

Multivariate Statistics

Lecture 04

Fudan University

- 1 Multivariate Normal Distribution (Conditional Distribution)
- 2 Characteristic Function
- 3 Maximum Likelihood Estimator of Mean and Covariance

- 1 Multivariate Normal Distribution (Conditional Distribution)
- 2 Characteristic Function
- 3 Maximum Likelihood Estimator of Mean and Covariance

Multivariate Normal Distribution (Conditional Distribution)

Let \mathbf{x} be distributed according to $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \succ \mathbf{0}$. Let us partition

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix} \quad \text{with } \mathbf{x}^{(1)} \in \mathbb{R}^q \text{ and } \mathbf{x}^{(2)} \in \mathbb{R}^{p-q}.$$

The joint density of $\mathbf{y}^{(1)} = \mathbf{x}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}^{(2)}$ and $\mathbf{y}^{(2)} = \mathbf{x}^{(2)}$ is

$$g(\mathbf{y}) = n(\mathbf{y}^{(1)} \mid \boldsymbol{\mu}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})n(\mathbf{y}^{(2)} \mid \boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22}).$$

Consider that

$$\begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix} = \mathbf{u}(\mathbf{x})$$

and use

$$f(\mathbf{x}) = g(\mathbf{u}(\mathbf{x}))|\det(\mathbf{J}(\mathbf{x}))| = g(\mathbf{u}(\mathbf{x})).$$

Multivariate Normal Distribution (Conditional Distribution)

The resulting joint density of $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ is

$$\begin{aligned} f(\mathbf{x}) &= f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \\ &= n(\mathbf{y}^{(1)} \mid \boldsymbol{\mu}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) n(\mathbf{x}^{(2)} \mid \boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22}) \\ &= \frac{1}{\sqrt{(2\pi)^q \det(\boldsymbol{\Sigma}_{11.2})}} \exp\left(-\frac{1}{2}(\mathbf{x}_{11.2} - \boldsymbol{\mu}_{11.2})^\top \boldsymbol{\Sigma}_{11.2}^{-1}(\mathbf{x}_{11.2} - \boldsymbol{\mu}_{11.2})\right) \\ &\quad \cdot \frac{1}{\sqrt{(2\pi)^{p-q} \det(\boldsymbol{\Sigma}_{22})}} \exp\left(-\frac{1}{2}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})^\top \boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})\right) \end{aligned}$$

where

$$\begin{aligned} \mathbf{x}_{11.2} &= \mathbf{x}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}^{(2)}, \\ \boldsymbol{\mu}_{11.2} &= \boldsymbol{\mu}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}^{(2)}, \\ \boldsymbol{\Sigma}_{11.2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \end{aligned}$$

Multivariate Normal Distribution (Conditional Distribution)

The marginal density of $\mathbf{x}^{(2)}$ is

$$\begin{aligned} f(\mathbf{x}^{(2)}) &= n(\mathbf{y}^{(2)} \mid \boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22}) \\ &= \frac{1}{\sqrt{(2\pi)^{p-q} \det(\boldsymbol{\Sigma}_{22})}} \exp \left(-\frac{1}{2} \left(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)} \right)^\top \boldsymbol{\Sigma}_{22}^{-1} \left(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)} \right) \right). \end{aligned}$$

Hence, the conditional density of $\mathbf{x}^{(1)}$ given that $\mathbf{x}^{(2)}$ is

$$\begin{aligned} f(\mathbf{x}^{(1)} \mid \mathbf{x}^{(2)}) &= \frac{f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})}{f(\mathbf{x}^{(2)})} \\ &= \frac{1}{\sqrt{(2\pi)^q \det(\boldsymbol{\Sigma}_{11.2})}} \exp \left(-\frac{1}{2} (\mathbf{x}_{11.2} - \boldsymbol{\mu}_{11.2})^\top \boldsymbol{\Sigma}_{11.2}^{-1} (\mathbf{x}_{11.2} - \boldsymbol{\mu}_{11.2}) \right) \end{aligned}$$

Multivariate Normal Distribution (Conditional Distribution)

The conditional density of $\mathbf{x}^{(1)}$ given that $\mathbf{x}^{(2)}$ is

$$\begin{aligned} f(\mathbf{x}^{(1)} | \mathbf{x}^{(2)}) &= \frac{f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})}{f(\mathbf{x}^{(2)})} \\ &= \frac{1}{\sqrt{(2\pi)^q \det(\boldsymbol{\Sigma}_{11.2})}} \exp \left(-\frac{1}{2} (\mathbf{x}_{11.2} - \boldsymbol{\mu}_{11.2})^\top \boldsymbol{\Sigma}_{11.2}^{-1} (\mathbf{x}_{11.2} - \boldsymbol{\mu}_{11.2}) \right) \end{aligned}$$

Consider that $\mathbf{x}_{11.2} - \boldsymbol{\mu}_{11.2} = \mathbf{x}^{(1)} - (\boldsymbol{\mu}^{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}))$.

The density $f(\mathbf{x}^{(1)} | \mathbf{x}^{(2)})$ is a q -variate normal density with mean

$$\mathbb{E}[\mathbf{x}^{(1)} | \mathbf{x}^{(2)}] = \boldsymbol{\mu}^{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}) = \boldsymbol{\nu}(\mathbf{x}^{(2)})$$

and covariance matrix (not depend on $\mathbf{x}^{(2)}$)

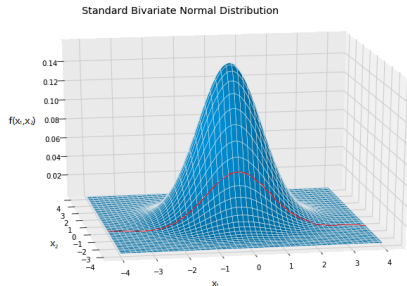
$$\begin{aligned} \text{Cov}[\mathbf{x}^{(1)} | \mathbf{x}^{(2)}] &= \mathbb{E}[(\mathbf{x}^{(1)} - \boldsymbol{\nu}(\mathbf{x}^{(2)}))(\mathbf{x}^{(1)} - \boldsymbol{\nu}(\mathbf{x}^{(2)}))^\top | \mathbf{x}^{(2)}] \\ &= \boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \preceq \boldsymbol{\Sigma}_{11}. \end{aligned}$$

Multivariate Normal Distribution (Conditional Distribution)

The density $f(x_1, x_2)$ can be thought of as a surface $z = f(x_1, x_2)$ over the x_1, x_2 -plane.

If we intersect this surface with the plane $x_2 = c$, we obtain a curve $z = f(x_1, c)$ over the line $x_2 = c$ in the x_1, x_2 -plane.

The ordinate of this curve is proportional to the conditional density of x_1 given $x_2 = c$; that is, it is proportional to the ordinate of the curve of a univariate normal distribution.



Correlation Coefficient

Recall that for random vector $\mathbf{x} = [x_1, x_2, \dots, x_p]^\top$, we define the covariance matrix as

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix} \in \mathbb{R}^{p \times p}$$

and the correlation coefficient between x_i and x_j as (suppose $\mathbf{\Sigma} \succ \mathbf{0}$)

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}.$$

Partial Correlation Coefficient

Now consider the partition

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix} \quad \text{with } \mathbf{x}^{(1)} \in \mathbb{R}^q \text{ and } \mathbf{x}^{(2)} \in \mathbb{R}^{p-q}.$$

Let

$$\Sigma_{11.2} = \begin{bmatrix} \sigma_{11 \cdot q+1, \dots, p} & \sigma_{12 \cdot q+1, \dots, p} & \dots & \sigma_{1q \cdot q+1, \dots, p} \\ \sigma_{21 \cdot q+1, \dots, p} & \sigma_{22 \cdot q+1, \dots, p} & \dots & \sigma_{2q \cdot q+1, \dots, p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1 \cdot q+1, \dots, p} & \sigma_{q2 \cdot q+1, \dots, p} & \dots & \sigma_{qq \cdot q+1, \dots, p} \end{bmatrix} \in \mathbb{R}^{q \times q}.$$

We define

$$\rho_{ij \cdot q+1, \dots, p} = \frac{\sigma_{ij \cdot q+1, \dots, p}}{\sqrt{\sigma_{ii \cdot q+1, \dots, p}} \sqrt{\sigma_{jj \cdot q+1, \dots, p}}}$$

as the partial correlation between x_i and x_j holding x_{q+1}, \dots, x_p fixed.

Multiple Correlation Coefficient

We again consider $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ such that

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \succ \mathbf{0}.$$

Then, we study some properties of $\mathbf{B}\mathbf{x}^{(2)}$, where

$$\mathbf{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$$

is the matrix of regression coefficients of $\mathbf{x}^{(1)}$ on $\mathbf{x}^{(2)}$.

The vector $\mathbb{E}[\mathbf{x}^{(1)} \mid \mathbf{x}^{(2)}] = \boldsymbol{\mu}^{(1)} + \mathbf{B}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})$ is called the regression function.

Multiple Correlation Coefficient

The vector

$$\mathbf{x}^{(11.2)} = \mathbf{x}^{(1)} - (\boldsymbol{\mu}^{(1)} + \mathbf{B}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}))$$

is the vector of residuals of $\mathbf{x}^{(1)}$ from its regression on $\mathbf{x}^{(2)}$.

The components of $\mathbf{x}^{(11.2)}$ are uncorrelated with the components of $\mathbf{x}^{(2)}$ since we have

$$\mathbf{x}^{(11.2)} = \mathbf{y}^{(1)} - \mathbb{E}[\mathbf{y}^{(1)}],$$

such that

$$\begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)} - \mathbf{B}\mathbf{x}^{(2)} \\ \mathbf{x}^{(2)} \end{bmatrix}.$$

Multiple Correlation Coefficient

Theorem 1

For $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and every vector $\boldsymbol{\alpha} \in \mathbb{R}^{(p-q)}$, we have

$$\text{Var}(x_i^{(11.2)}) \leq \text{Var}(x_i - \boldsymbol{\alpha}^\top \mathbf{x}^{(2)}),$$

where $x_i^{(11.2)}$ and x_i are the i -th entry of $\mathbf{x}^{(11.2)}$ and the i -th entry of \mathbf{x} respectively.

Observe that

$$\mathbb{E}[x_i] = \mu_i + \boldsymbol{\alpha}^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}),$$

which means

$$\mu_i + \boldsymbol{\beta}_{(i)}^\top (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})$$

is the best linear predictor of x_i in all functions of the form $\boldsymbol{\alpha}^\top \mathbf{x}^{(2)} + c$, the mean squared error of the above is a minimum.

Multiple Correlation Coefficient

The correlation of two variables z_1 and z_2 is defined as

$$\text{Corr}(z_1, z_2) = \frac{\text{Cov}[z_1, z_2]}{\sqrt{\text{Var}[z_1]\text{Var}[z_2]}}.$$

The maximum correlation between x_i and the linear combination $\alpha^\top \mathbf{x}^{(2)}$ is called the multiple correlation coefficient between x_i and $\alpha^\top \mathbf{x}^{(2)}$.

Corollary 1

Under the setting of Theorem 1, prove that

$$\text{Corr} \left[x_i, \beta_{(i)}^\top \mathbf{x}^{(2)} \right] \geq \text{Corr} \left[x_i, \alpha^\top \mathbf{x}^{(2)} \right]$$

for every $\alpha \in \mathbb{R}^{(p-q)}$.

- 1 Multivariate Normal Distribution (Conditional Distribution)
- 2 Characteristic Function
- 3 Maximum Likelihood Estimator of Mean and Covariance

Characteristic Function

The characteristic function of a p -dimensional random vector \mathbf{x} is

$$\phi(\mathbf{t}) = \mathbb{E} \left[\exp(\mathbf{i} \mathbf{t}^\top \mathbf{x}) \right]$$

defined for every real vector $\mathbf{t} \in \mathbb{R}^p$.

For the complex-valued function $g(z)$ be written as

$$g(z) = g_1(z) + \mathbf{i} g_2(z),$$

where $g_1(z)$ and $g_2(z)$ are real-valued, the expected value of $g(z)$ is

$$\mathbb{E}[g(z)] = \mathbb{E}[g_1(z)] + \mathbf{i} \mathbb{E}[g_2(z)].$$

Characteristic Function

Let $\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix}$ be a p -dimensional random vector. If $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are independent and $g(\mathbf{x}) = g^{(1)}(\mathbf{x}^{(1)})g^{(2)}(\mathbf{x}^{(2)})$, then we have

$$\mathbb{E}[g(\mathbf{x})] = \mathbb{E}[g^{(1)}(\mathbf{x}^{(1)})] \mathbb{E}[g^{(2)}(\mathbf{x}^{(2)})].$$

If the components of \mathbf{x} are mutually independent, then

$$\mathbb{E}[\exp(\mathbf{i} \mathbf{t}^\top \mathbf{x})] = \mathbb{E} \left[\prod_{j=1}^p \exp(\mathbf{i} t_j x_j) \right].$$

Characteristic Function

Theorem 2

The characteristic function of \mathbf{x} distributed according to $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\phi(\mathbf{t}) = \exp \left(i \mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2} \mathbf{t}^\top \boldsymbol{\Sigma} \mathbf{t} \right).$$

for every $\mathbf{t} \in \mathbb{R}^p$.

Sketch of the proof

- 1 The characteristic function of $\mathbf{y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ is $\phi_0(\mathbf{t}) = \exp \left(-\frac{1}{2} \mathbf{t}^\top \mathbf{t} \right)$.
- 2 For $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have $\mathbf{x} = \mathbf{A}\mathbf{y} + \boldsymbol{\mu}$ such that $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top$.
- 3 Using $\phi_0(\mathbf{t})$ to present the characteristic function of $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Theorem 2

The characteristic function of \mathbf{x} distributed according to $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\phi(\mathbf{t}) = \exp \left(\mathbf{i} \mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2} \mathbf{t}^\top \boldsymbol{\Sigma} \mathbf{t} \right).$$

for every $\mathbf{t} \in \mathbb{R}^p$.

We can use this theorem to prove $\mathbf{z} = \mathbf{D}\mathbf{x} \sim \mathcal{N}(\mathbf{D}\boldsymbol{\mu}, \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top)$ easily.

Characteristic Function

The following theorem can be viewed as another definition of multivariate normal distribution.

Theorem 3

If every linear combination of the components of a random vector \mathbf{y} is normally distributed, then \mathbf{y} is normally distributed.

In other words, if the p -dimensional random vector \mathbf{y} leads to the univariate random variable

$$\mathbf{u}^T \mathbf{y}$$

is normally distributed for any fixed $\mathbf{u} \in \mathbb{R}^p$, then \mathbf{y} is normally distributed.

Characteristic Function

Problem in Exam

Let $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\mathbf{y} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ and $\mathbf{z} = \mathbf{x} + \mathbf{y}$. Suppose that \mathbf{x} and \mathbf{y} are independent. Prove $\mathbf{z} \sim \mathcal{N}_p(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$.

Use characteristic function to avoid using density.

Characteristic Function and Density

The characteristic function determines the density function uniquely (if the density exists).

Theorem 4

If the p -dimensional random vector \mathbf{x} has the density $f(\mathbf{x})$ and the characteristic function $\phi(\mathbf{t})$, then

$$f(\mathbf{x}) = \frac{1}{(2\pi)^p} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp(-i \mathbf{t}^\top \mathbf{x}) \phi(\mathbf{t}) dt_1 \cdots dt_p.$$

See the proof in Section 10.6 of “Cramer, H. (1946). Mathematical Methods of Statistics. Princeton University Press”.

Characteristic Function and Probability

If \mathbf{x} does not have a density, the characteristic function uniquely defines the probability of any continuity interval.

Theorem 5

Let $\{F_j(x)\}$ be a sequence of cdfs, and let $\{\phi_j(t)\}$ be the sequence of corresponding characteristic functions. A necessary and sufficient condition for $F_j(x)$ to converge to a cdf $F(x)$ is that, for every t , $\phi_j(t)$ converges to a limit $\phi(t)$ that is continuous at $t = 0$. When this condition is satisfied, the limit $\phi(t)$ is identical with the characteristic function of the limiting distribution $F(x)$.

See the proof in Section 10.7 of “Cramer, H. (1946). Mathematical Methods of Statistics. Princeton University Press”

Characteristic Function and Moments

If the n -th moment of random variable x , denoted by $\mathbb{E}[x^n]$, exists and is finite, then its characteristic function is n times continuously differentiable and

$$\mathbb{E}[x^n] = \frac{1}{i^n} \left. \frac{d^n \phi(t)}{dt^n} \right|_{t=0},$$

which is because of

$$\begin{aligned} \frac{d^n \phi(t)}{dt^n} &= \frac{d^n}{dt^n} \mathbb{E}[\exp(i tx)] \\ &= \mathbb{E} \left[\frac{d^n}{dt^n} \exp(i tx) \right] \\ &= \mathbb{E} [(i x)^n \exp(i tx)] \\ &= i^n \mathbb{E} [x^n \exp(i tx)]. \end{aligned}$$

Characteristic Function and Moments

For normal distributed random vector $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and its characteristic function $\phi(\mathbf{t}) = \exp(\mathbf{i} \mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2} \mathbf{t}^\top \boldsymbol{\Sigma} \mathbf{t})$, we have

$$\mathbb{E}[x_h] = \left. \frac{1}{\mathbf{i}} \frac{d\phi(\mathbf{t})}{dt_h} \right|_{\mathbf{t}=\mathbf{0}} = \frac{1}{\mathbf{i}} \left(\mathbf{i} \mu_h - \sum_{j=1}^p \sigma_{hj} t_j \right) \phi(\mathbf{t}) \Big|_{\mathbf{t}=\mathbf{0}} = \mu_h$$

and

$$\begin{aligned} \mathbb{E}[x_h x_j] &= \left. \frac{1}{\mathbf{i}^2} \frac{\partial^2 \phi(\mathbf{t})}{\partial t_h \partial t_j} \right|_{\mathbf{t}=\mathbf{0}} \\ &= \frac{1}{\mathbf{i}^2} \left(\left(- \sum_{k=1}^p \sigma_{hk} t_k + \mathbf{i} \mu_h \right) - \sigma_{hj} \right) \left(\left(- \sum_{k=1}^p \sigma_{kj} t_k + \mathbf{i} \mu_j \right) - \sigma_{hj} \right) \phi(\mathbf{t}) \Big|_{\mathbf{t}=\mathbf{0}} \\ &= \sigma_{hj} + \mu_h \mu_j. \end{aligned}$$

Thus, we have

$$\begin{aligned} \text{Var}(x_h) &= \mathbb{E}[x_h - \mu_h]^2 = \mathbb{E}[x_h^2] - \mu_h^2 = \sigma_{hh}, \\ \text{Cov}(x_h, x_j) &= \mathbb{E}[(x_h - \mu_h)(x_j - \mu_j)] = \mathbb{E}[x_h x_j] - \mu_h \mu_j = \sigma_{hj}. \end{aligned}$$

Characteristic Function and Moments

If all the moments of a distribution exist, then the cumulants are the coefficients κ in

$$\log \phi(\mathbf{t}) = \sum_{s_1=0}^{\infty} \cdots \sum_{s_p=0}^{\infty} \kappa_{s_1 \dots s_p} \frac{(it_1)^{s_1} \dots (it_p)^{s_p}}{s_1! \dots s_p!}.$$

In the case of the multivariate normal distribution, we have

$$\kappa_{100\dots 0} = \mu_1, \quad \kappa_{010\dots 0} = \mu_2, \quad \dots \quad \kappa_{000\dots 1} = \mu_p,$$

and

$$\kappa_{200\dots 0} = \sigma_{11}, \quad \kappa_{110\dots 0} = \sigma_{12}, \quad \dots \quad \kappa_{000\dots 2} = \sigma_{pp}.$$

The cumulants for which $\sum s_i > 2$ are 0.

- 1 Multivariate Normal Distribution (Conditional Distribution)
- 2 Characteristic Function
- 3 Maximum Likelihood Estimator of Mean and Covariance

The Maximum Likelihood Estimators

Given a sample of (vector) observations from a p -variate (non-singular) normal distribution, we ask for estimators of the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of the distribution.

Suppose our sample of N observations on the $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$, which are distributed according to $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $N > p$. The likelihood function is

$$\begin{aligned} L &= \prod_{\alpha=1}^N n(\mathbf{x}_{\alpha} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \frac{1}{(2\pi)^{\frac{pN}{2}} (\det(\boldsymbol{\Sigma}))^{\frac{N}{2}}} \exp \left[-\frac{1}{2} \sum_{\alpha=1}^N (\mathbf{x}_{\alpha} - \boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_{\alpha} - \boldsymbol{\mu}) \right]. \end{aligned}$$

The Maximum Likelihood Estimators

The likelihood function is

$$L = \frac{1}{(2\pi)^{\frac{pN}{2}} (\det(\mathbf{\Sigma}))^{\frac{N}{2}}} \exp \left[-\frac{1}{2} \sum_{\alpha=1}^N (\mathbf{x}_{\alpha} - \boldsymbol{\mu})^{\top} \mathbf{\Sigma}^{-1} (\mathbf{x}_{\alpha} - \boldsymbol{\mu}) \right].$$

The vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ are fixed at the sample values and L is a function of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$.

The logarithm of the likelihood function is

$$\ln L = -\frac{PN}{2} \ln 2\pi - \frac{N}{2} \ln (\det(\mathbf{\Sigma})) - \frac{1}{2} \sum_{\alpha=1}^N (\mathbf{x}_{\alpha} - \boldsymbol{\mu})^{\top} \mathbf{\Sigma}^{-1} (\mathbf{x}_{\alpha} - \boldsymbol{\mu}).$$

Since $\ln L$ is an increasing function of L , the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are the vector and the positive definite matrix that maximize $\ln L$.

The Maximum Likelihood Estimators

Let the mean vector be

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{\alpha=1}^N \mathbf{x}_{\alpha} = \begin{bmatrix} \frac{1}{N} \sum_{\alpha=1}^N x_{1\alpha} \\ \vdots \\ \frac{1}{N} \sum_{\alpha=1}^N x_{p\alpha} \end{bmatrix} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

where

$$\mathbf{x}_{\alpha} = \begin{bmatrix} x_{1\alpha} \\ \vdots \\ x_{p\alpha} \end{bmatrix} \quad \text{and} \quad \bar{x}_i = \frac{1}{N} \sum_{\alpha=1}^N x_{i\alpha}.$$

Let the matrix of sums of squares and cross products of deviations about the mean be

$$\mathbf{A} = \frac{1}{N} \sum_{\alpha=1}^N (\mathbf{x}_{\alpha} - \bar{\mathbf{x}})(\mathbf{x}_{\alpha} - \bar{\mathbf{x}})^{\top}$$

The Maximum Likelihood Estimators

Theorem 6

If $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ constitute a sample from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $p < N$, the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{N} \sum_{\alpha=1}^N \mathbf{x}_{\alpha} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{\alpha=1}^N (\mathbf{x}_{\alpha} - \bar{\mathbf{x}})(\mathbf{x}_{\alpha} - \bar{\mathbf{x}})^{\top}$$

respectively.

Lemma 1

If $\mathbf{D} \in \mathbb{R}^{p \times p}$ is positive definite, the maximum of

$$f(\mathbf{G}) = -N \ln \det(\mathbf{G}) - \text{tr}(\mathbf{G}^{-1} \mathbf{D})$$

with respect to positive definite matrices \mathbf{G} exists, occurs at $\mathbf{G} = \frac{1}{N} \mathbf{D}$.

The Maximum Likelihood Estimators

The maximum likelihood estimators of functions of the parameters are those functions of the maximum likelihood estimators of the parameters.

Theorem 7

Let $f(\theta)$ be a real-valued function defined on a set \mathcal{S} and let ϕ be a single-valued function, with a single-valued inverse, on \mathcal{S} to a set \mathcal{S}^* . Let

$$g(\theta^*) = f(\phi^{-1}(\theta^*)).$$

Then if $f(\theta)$ attains a maximum at $\theta = \theta_0$, then $g(\theta^*)$ attains a maximum at $\theta^* = \theta_0^* = \phi(\theta_0)$. If the maximum of $f(\theta)$ at θ_0 is unique, so is the maximum of $g(\theta^*)$ at θ_0^* .

The Maximum Likelihood Estimators

Corollary 2

If on the basis of a given sample $\hat{\theta}_1, \dots, \hat{\theta}_m$ are maximum likelihood estimators of the parameters $\theta_1, \dots, \theta_m$ of a distribution, then $\phi_1(\hat{\theta}_1, \dots, \hat{\theta}_m), \dots, \phi_m(\hat{\theta}_1, \dots, \hat{\theta}_m)$ are maximum likelihood estimator of $\phi_1(\theta_1, \dots, \theta_m), \dots, \phi_m(\theta_1, \dots, \theta_m)$ if the transformation from $\theta_1, \dots, \theta_m$ to ϕ_1, \dots, ϕ_m is one-to-one. If the estimators of $\theta_1, \dots, \theta_m$ are unique, then the estimators of ϕ_1, \dots, ϕ_m are unique.

Corollary 3

If $\mathbf{x}_1, \dots, \mathbf{x}_N$ constitutes a sample from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, let $\rho_{ij} = \sigma_{ij}/(\sigma_i\sigma_j)$. Then the maximum likelihood estimator of ρ_{ij} is

$$\hat{\rho}_{ij} = \frac{\sum_{\alpha=1}^N (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j)}{\sqrt{\sum_{\alpha=1}^N (x_{i\alpha} - \bar{x}_i)^2} \sqrt{\sum_{\alpha=1}^N (x_{j\alpha} - \bar{x}_j)^2}}$$

The Maximum Likelihood Estimators

If $\phi : \mathcal{S} \rightarrow \mathcal{S}^*$ is not one-to-one, we let

$$\phi^{-1}(\theta^*) = \{\theta : \theta^* = \phi(\theta)\}.$$

and define (the induced likelihood function)

$$g(\theta^*) = \sup\{f(\theta) : \theta^* = \phi(\theta)\}.$$

If $\theta = \hat{\theta}$ maximize $f(\theta)$, then $\theta^* = \phi(\hat{\theta})$ also maximize $g(\theta^*)$.