The multivariate normal distribution

Multivariate normal distribution has many advantageous and revealing properties.

There are several ways of introducing multivariate normal. One way is to directly define the joint density of multivariate normal, such as in Johnson and Wichern.

In this course we introduce multivariate normal via linear combinations of its univariate components.

First let's review the univariate normal distribution.

1 The univariate normal distribution

Definition of univariate standard normal random variable

Z is a standard normal random variable, denoted as $Z \sim N(0,1)$, if its density function is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad z \in \mathbb{R}.$$

(Notation: $\mathbb{R} = (-\infty, \infty)$. $z \in \mathbb{R}$ means z is a real number.) By convention, ϕ is used for the density instead of the generic f.

A common notation for the cumulative distribution function of standard normal is

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt, \quad z \in \mathbb{R}$$

Definition of univariate normal random variable

For any real constants a, b, if $Y = aZ + b$, that is, if Y is a (affine) linear transformation of the standard normal Z, then by the properties of expectation (via the properties of integration), the expectation and variance are

$$\mathbb{E}(Y) = a\,\mathbb{E}(Z) + b = b, \qquad var(Y) = a^2\, var(Z) = a^2.$$

By variable substitution, for $a \neq 0$, Y has the density function

$$f_Y(y) = \frac{1}{a}\phi\left(\frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi a^2}} e^{-\frac{(y-b)^2}{2a^2}}, \quad x \in \mathbb{R}.$$

Then Y is a normal random variable with mean b and variance a^2, denoted as $Y \sim N(b, a^2)$.

Remarks on univariate normal

- The degenerate case $a = 0$ yields $Y \sim N(b, 0)$, which represents the point mass distribution at b.
- A univariate normal distribution $N(\mu, \sigma^2)$ is uniquely determined by its mean μ and variance σ^2. Therefore the distribution of $Y = aZ + b$ with $Z \sim N(0,1)$ is completely and uniquely determined by the values of b and a^2.
- The cumulative distribution function of normal variable does not have closed form, it has to be expressed as integrals or infinite series of known functions (e.g., polynomials).
- If X, Y are independent normal, then $X + Y$ is normal. (Exercise)
  - Consequently, if $X_i, i = 1, \cdots, k$ are independent normal, then their sum $X_1 + \cdots + X_k$ is normal. This important facts will be used frequently.

  - Furthermore, the linear combination of independent normal $a_1X_1 + \cdots + a_kX_k$ is normal, since each a_iX_i is normal. In general a_i's and other constants used in this course are real (rather than complex).
- However if X, Y are normal but not independent, then $X + Y$ is not necessarily normal.

  An example

  Suppose $X \sim N(0,1)$. Let random variable $S = \pm 1$ with probability $\frac{1}{2}$ each, and S be independent of X. Such an S is a random sign. Let $Y = SX$, so

  $$Y = \begin{cases} X, & \text{if } S = 1, \\ -X, & \text{if } S = -1. \end{cases}$$

  Then

  $$\begin{aligned} \mathbb{P}(Y \le y) &= \mathbb{P}(Y \le y, S = 1) + \mathbb{P}(Y \le y, S = -1) \\ &= \mathbb{P}(Y \le y | S = 1)\,\mathbb{P}(S = 1) + \mathbb{P}(Y \le y | S = -1)\,\mathbb{P}(S = -1) \\ &= \mathbb{P}(X \le y | S = 1)\frac{1}{2} + P(-X \le y | S = -1)\frac{1}{2} \\ &= \mathbb{P}(X \le y)\frac{1}{2} + \mathbb{P}(-X \le y)\frac{1}{2} && since \quad X \perp\!\!\!\perp S \\ &= \mathbb{P}(X \le y)\frac{1}{2} + \mathbb{P}(X \le y)\frac{1}{2} && since \quad -X \sim N(0,1) \\ &= \mathbb{P}(X \le y) \end{aligned}$$

  Hence $Y \sim N(0,1)$. Therefore the bivariate vector (X, Y) is marginally normal, that is, each component on its own is normal. However the joint distribution of (X, Y) is not bivariate normal. In fact,

  $$X + Y = \begin{cases} 0, & \text{with probability } \frac{1}{2}, \\ 2X, & \text{with probability } \frac{1}{2}. \end{cases}$$

  So $0 < \mathbb{P}(X + Y = 0) < 1$, which means $X + Y$ is not a univariate normal. Consequently (X, Y) is not bivariate normal, because a linear combination of the components $(X + Y)$ is not of normal distribution.

  Remark $X + Y$ is a mixture of a discrete (Bernoulli) distribution and a continuous $(2X)$ distribution which is normal. Furthermore, X and Y are uncorrelated but not independent.

2 Multivariate normal distribution via linear combinations

In the following we give a definition of multivariate normal in terms of the linear combinations of its individual components, which is the familiar univariate normal. This definition provides a convenient tool to construct multivariate normal vectors, as shown in the examples.

Definition

A p-variate random vector $\mathbf{X} = (X_1, \cdots, X_p)'$ is **multivariate normal** if for any $\boldsymbol{b} \in \mathbb{R}^p$, the linear combination $\boldsymbol{b}'\mathbf{X} = b_1X_1 + \cdots + b_pX_p$ is univariate normal (including the degenerate univariate normal, the point mass). (*)

Examples

- ***p*-variate standard normal distribution.**

  Let $Z_1, \cdots, Z_p$ be independent univariate standard normal random variables, so $Z_i \sim N(0,1)$.

By the independence, the p-variate variable $\mathbf{Z} = (Z_1, \cdots, Z_p)$ has joint density

$$f_{\mathbf{Z}}(\boldsymbol{z}) = f_{Z_1}(z_1) f_{Z_2}(z_2) \cdots f_{Z_p}(z_p)$$
$$= \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi}} e^{-z_j^2/2} = \frac{1}{\sqrt{2\pi}} e^{-\sum_{j=1}^{p} z_j^2/2}$$
$$= \frac{1}{(2\pi)^{p/2}} e^{-\boldsymbol{z}'\boldsymbol{z}/2} = \frac{1}{(2\pi)^{p/2}} e^{-\|\boldsymbol{z}\|^2/2}$$

where

$$\|\boldsymbol{z}\| = \boldsymbol{z}'\boldsymbol{z} = \sqrt{z_1^2 + \cdots + z_p^2}$$

denotes the Euclidean norm of vector $\boldsymbol{z}$ in $\mathbb{R}^p$.

$\boldsymbol{Z}$ is of multivariate normal distribution, and is termed as the p-**variate standard normal**.

Proof. We show that the p vector $\boldsymbol{Z}$ with independent standard normal components is of p-variate normal distribution.
By definition (*), it is sufficient to show that any linear combination of the components of the p vector $\boldsymbol{Z}$ is a univariate normal random variable.
For any linear combination $\boldsymbol{b}'\mathbf{Z} = b_1 Z_1 + \cdots b_p Z_p$, since $b_j Z_j \sim N(0, b_j^2)$, $b_j Z_j$'s are independent for $j = 1, \cdots, p$. By the fact that sum of independent univariate normal variables is still an univariate normal variable, we can conclude that the linear combination $\boldsymbol{b}'\mathbf{Z}$ is of univariate normal. Since this is true for any $\boldsymbol{b} \in \mathbb{R}^p$, this proves that $\mathbf{Z}$ is of p-variate normal.

□

- Correlated bivariate normal constructed from linear combinations of independent standard normal.

  Let Z_1, Z_2 be independent standard normal random variables. Then the random vector

  $$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} Z_1 + 2Z_2 \\ 3Z_1 + 4Z_2 \end{bmatrix}$$

  is of **bivariate normal**, because any linear combination of the components X, Y is a linear combination of Z_1, Z_2:

  $$b_1 X + b_2 Y = b_1(Z_1 + 2Z_2) + b_2(3Z_1 + 4Z_2) = (b_1 + 3b_2)Z_1 + (2b_1 + 4b_2)Z_2,$$

  which is univariate normal, again from the normality of sums of independent univariate normal variables.

  Furthermore, because $cov(Z_1, Z_2) = 0$, the covariance

  $$cov(X, Y) = cov(Z_1 + 2Z_2, 3Z_1 + 4Z_2) = 3V(Z_1) + 8V(Z_2) = 11.$$

  Hence (X, Y) is of correlated bivariate normal distribution constructed from independent, uncorrelated standard normal variables.

  $$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 5 & 11 \\ 11 & 25 \end{bmatrix} \right)$$

- Uncorrelated bivariate normal from linear combinations of independent variables.

  If any pair of random variables X, Y are independent and of equal variance, then $X + Y$ and $X - Y$ are uncorrelated:

  $$cov(X + Y, X - Y) = var(X) + cov(X, Y) - cov(Y, X) - var(Y) = var(X) - var(Y) = 0$$

  (Note: the independence of X and Y is not needed here but is needed for next step.)

If in addition to equal-variance, both X and Y are of normal distributions, then the independence of X and Y yields that any linear combination $b_1(X+Y) + b_2(X-Y) = (b_1 + b_2)X + (b_1 - b_2)Y$ is of univariate normal. By the definition of multivariate normal, $(X + Y, X - Y)$ is bivariate normal, where its components $X + Y$ and $X - Y$ are uncorrelated (consequently independent, due to the normality property).

3 Important properties of multivariate normal distribution

Theorem (from p-variate normal to k-variate normal)

If $\boldsymbol{X} \in \mathbb{R}^p$ is p-variate normal, and $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{c}$, where A is a matrix of dimension $k \times p$ and $\boldsymbol{c}$ is a k-vector of constants, then $\boldsymbol{Y} \in \mathbb{R}^k$ is of k-variate normal distribution.

Proof. For any k-vector $\boldsymbol{b}$, we need to show that $\boldsymbol{b}'\boldsymbol{Y}$ is univariate normal. Write

$$\boldsymbol{b}'\boldsymbol{Y} = \boldsymbol{b}'A\boldsymbol{X} + \boldsymbol{b}'\boldsymbol{c} = \boldsymbol{b}^{*\prime}\boldsymbol{X} + \boldsymbol{c}^{*\prime}$$

where $\boldsymbol{b}^* = (\boldsymbol{b}'A)' = A'_{p\times k}\boldsymbol{b}_{k\times 1}$ is a vector in $\mathbb{R}^p$, and $\boldsymbol{c}^* = (\boldsymbol{b}'\boldsymbol{c})' = \boldsymbol{c}'\boldsymbol{b}$ is a constant in $\mathbb{R}$. $\boldsymbol{X}$ is p-variate normal, so any linear combination $\boldsymbol{b}^{*\prime}\boldsymbol{X}$ of the components of $\mathbf{X}$ is of univariate normal. Since any linear combination of the k component of $\boldsymbol{Y}$ is a linear combination of the components of $\boldsymbol{X}$ plus a constant, thus univariate normal, we conclude that $\boldsymbol{Y}$ is of k-variate normal. □

Remarks

- By the mean and variance formulas we proved in the previous notes on "Multivariate random sample matrices",

  $$\mathbb{E}(\boldsymbol{Y}) = A\mathbb{E}(\boldsymbol{X}) + \boldsymbol{c} = A\boldsymbol{\mu}_x + \boldsymbol{c}, \qquad Cov(\boldsymbol{Y}) = ACov(\boldsymbol{X})A' = A\Sigma_x A'$$

  Thus,

  $$\boldsymbol{Y} \sim N_k(A\boldsymbol{\mu}_x + \boldsymbol{c}, A\Sigma_x A') \tag{1}$$

- In writing up (1) the uniqueness of multivariate normal with given mean and variance is implicitly assumed. The theorem (uniqueness of multivariate normal) is stated near the end in this section with a brief outline of the proof.
- Each component of the k-variate $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{c}$ is a linear combination of the p-components of $\boldsymbol{X}$ plus a constant.
- In applications, usually $rank(A) = k \leq p$ to avoid degenerate cases.

Corollary (**Construct k-variate normal** with any mean and any legit covariance)

If $\boldsymbol{Z}$ is p-variate standard normal, and $\boldsymbol{Y} = A\boldsymbol{Z} + \boldsymbol{c}$, where matrix A is $k \times p$ and $\boldsymbol{c} \in \mathbb{R}^k$, then $\boldsymbol{Y}$ is k-variate normal with

$$E(\boldsymbol{Y}) = \boldsymbol{c}, \qquad Cov(\boldsymbol{Y}) = AA', \qquad \text{that is, } \boldsymbol{Y} \sim N(\boldsymbol{c}, AA')$$

Corollary (**Marginals of multivariate normal** are multivariate normal)

Suppose $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$. Write $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}$, where $\boldsymbol{X}_1$ consists of the first $q < p$ components of $\boldsymbol{X}$, so $\boldsymbol{X}_1$ is q-variate, consequently $\boldsymbol{X}_2$ is $(p - q)$-variate. Partition $\boldsymbol{\mu}$ and Σ accordingly,

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{2}$$

Then

$$\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \Sigma_{11}), \qquad \boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \Sigma_{22}).$$

Proof. We are to derive that $\boldsymbol{X}_1, \boldsymbol{X}_2$ are normal with the stated means and covariance. $E(\boldsymbol{X}_i) = \boldsymbol{\mu}_i$ and $Cov(\boldsymbol{X}_i) = \Sigma_{ii}$ $(i = 1, 2)$ can be verified directly. To show $\boldsymbol{X}_i$ is normal, we express each $\boldsymbol{X}_i$ as a linear transformation of $\boldsymbol{X}$.

$$\boldsymbol{X}_1 = [\boldsymbol{I}_q \ \boldsymbol{O}_{p-q}] \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} = \boldsymbol{A}_1 \boldsymbol{X}, \qquad \boldsymbol{X}_2 = [\boldsymbol{O}_p \ \boldsymbol{I}_{p-q}] \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} = \boldsymbol{A}_2 \boldsymbol{X}$$

where $\boldsymbol{I}$ and $\boldsymbol{O}$ are the identity matrix and the components $= 0$ matrix with the corresponding dimensions. Recall $\boldsymbol{AX} + \boldsymbol{c}$ is of multivariate normal for any $\boldsymbol{A}$ and $\boldsymbol{c}$, as formulated in (1). Therefore $\boldsymbol{X}_i = \boldsymbol{A}_i \boldsymbol{X}$ $(i = 1, 2)$ are of multivariate normal.

□

Remarks

- An immediate conclusion from the second corollary is that each component of $\boldsymbol{X}$ is univariate normal,

  $$X_j \sim N_1(\mu_j, \sigma_{jj})$$

  where σ_{jj} is the (j, j)th element or the ith diagonal entry of Σ.
- The converse is not necessarily true. If each component X_j's is normal, $\mathbf{X} = (X_1, \cdots, X_p)$ may not be p-variate normal, as illustrated by the example given earlier.
- For normal distributions, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent if and only if $Cov(\mathbf{X_1}, \mathbf{X_2}) = \mathbf{0}$, which can be derived from the form of the joint density function. (exercise)

Corollary (**Conditional distributions of multivariate normal** are multivariate normal)

Assume that $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \Sigma_{11}), \boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \Sigma_{22})$, and $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)'$ is p-variate normal with mean $\boldsymbol{\mu}$ and variance Σ of the partitioned form in (2). Then the conditional distribution of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is normal, with conditional mean and covariance as in the following.

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \quad \sim \quad N_q\left(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

The above result can be derived from the form of the conditional density.

Example: For $p = 2$, $q = 1$,

$$\boldsymbol{X} \ \sim \ N_2(\boldsymbol{\mu}, \Sigma), \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\ \sigma_1\sigma_2 \\ \rho\ \sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

The notations imply

$$X_1 \ \sim \ N(\mu_1, \sigma_1^2), \qquad X_2 \ \sim \ N(\mu_2, \sigma_2^2), \qquad corr(X_1, X_2) = \rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$$

The conditional distribution of X_1 given that $X_2 = x_2$ has the more familiar expression

$$X_1 \mid X_2 = x_2 \ \sim \ N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2),\ \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}\right) = N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2),\ \sigma_1^2\left(1 - \rho^2\right)\right)$$

Theorem* (**uniqueness of multivariate normal**)*

For and vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and symmetric positive definite matrix Σ of dimension $p \times p$, there exists a unique multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

- The uniqueness is a subtle key step in introducing multivariate normal via (affine) linear transformation.
- To prove the uniqueness, we would need to show that if $\boldsymbol{X}$, $\boldsymbol{Y}$ are p-variate normal with the same mean and covariance matrix, then $\boldsymbol{X}$ and $\boldsymbol{Y}$ are of the same distribution.
- The proof may start with the fact that linear transformation transforms normal from one dimension to another. So we may write $\boldsymbol{X} = A\boldsymbol{Z} + \boldsymbol{c}$, $\boldsymbol{Y} = B\boldsymbol{W} + \boldsymbol{d}$, where $\boldsymbol{Z}, \boldsymbol{W}$ are multivariate standard normal.
- $\boldsymbol{c} = E(\boldsymbol{X}) = E(\boldsymbol{Y}) = \boldsymbol{d}$. WLOG we may only consider the case $\boldsymbol{c} = \boldsymbol{d} = 0$.
- We only consider the case the A and B both are $p \times k$ (otherwise add zero columns to one of them).
- $AA^T = Cov(\boldsymbol{X}) = Cov(\boldsymbol{Y}) = BB^T$.
- Show that then $B^T = QA^T$ for some $QQ^T = I$.
- Then $Y = B\boldsymbol{W} = A(Q^T\boldsymbol{W})$
- $Q^T\boldsymbol{W} = Z^*$, where Z^* are multivariate standard normal.
- Then $Y = AZ^*$, $X = AZ$, same transformation of multivariate standard normal, are of the some distribution.
- A detailed proof of the uniqueness theorem was provided by Miurhead (ref. Miurhead 1982).

Theorem (Joint density of multivariate normal)

If $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ is p-variate normal with mean $\boldsymbol{\mu}$ and covariance matrix Σ, where Σ is symmetric and positive definite, then $\mathbf{X}$ has density function

$$f_X(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

Proof.

We need to show that there is a normal random vector with the given mean and covariance matrix, its density is of the desired form.

Since Σ is positive definite, by matrix theory, there is a $p \times p$ invertible matrix A such that

$$\Sigma = AA'$$

with

$$\Sigma^{-1} = A'^{-1}A^{-1}, \qquad |\Sigma| = det(\Sigma) = (det(A))^2$$

Let $\boldsymbol{Z}_p \sim N(0, I_p)$ be the p-variate standard normal, and define

$$X^* = A\boldsymbol{Z}_p + \boldsymbol{\mu}$$

Then

$$X^* \sim N_p(\boldsymbol{\mu}, AA') = N_p(\boldsymbol{\mu}, \Sigma)$$

By the uniqueness theorem, X and $X^* = A\boldsymbol{Z}_p + \boldsymbol{\mu}$ are of the same mean, same covariance matrix, thus they are of the same multivariate normal distribution.

By variable transformation, we may derive the common density function of X and X^* from the known density of $\boldsymbol{Z} = A^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$, by the method of variable substitution for continuous functions. Denote the variable transformation Jacobian as

$$J = \left[\frac{\partial z_i}{\partial x_j}\right]_{i,j=1,\cdots,p} = A^{-1}$$

then

$$\begin{aligned} f_X(\boldsymbol{x}) &= |det(J)| f_Z(A^{-1}(\boldsymbol{x}-\boldsymbol{\mu})) \\ &= |det(A^{-1})| \frac{1}{(2\pi)^{p/2}} e^{\frac{1}{2}(A^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))'(A^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))} \\ &= \frac{1}{|det(A)|} \frac{1}{(2\pi)^{p/2}} e^{\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})' A'^{-1} A^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \\ &= \frac{1}{|\Sigma|^{1/2}} \frac{1}{(2\pi)^{p/2}} e^{\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \end{aligned}$$

□

Remarks

This theorem gives a necessary and sufficient condition for multivariate normal random vectors, thus often serves as the definition.

4 Moment generation functions

The moment generating function for a univariate random variable X is

$$M_X(t) = \mathbb{E}\left(e^{tX}\right)$$

for t values at which the expectation exists. The most important property is that the kth derivative of moment generation function at the origin is

$$M^{(k)}(0) = \mathbb{E}(X^k),$$

the kth moment of X (thus the name). For $X \sim N(\mu, \sigma^2)$, the moment generating function is

$$M_X(t) = \mathbb{E}\left(e^{tX}\right) = e^{t\mu + \frac{1}{2}\sigma^2 t^2}$$

The moment-generating function of a p-variate random vector $\boldsymbol{X} = (X_1, \cdots, X_p)'$ has the form

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}(e^{\boldsymbol{t}'\boldsymbol{X}}) = \mathbb{E}\left(e^{t_1 X_1 + \cdots + t_p X_p}\right)$$

which is defined for $\boldsymbol{t} = (t_1, \cdots, t_p)'$ wherever the expectation exists. The moment-generating function of p-variate normal $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ has the simple expression

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = e^{\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t}} = e^{t_1\mu_1 + \cdots + t_p\mu_p + \frac{1}{2}Var(t_1X_1 + \cdots + t_pX_p)},$$

which can be derived from the moment generating function for univariate normal,

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = M_{\boldsymbol{t}'\boldsymbol{X}}(1)$$

Remarks

In this course we rarely use moment generating functions directly. However many theoretical results we assume or use in this course, such as the central limit theorem and its various versions, are relatively easier to derive using moment generating functions or characteristic functions.

5 Maximum likelihood estimation

There are many methods to estimate the population parameters $\boldsymbol{\mu}$ and Σ from sample data. Maximum likelihood method is a popular and powerful method.

Let $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ be a random sample of independent p-variate random vectors of distribution $N_p(\boldsymbol{\mu}, \Sigma)$. Then each p-variate sample vector $\boldsymbol{X}_j$ has density function

$$f(\boldsymbol{x}_j) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}_j-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu})}$$

By independence, $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ has joint density function

$$f(\boldsymbol{x}_1, \cdots, x_n; \boldsymbol{\mu}, \Sigma) = \prod_{j=1}^{n} f(\boldsymbol{x}_j) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu})}$$

Given sample data $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$, the above joint density becomes a function with unknown parameters μ_i's and σ_{ik}'s $(i, k = 1, \cdots, p)$ in $\boldsymbol{\mu}$ and Σ. Express the joint density as a function of the unknown parameters given the data, we write

$$L(\boldsymbol{\mu}, \Sigma) = L(\boldsymbol{\mu}, \Sigma | \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) = \prod_{j=1}^{n} f(\boldsymbol{x}_j) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu})}$$

This function is called the likelihood function of the sample data of size n.

The values of μ_i's and σ_{ik}'s, thus the values of $\boldsymbol{\mu}$ and Σ that maximize the likelihood function are the MLE's (maximum likelihood estimates).

For p-variate normal, the MLE of the population mean $\boldsymbol{\mu}$ is the sample mean

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j$$

and the MLE of the population covariance Σ is the sample variance matrix with denominator n (note: not $n-1$)

$$\hat{\Sigma} = \boldsymbol{S}_n = \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'$$

The MLE $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ maximize the likelihood function:

$$L(\bar{\boldsymbol{x}}, \boldsymbol{S}_n) = L(\hat{\boldsymbol{\mu}}, \hat{\Sigma}) = \max_{\boldsymbol{\mu}, \Sigma} L(\boldsymbol{\mu}, \Sigma)$$

where

$$L(\boldsymbol{\mu}, \Sigma) = L(\boldsymbol{\mu}, \Sigma | \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) = f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n; \boldsymbol{\mu}, \Sigma) = \prod_{j=1}^{n} f(\boldsymbol{x}_j)$$

Proof. The derivation of MLE $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$ and $\hat{\Sigma} = \boldsymbol{S}_n$ consists of two parts:

The $\boldsymbol{\mu}$ part

$\boldsymbol{\mu}$ only occurs in the exponent. We will show that, for any Σ, the exponent is maximized when the estimator of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$.

$$L(\hat{\boldsymbol{\mu}}, \Sigma) = \max_{\boldsymbol{\mu}} L(\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} e^{-\frac{1}{2}tr\{\Sigma^{-1}(n\boldsymbol{S}_n)\}} \quad (3)$$

The Σ part

We will show that, among all positive definite Σ, $L(\hat{\boldsymbol{\mu}}, \Sigma)$ is maximized when Σ is $\hat{\Sigma} = \boldsymbol{S}_n$, which yields

$$L(\hat{\boldsymbol{\mu}}, \hat{\Sigma}) = \max_{\Sigma} L(\hat{\boldsymbol{\mu}}, \Sigma) = \max_{\boldsymbol{\mu}, \Sigma} L(\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{S}_n|^{n/2}} e^{-\frac{np}{2}} \tag{4}$$

The following are the steps of the proof.

Proof of the $\boldsymbol{\mu}$ part

First we need to rewrite the exponent in terms of matrix trace, which is the sum of the diagonal elements of a square matrix.

- Use the property $tr(AB) = tr(BA)$ and $tr(c) = c$ for any number c, the scalar

$$(\boldsymbol{x}_j - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}) = tr\left\{(\boldsymbol{x}_j - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu})\right\} = tr\left\{\Sigma^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu})(\boldsymbol{x}_j - \boldsymbol{\mu})'\right\}$$

- Using the property $tr(A + B) = tr(A) + tr(B)$,

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

- Regrouping, and noting that $\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = 0$,

$$\begin{aligned}
\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})' &= \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}} + \bar{\boldsymbol{x}} - \boldsymbol{\mu})(\boldsymbol{x}_i - \bar{\boldsymbol{x}} + \bar{\boldsymbol{x}} - \boldsymbol{\mu})' \\
&= \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' + \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})' + \sum_{i=1}^{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' + \sum_{i=1}^{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})' \\
&= \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})' \\
&= n\boldsymbol{S}_n + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})'
\end{aligned}$$

- The exponent of the likelihood function becomes

$$-\frac{1}{2}tr\left\{\Sigma^{-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})\right\} = -\frac{1}{2}tr\left[\Sigma^{-1}(n\boldsymbol{S}_n)\right] - \frac{1}{2}tr\left\{\Sigma^{-1}\left[n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})'\right]\right\} \tag{5}$$

By the positive definiteness of Σ^{-1}, the last term in (5) is non-negative:

$$\frac{1}{2}tr\left\{\Sigma^{-1}\left[n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})'\right]\right\} = \frac{n}{2}tr\left\{(\bar{\boldsymbol{x}} - \boldsymbol{\mu})'\Sigma^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})\right\} \begin{cases} = 0, & \boldsymbol{\mu} = \bar{\boldsymbol{x}}, \\ > 0, & \boldsymbol{\mu} \neq \bar{\boldsymbol{x}}, \end{cases}$$

Thus the exponent of the likelihood function

$$-\frac{1}{2}tr\Big(\Sigma^{-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})\Big) \begin{cases} = -\frac{1}{2}tr(\Sigma^{-1}n\boldsymbol{S}_n), & \boldsymbol{\mu} = \bar{\boldsymbol{x}}, \\ < -\frac{1}{2}tr(\Sigma^{-1}n\boldsymbol{S}_n), & \boldsymbol{\mu} \neq \bar{\boldsymbol{x}}. \end{cases}$$

That is, the exponent is maximized when $\mu = \bar{\boldsymbol{x}}$, for any Σ. So (3) is proved.

Proof of the Σ part

For any symmetric positive definite Σ, the matrix $\Sigma^{-1}(n\boldsymbol{S}_n)$ is also positive definite (or positive semi-definite when $\boldsymbol{S}$ is so). Denote the eigenvalues of $\Sigma^{-1}(n\boldsymbol{S}_n)$ as $\lambda_k > 0$ (≥ 0 if $\boldsymbol{S}_n$ is positive semi-definite), $k = 1, \cdots, p$.

By the relationship of matrix trace and matrix eigenvalues,

$$tr\{\Sigma^{-1}(n\boldsymbol{S}_n)\} = \sum_{k=1}^{p}\lambda_k$$

By the relationship of matrix determinant and matrix eigenvalues,

$$det\{\Sigma^{-1}(n\boldsymbol{S}_n)\} = \frac{det(n\boldsymbol{S}_n)}{det(\Sigma)} = \prod_{k=1}^{p}\lambda_k$$

Thus

$$\frac{1}{|\Sigma|^{n/2}}e^{-\frac{1}{2}tr\{\Sigma^{-1}(n\boldsymbol{S}_n)\}} = \left(\frac{\prod_{k=1}^{p}\lambda_k}{|n\boldsymbol{S}_n|}\right)^{n/2} e^{-\frac{1}{2}\sum_{k=1}^{p}\lambda_k} = \frac{1}{n^{np/2}|\boldsymbol{S}_n|^{n/2}}\prod_{k=1}^{p}\lambda_k^{n/2}e^{-\frac{1}{2}\lambda_k}$$

Since the function $t^{n/2}e^{-t/2}$ achieves maximum $n^{n/2}e^{-n/2}$ at $t = n$ for given n (calculus exercise),

$$t^{n/2}e^{-t/2} \leq n^{n/2}e^{-\frac{n}{2}} \qquad \Longrightarrow \qquad \lambda_k^{n/2}e^{-\frac{1}{2}\lambda_k} \leq n^{n/2}e^{-\frac{n}{2}}$$

We have

$$\frac{1}{|\Sigma|^{n/2}}e^{-\frac{1}{2}tr\{\Sigma^{-1}(n\boldsymbol{S}_n)\}} \leq \frac{1}{n^{np/2}|\boldsymbol{S}_n|^{n/2}}\prod_{k=1}^{p}n^{n/2}e^{-\frac{n}{2}} = \frac{1}{|\boldsymbol{S}_n|^{n/2}}e^{-\frac{np}{2}}$$

The inequality holds if $\lambda_k \equiv n, k = 1, \cdots, p$, when and only when

$$\Sigma^{-1}(n\boldsymbol{S}_n) = nI_p \qquad \Longleftrightarrow \qquad \Sigma = \boldsymbol{S}_n$$

Therefore

$$\max_{\boldsymbol{\mu}, \Sigma} L(\boldsymbol{\mu}, \Sigma) = \max_{\Sigma} L(\hat{\boldsymbol{\mu}}, \Sigma) = \max_{\Sigma} \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}e^{-\frac{1}{2}tr\{\Sigma^{-1}(n\boldsymbol{S}_n)\}} = \frac{1}{(2\pi)^{np/2}|\boldsymbol{S}_n|^{n/2}}e^{-\frac{np}{2}} = L(\hat{\boldsymbol{\mu}}, \hat{\Sigma}),$$

which is (4).

We have proved that the maximum likelihood is achieved at $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$ and $\hat{\Sigma} = \boldsymbol{S}_n$.

□

Remarks on MLE

- Recall the sample generalized variance is $|\boldsymbol{S}| = \frac{n}{n-1}|\boldsymbol{S}_n|$, the maximum likelihood can be written as

$$\frac{1}{(2\pi)^{np/2}|\boldsymbol{S}_n|^{n/2}}e^{-\frac{np}{2}} = C_1|\boldsymbol{S}_n|^{-n/2} = C|\boldsymbol{S}|^{-n/2} = \text{constant} \times (\text{sample generalized variance})^{-n/2}$$

This expression will be useful in deriving likelihood based tests.

- Maximum likelihood estimates have several nice properties, such as functional invariance, consistency, and asymptotic normality.
- The estimators match the univariate maximum likelihood estimators, where

$$\hat{\mu} = \bar{x}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{n-1}{n}s^2$$

That is, while the MLE of μ is an unbiased estimator, the MLE of σ^2 is not unbiased.