal

$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$ is MVN, then $\begin{pmatrix} X_3 \\ X_5 \end{pmatrix}$ is MVN.

because $\begin{pmatrix} X_3 \\ X_5 \end{pmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$

All subsets of $\mathbf{X}$ are normally distributed. If we respectively partition $\mathbf{X}$, its mean vector $\boldsymbol{\mu}$, and its covariance matrix $\boldsymbol{\Sigma}$ as

$Cov\begin{pmatrix} X_3 \\ X_5 \end{pmatrix} = \begin{bmatrix} \sigma_{33} & \sigma_{35} \\ \sigma_{35} & \sigma_{55} \end{bmatrix}$

$$\underset{(p\times 1)}{\mathbf{X}} = \begin{bmatrix} \underset{(q\times 1)}{\mathbf{X}_1} \\ \underset{((p-q)\times 1)}{\mathbf{X}_2} \end{bmatrix} \quad \underset{(p\times 1)}{\boldsymbol{\mu}} = \begin{bmatrix} \underset{(q\times 1)}{\boldsymbol{\mu}_1} \\ \underset{((p-q)\times 1)}{\boldsymbol{\mu}_2} \end{bmatrix}$$

and

$$\underset{(p\times p)}{\boldsymbol{\Sigma}} = \begin{bmatrix} \underset{(q\times q)}{\boldsymbol{\Sigma}_{11}} & \underset{(q\times(p-q))}{\boldsymbol{\Sigma}_{12}} \\ \underset{((p-q)\times q)}{\boldsymbol{\Sigma}_{21}} & \underset{((p-q)\times(p-q))}{\boldsymbol{\Sigma}_{22}} \end{bmatrix}$$

then $\mathbf{X}_1$ is distributed as $N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

Property 3

If $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is MVN, $CoV(X_1, X_2) = 0$, then $X_1 \perp\!\!\!\perp X_2$

because $\begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}$, joint density $= \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{(X-\mu)^T \Sigma^{-1} (X-\mu)}{2}} = \frac{1}{(2\pi)^{\frac{k}{2}} |\Sigma_1|} e^{-\frac{(x_1-\mu_1)^T \Sigma_1^{-1} (x_1-\mu_1)}{2}} \cdot \frac{1}{(2\pi)^{\frac{p-k}{2}} |\Sigma_2|} e^{-\frac{(x_2-\mu_2)^T \Sigma_2^{-1} \cdots}{2}}$

① here $|\Sigma| = |\Sigma_1| \cdot |\Sigma_2|$

② $(X-\mu)^T \Sigma^{-1} (X-\mu)$

$= (X_1-\mu_1, X_2-\mu_2) \begin{pmatrix} \Sigma_1^{-1} & \\ & \Sigma_2^{-1} \end{pmatrix} \begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \end{pmatrix}$

$= (X_1-\mu_1)^T \Sigma_1^{-1} (X_1-\mu_1) + (X_2-\mu_2)^T \Sigma_2^{-1} (X_2-\mu_2)$

$\Leftarrow \times$ generally not true. $\Rightarrow$

- If $\mathbf{X}_1 \in \mathbb{R}^{q_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{q_2}$ are independent, then $\text{Cov}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{0}$, a $q_1 \times q_2$ matrix of zeros.
- If $\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$ is $N_{q_1+q_2}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$, then $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent if and only if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.
- Note that if two normal vectors have a covariance of zero, it does not imply that they are independent. An counterexample is by taking $\mathbf{X}_2 = Z\mathbf{X}_1$ for some Z independent of $\mathbf{X}_1$ and has distribution $P(Z = \pm 1) = 1/2$.

If X_1 is MVN and X_2 is MVN and $CoV(X_1, X_2) = 0$ then $X_1 \perp\!\!\!\perp X_2$ ✗

$CoV\begin{pmatrix} X_1 \in \mathbb{R}^k \\ X_2 \in \mathbb{R}^{p-k} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$

$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}^T + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}^T$

Example

Let $\underset{(3\times 1)}{\mathbf{X}}$ be $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\vec{X} \in R^{3\times 1} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$ $Cov(X) = \Sigma$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

$Cov(X_1, X_2) = Cov(X_2, X_1) = 1 \neq 0$

Are X_1 and X_2 independent? What about (X_1, X_2) and X_3 ?

$$Cov\left(\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, X_3\right) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} Cov(X_1, X_3) \\ Cov(X_2, X_3) \end{pmatrix}$$

and (X_1, X_2) and X_3 are jointly normal because $\vec{X}$ is $N_3(\mu, \Sigma)$

Example

Let $\underset{(3\times 1)}{\mathbf{X}}$ be $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Are X_1 and X_2 independent? What about (X_1, X_2) and X_3 ?
X_1 and X_2 are not independent; (X_1, X_2) and X_3 are independent.

Property 4: Conditional distributions of components are normal

$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is MVN

then $X_1 | X_2 = x_2$ is MVN

$X_2 | X_1 = x_1$ is MVN

Let $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$ be distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$, and $|\boldsymbol{\Sigma}_{22}| > 0$. Then the conditional distribution of $\mathbf{X}_1$, given that $\mathbf{X}_2 = \mathbf{x}_2$, is normal and has

$$\text{Mean} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

$X_1 | X_2 = x_2$ $\quad E(\cdots)$

$EX = \begin{pmatrix} EX_1 \\ EX_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, $CovX = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$

and

$$\text{Covariance} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Let $Y = AX$ where $A = \begin{bmatrix} I_k & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_{p-k} \end{bmatrix}$, $Y = \begin{pmatrix} X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \\ X_2 \end{pmatrix} \overset{?}{\sim} N\left(\begin{pmatrix} \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}\right)$

$X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$ ✓

$X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 | X_2 = x_2$

then we know

$X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 | X_2 = x_2$

$E(Y_1)$

So $X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \perp\!\!\!\perp X_2$

$E[X_1 | X_2 = x_2] = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \Sigma_{12}\Sigma_{22}^{-1}x_2$

Distribution of Sample mean

For any $a \in R^p$, $a^T\bar{x} = \frac{1}{n}\sum_{i=1}^n a^T x_i$ is normal

So $\bar{x}$ is normal

$E\bar{x} = \mu$

$\begin{bmatrix} a \\ a \\ \vdots \\ a \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

- If $\mathbf{x}_i \overset{i.i.d.}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then for any positive integer n,

$Cov(\bar{x}) = Cov(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{n}\Sigma$

remain same

$$\sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \sim N_p(0, \boldsymbol{\Sigma}).$$

$\bar{x} - \mu \sim N_p(0, \frac{1}{n}\Sigma)$ $\therefore \sqrt{n}(\bar{x} - \mu) \sim N_p(0, \Sigma)$

(Proof hint: Use Definition 3, show that $\mathbf{c}^\top\bar{\mathbf{x}}$ is normal first.)

- For real data, the multivariate normal assumption may be too strong. Fortunately, the Central Limit Theorem implies that the above result still holds asymptotically as the sample size n approaches infinity.

Central Limit Theorem

(Multivariate) Central Limit Theorem

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be independent observations from any population with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Then

$$\sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \xrightarrow{d} N_p(0, \boldsymbol{\Sigma}) \text{ as } n \to \infty,$$

where convergence in distribution $\to^d$ means that for any set $\Omega \in \mathbb{R}^p$,

$$\lim_{n\to\infty} \mathbb{P}\left(S_n \in \Omega\right) = \mathbb{P}(Y \in \Omega)$$

for a random vector Y with normal distribution $N_p(0, \Sigma)$.

Why MVN

Why emphasis the MVN?

- Genuinely good population model for some natural phenomena
- Even for nonnormal data, MVN is often useful approximation - such as for inferences involving sample mean vectors, which are asymptotically normal due to CLT last page
- Theoretical elegance, such as
    - MVN is completely characterized by its mean and covariance
    - Linear transformations of MVN are normal
    - Uncorrelated components of MVN are independent

Maximum Likelihood Estimation

- Maximum likelihood estimation (MLE) selects parameter values that maximize the joint density of the observed data.

- In the next few pages, we derive the MLE for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for multivariate normal data $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- This optimization leads to the MLEs: $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$.

Maximum likelihood estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
Joint density:

$$
\begin{aligned}
f(\mathbf{x}_1, \ldots, \mathbf{x}_n) &= \prod_{i=1}^{n} f(\mathbf{x}_i) \\
&= \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\} \\
&= \frac{1}{(2\pi)^{\frac{np}{2}} |\boldsymbol{\Sigma}|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\sum_{i}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\} \\
&= L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = L(x_1, \cdots, x_n)
\end{aligned}
$$

Goal: Find values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that maximize the likelihood of observing $\mathbf{x}_1, \ldots, \mathbf{x}_n$. $\hat{\mu}, \hat{\Sigma} = \underset{\mu, \Sigma}{\operatorname{argmax}} L(x_1, \cdots, x_n)$

$\hat{\mu} = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$

Consider $\sum_{i=1}^n (X_i-\mu)^T \Sigma^{-1} (X_i-\mu)$ $(\operatorname{tr} AB = \operatorname{tr} BA)$

$= \operatorname{tr} \sum_{i=1}^n (X_i-\mu)^T \Sigma^{-1} (X_i-\mu) = \operatorname{tr} \sum_{i=1}^n \Sigma^{-1} (X_i-\mu)(X_i-\mu)^T$

$= \operatorname{tr} \sum_{i=1}^n \Sigma^{-1} (X_i-\bar{X}+\bar{X}-\mu)(\cdots)^T$

$= \operatorname{tr}\left[\sum_{i=1}^n \Sigma^{-1} (X_i-\bar{X})(X_i-\bar{X})^T + n\Sigma^{-1}(\bar{X}-\mu)(\bar{X}-\mu)^T\right]$

Use

$$\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right) = \operatorname{tr}\left\{\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\right\} = \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^\top\right\},$$

rewrite

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

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

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

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

So,

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{np}{2}} |\boldsymbol{\Sigma}|^{\frac{n}{2}}} \times$$
$$\exp\left\{ -\frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1} \left[\sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}}) (\mathbf{x}_i - \overline{\mathbf{x}})^\top + n(\overline{\mathbf{x}} - \boldsymbol{\mu})(\overline{\mathbf{x}} - \boldsymbol{\mu})^\top \right] \right) \right\}$$

Note that the value of $\boldsymbol{\mu}$ maximizing $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the value minimizing $\operatorname{tr}\left\{ n\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu})(\overline{\mathbf{x}} - \boldsymbol{\mu})^\top \right\}$. Since $\boldsymbol{\Sigma}^{-1}$ is positive definite,

$$\operatorname{tr}\left\{ n\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu})(\overline{\mathbf{x}} - \boldsymbol{\mu})^\top \right\} = n(\overline{\mathbf{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}) > 0$$

unless

$$\boldsymbol{\mu} = \overline{\mathbf{x}}.$$

Thus, the likelihood is maximized with respect to $\boldsymbol{\mu}$ at $\hat{\boldsymbol{\mu}} = \overline{\mathbf{x}}$.

It remains to maximize

$$L(\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} e^{-\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\overline{\mathbf{x}})(\mathbf{x}_j-\overline{\mathbf{x}})^\top\right)\right]/2}$$

over $\boldsymbol{\Sigma}$.

not important

Result: Given a $p \times p$ symmetric positive definite matrix $\mathbf{B}$ and a scalar $b > 0$, it follows that

$$\frac{1}{|\boldsymbol{\Sigma}|^{b}} e^{-\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{B}\right)/2} \leq \frac{1}{|\mathbf{B}|^{b}} (2b)^{pb} e^{-bp}$$

for all positive definite $\boldsymbol{\Sigma}$, with equality holding only for $\boldsymbol{\Sigma} = (1/2b)\mathbf{B}$.

By the above result with $b = n/2$ and $\mathbf{B} = \sum_{j=1}^{n}\left(\mathbf{x}_j - \overline{\mathbf{x}}\right)\left(\mathbf{x}_j - \overline{\mathbf{x}}\right)^\top$, the maximum occurs at $\hat{\boldsymbol{\Sigma}} = (1/n)\sum_{j=1}^{n}\left(\mathbf{x}_j - \overline{\mathbf{x}}\right)\left(\mathbf{x}_j - \overline{\mathbf{x}}\right)^\top = (n-1)\mathbf{S}/n$.