Multivariate Statistical Analysis

Lecture 03

Fudan University

luoluo@fudan.edu.cn

Outline

Outline

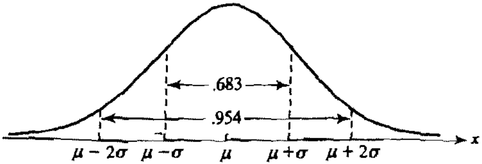
Univariate Normal Distribution

A random variable x is normally distributed with mean μ and standard deviation $\sigma > 0$ can be written in the following notation

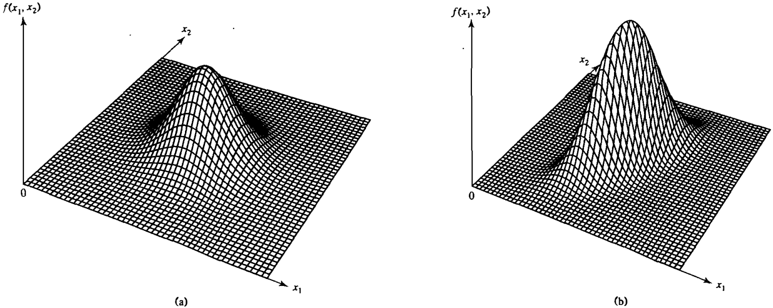
$$x \sim \mathcal{N}(\mu, \sigma^2).$$

The probability density function of univariate normal distribution is

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



Bivariate Normal Density



Two bivariate normal distributions:

- (a) $\sigma_1 = \sigma_2$ and $\rho_{12} = 0$
- (b) $\sigma_1 = \sigma_2$ and $\rho_{12} = 0.75$

The Central Limit Theorem

Let $x_1, \dots, x_n$ be independent and identically distributed random variables with the same arbitrary distribution, mean μ, and variance σ^2.

Let $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$, then the random variable

$$z = \lim_{n \to \infty} \sqrt{n} \left(\frac{\bar{x}_n - \mu}{\sigma} \right)$$

is a standard normal distribution.

The standard normal distribution is a normal distribution with a mean of 0 and standard deviation of 1.

What about multivariate case?

The Central Limit Theorem

$$\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}$$

$$\Rightarrow$$

Multivariate Normal Distribution

The multivariate normal distribution of a p-dimensional random vector $\mathbf{x} = [x_1, \ldots, x_p]^\top$ can be written in the following notation:

$$\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

or to make it explicitly known that $\mathbf{x}$ is p-dimensional

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

with p-dimensional mean vector

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] = \begin{bmatrix} \mathbb{E}[x_1] \\ \vdots \\ \mathbb{E}[x_p] \end{bmatrix} \in \mathbb{R}^p$$

and covariance matrix

$$\boldsymbol{\Sigma} = \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\right] \in \mathbb{R}^{p \times p}.$$

Multivariate Normal Distribution

The density function of univariate normal distribution is

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

where μ is the mean and σ^2 is the variance with $\sigma > 0$.

The density function of non-singular p-dimensional multivariate normal distribution is

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the mean and $\boldsymbol{\Sigma}$ is the $p \times p$ (non-singular) covariance matrix.

Density Function of Multivariate Normal Distribution

Theorem

Suppose the p-dimensional random vector $\mathbf{x}$ *has the density function*

$$f(\mathbf{x}) = K \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{b})^\top \mathbf{A}(\mathbf{x}-\mathbf{b})\right),$$

where $K \in \mathbb{R}$, $\mathbf{b} \in \mathbb{R}^p$ *and* $\mathbf{A} \in \mathbb{R}^{p\times p}$ *is symmetric positive definite. Then*

$$K = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}}, \qquad \mathbf{b} = \boldsymbol{\mu} \qquad \textit{and} \quad \mathbf{A} = \boldsymbol{\Sigma}^{-1}.$$

The main idea of this section:

Multivariate Normal Distribution

If the density of a p-dimensional random vector $\mathbf{x}$ is

$$K \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{b})^\top \mathbf{A}(\mathbf{x}-\mathbf{b})\right),$$

where $\mathbf{A} \in \mathbb{R}^{p\times p}$ is symmetric positive definite, then $\mathbb{E}[\mathbf{x}] = \mathbf{b}$ and $\mathrm{Cov}[\mathbf{x}] = \mathbf{A}^{-1}$.

Conversely, given a vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and a positive definite matrix $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$, there is a multivariate normal density

$$n(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p \det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

Bivariate Normal Distribution

We consider the (non-singular) bivariate normal distribution $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}.$$

We have

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}.$$

Let $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$, then we have $\det(\boldsymbol{\Sigma}) = \sigma_{11}\sigma_{22}(1 - \rho^2)$ and

$$\begin{aligned} &(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \\ =& \frac{1}{1 - \rho^2} \left(\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right)^2 + \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right) \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right) \right). \end{aligned}$$

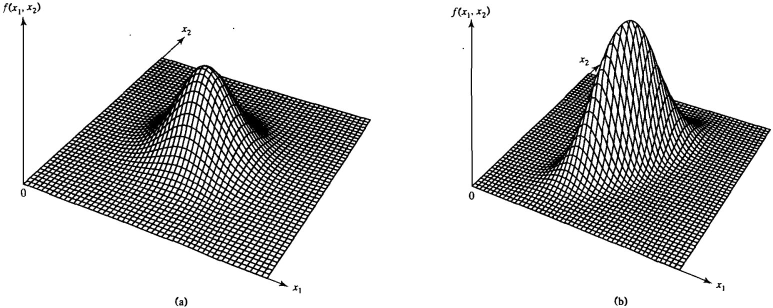
Bivariate Normal Density

The density function is

$$
\begin{aligned}
& f(x_1, x_2) \\
=& \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}\left(1-\rho^2\right)}} \\
& \times \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_{11}} + \frac{(x_2-\mu_2)^2}{\sigma_{22}} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sqrt{\sigma_{11}\sigma_{22}}}\right)\right).
\end{aligned}
$$

If $\rho = 0$, then the variables x_1 and x_2 are independent.

Bivariate Normal Density



Two bivariate normal distributions:

- (a) $\sigma_1 = \sigma_2$ and $\rho_{12} = 0$
- (b) $\sigma_1 = \sigma_2$ and $\rho_{12} = 0.75$

Probability Density Contour

The density of a p-dimensional normal variable

$$n(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

implies the multivariate normal density is constant on surfaces where the square of the distance $(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ is a constant.

These paths are called contours:

$$\left\{\mathbf{x} : (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2\right\} = \text{surface of an hyperellipsoid centered at } \boldsymbol{\mu},$$

where $c > 0$ is a fixed constant.

Probability Density Contour

Consider the hyperellipsoid with surface defined by $\mathbf{x} \in \mathbb{R}^p$ such that

$$(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2.$$

Denote the eigenvalue-eigenvector pairs of $\boldsymbol{\Sigma}$ by

$$(\lambda_1, \mathbf{u}_1),\ (\lambda_2, \mathbf{u}_2), \dots, (\lambda_p, \mathbf{u}_p),$$

then the hyperellipsoid is centered at $\boldsymbol{\mu}$ and have axes (vertices)

$$\pm c\sqrt{\lambda_1}\mathbf{u}_1,\ \ \pm c\sqrt{\lambda_2}\mathbf{u}_2, \dots,\ \ \pm c\sqrt{\lambda_p}\mathbf{u}_p.$$

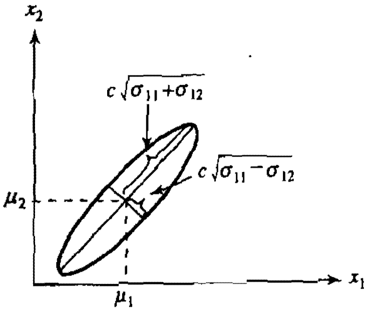
.

Probability Density Contour

For bivariate normal distribution with $\sigma_{11} = \sigma_{22}$, we have

$$\lambda_1 = \sigma_{11} + \sigma_{12}, \quad \mathbf{u}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \quad \lambda_2 = \sigma_{11} - \sigma_{12} \quad \text{and} \quad \mathbf{u}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

If we additionally suppose $\sigma_{12} > 0$, it leads to the following figure:



Probability Density Contour

For the hyperellipsoid with surface defined by $\mathbf{x} \in \mathbb{R}^p$ such that

$$(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2,$$

how the connect the eigenvalue-eigenvector pairs of $\boldsymbol{\Sigma}$ to the vertices of the hyperellipsoid?

The main idea:

Normally Distributed Variables

Some properties of normally distributed variables:

1. The linear transform of multivariate normal variates are normally distributed.
2. The marginal distributions derived from multivariate normal distributions are also normal distributions.
3. The conditional distributions derived from multivariate normal distributions are also normal distributions.

Outline

Theorem 1

Let $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} \succ \mathbf{0}$. Then

$$\mathbf{y} = \mathbf{C}\mathbf{x}$$

is distributed according to $\mathcal{N}_p(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top)$ for non-singular $\mathbf{C} \in \mathbb{R}^{p \times p}$.

Sketch of the proof:

1. Let $f(\mathbf{x})$ be the density function of $\mathbf{x}$.
2. Let $g(\mathbf{y})$ be the density function of $\mathbf{y}$.
3. The relation $\mathbf{x} = \mathbf{C}^{-1}\mathbf{y}$ implies

$$g(\mathbf{y}) = f(\mathbf{u}^{-1}(\mathbf{y}))|\det(\mathbf{J}^{-1}(\mathbf{y}))|$$

with $\mathbf{u}(\mathbf{x}) = \mathbf{C}\mathbf{x}$, $\mathbf{u}^{-1}(\mathbf{y}) = \mathbf{C}^{-1}\mathbf{y}$ and $\mathbf{J}^{-1}(\mathbf{y}) = \mathbf{C}^{-1}$.

Outline

Independence and Uncorrelatedness

Theorem

If $\mathbf{x} = [x_1, \ldots, x_p]^\top \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *with* $\boldsymbol{\Sigma} \succ \mathbf{0}$. *Let*

$$\mathbf{x}^{(1)} = [x_1, \ldots, x_q]^\top \quad \text{and} \quad \mathbf{x}^{(2)} = [x_{q+1}, \ldots, x_p]^\top$$

for $q < p$. *A necessary and sufficient condition for $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ to be independent is that each covariance of a variable from $\mathbf{x}^{(1)}$ and a variable from $\mathbf{x}^{(2)}$ is* 0.

1. The random vectors $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ can be replaced by any subset of $\mathbf{x}$ the subset consisting of the remaining variables respectively.
2. The necessity does not depend on the assumption of normality.

Quiz

If two random variables are normally distributed and uncorrelated, can we say they are independent?

Marginal Distribution

Corollary

We use the notation in above theorem such that

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$

It shows that if $\mathbf{x}^{(1)}$ is uncorrelated with $\mathbf{x}^{(2)}$, the marginal distribution of $\mathbf{x}^{(1)}$ is $\mathcal{N}(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}_{11})$ and the marginal distribution of $\mathbf{x}^{(2)}$ is $\mathcal{N}(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22})$.

In fact, this result holds even if the two sets are NOT uncorrelated.

Marginal Distribution

Theorem

If $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *with* $\boldsymbol{\Sigma} \succ \mathbf{0}$, *the marginal distribution of any set of components of*

$$\mathbf{x} = [x_1, x_2, \ldots, x_p]^\top$$

is multivariate normal with means, variances, and covariances obtained by taking the corresponding components of $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$, *respectively.*