MULTIVARIATE STATISTICAL ANALYSIS

Lecture 2
The Basic Concepts of MSA

Lecturer: Associate Professor Lý Quốc Ngọc

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

fit@hcmus

Contents

Contents

2.1. Multivariate Data

2.1.1. Arrays

$$
\begin{array}{cc}
 & \begin{array}{ccccc} V1 & V2 & \ldots\ldots Vk & \ldots\ldots Vp \end{array} \\
X = & \begin{bmatrix}
x_{11} & x_{12} \ldots\ldots & x_{1k} \ldots\ldots & x_{1p} \\
x_{21} & x_{22} \ldots\ldots & x_{2k} \ldots\ldots & x_{2p} \\
. & . & . & . \\
. & . & . & . \\
x_{j1} & x_{j2} \ldots\ldots & x_{jk} \ldots\ldots & x_{jp} \\
. & . & . & . \\
x_{n1} & x_{n2} \ldots\ldots & x_{nk} \ldots\ldots & x_{np}
\end{bmatrix}
\begin{array}{l} \text{Item 1} \\ \text{Item 2} \\ . \\ . \\ \text{Item j} \\ . \\ \text{Item n} \end{array}
\end{array}
$$

2.1. Multivariate Data

2.1.2. Descriptive Statistics

- Sample mean
- Sample variance
- Sample covariance
- Sample correlation coefficient

2.1. Multivariate Data

2.1.2. Descriptive Statistics

- Sample mean

$$\overline{x_k} = \frac{1}{n}\sum_{j=1}^{n} x_{jk}, \quad \text{k} = 1,2,\ldots,\text{p}$$

$$\overline{x} = \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \\ . \\ . \\ . \\ \overline{x_p} \end{bmatrix}$$

2.1. Multivariate Data

2.1.2. Descriptive Statistics

- Sample variance

$$S_k^2 = S_{kk} = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \overline{x_k})^2,\ \mathrm{k} = 1,2,\ldots,\mathrm{p}$$

$$S_k = \sqrt{S_{kk}}\ : \text{sample standard deviation}$$

2.1. Multivariate Data

2.1.2. Descriptive Statistics

❖ Sample covariance

$$S_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \overline{x_i})(x_{jk} - \overline{x_k}),\ \mathrm{i,k} = 1,2,\ldots,\mathrm{p}$$

$$S_{ik} = S_{ki}$$

2.1. Multivariate Data

2.1.2. Descriptive Statistics

- Sample variance & covariance

$$S_n = \begin{bmatrix} s_{11} & s_{12} \ldots\ldots\ldots s_{1p} \\ s_{21} & s_{22} \ldots\ldots\ldots s_{2p} \\ . & . & . \\ . & . & . \\ s_{j1} & s_{j2} \ldots\ldots\ldots s_{jp} \end{bmatrix}$$

2.1. Multivariate Data

2.1.2. Descriptive Statistics

❖ Sample correlation coefficient

$$r_{ik} = \frac{S_{ik}}{\sqrt{S_{ii}}\sqrt{S_{kk}}} = \frac{\sum_{j=1}^{n}(x_{ji} - \overline{x_i})(x_{jk} - \overline{x_k})}{\sqrt{\sum_{j=1}^{n}(x_{ji} - \overline{x_i})^2}\sqrt{\sum_{j=1}^{n}(x_{jk} - \overline{x_k})^2}},$$

$$\mathrm{i}, \mathrm{k} = 1, 2, \ldots, \mathrm{p}$$

$$r_{ik} = r_{ki} \quad \forall i, k$$

2.1. Multivariate Data

2.1.2. Descriptive Statistics

- Sample correlation coefficient

$$R_p = \begin{bmatrix} 1 \; r_{12} \ldots\ldots\ldots r_{1p} \\ r_{21} \; 1 \ldots\ldots\ldots r_{2p} \\ . \quad\quad . \quad\quad . \\ . \quad\quad . \quad\quad . \\ r_{p1} \; r_{p2} \ldots\ldots 1 \end{bmatrix}$$

2.1. Multivariate Data

2.1.3. Graphical Techniques

- Dot diagrams + Scatter plot
- Multiple scatter plot
- 3D scatter plot (for group structure)
- Graph of growth curves
- Stars
- Chernoff Faces

2.1. Multivariate Data

2.1.4. Statistical Distance

- Euclide Distance

$$P=(x_1,x_2,\ldots,x_p),\ Q=(y_1,y_2,\ldots,y_p)$$

$$d(P,Q)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_p-y_p)^2}$$

2.1. Multivariate Data

2.1.4. Statistical Distance

- Statistical Distance

$$d(P,Q)=\sqrt{\frac{(x_1-y_1)^2}{s_{11}}+\frac{(x_2-y_2)^2}{s_{22}}+\dots+\frac{(x_p-y_p)^2}{s_{pp}}}$$

$$d(P,Q)=\sqrt{\begin{array}{l}a_{11}(x_1-y_1)^2+a_{22}(x_2-y_2)^2+\dots+a_{pp}(x_p-y_p)^2+\\2a_{12}(x_1-y_1)(x_2-y_2)+2a_{13}(x_1-y_1)(x_3-y_3)+\dots+\\2a_{p-1p}(x_{p-1}-y_{p-1})(x_p-y_p)\end{array}}$$

2.1. Multivariate Data

2.1.4. Statistical Distance

- Statistical Distance

$$d(P,Q)=\begin{bmatrix}x_1-y_1 & x_2-y_2 \ldots & x_p-y_p\end{bmatrix}\begin{bmatrix}a_{11}\ a_{12}\ldots\ldots a_{1p}\\ a_{21}\ a_{22}\ldots\ldots a_{2p}\\ . \quad . \quad .\\ . \quad . \quad .\\ a_{j1}\ s_{j2}\ldots\ldots a_{jp}\end{bmatrix}\begin{bmatrix}x_1-y_1\\ x_2-y_2\\ .\\ .\\ x_p-y_p\end{bmatrix}$$

2.1. Multivariate Data

2.1.5. Sample Geometry

- Mean vector
- Deviation vector
- Variance
- Correlation Coefficient
- Generalized variance

2.1. Multivariate Data

2.1.5. Sample Geometry

- Mean vector

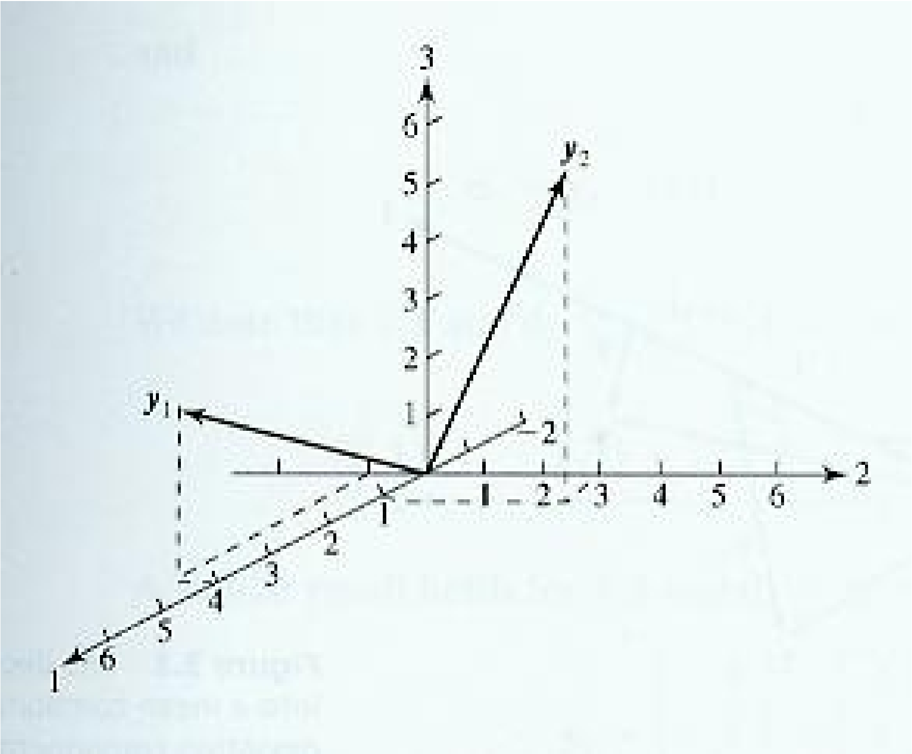
$$X = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

2.1. Multivariate Data

2.1.5. Sample Geometry

- Mean vector

$$X = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$



$$y_1 = [\ 4,\ -1, 3],$$
$$y_2 = [\ 1,\ 3,\ 5]$$

2.1. Multivariate Data

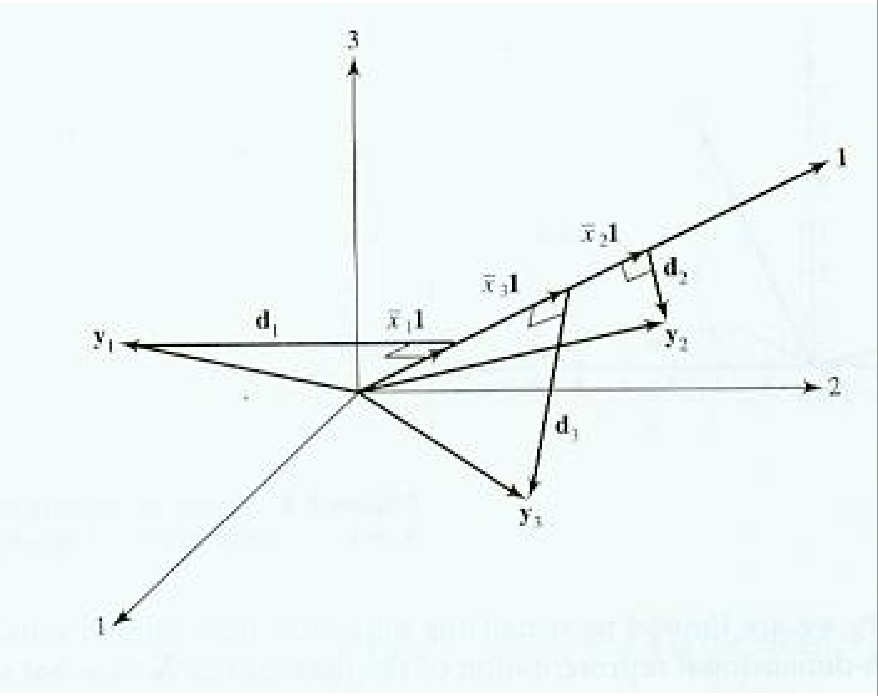
2.1.5. Sample Geometry

- Deviation vector

$$I_n = [1,1,\ldots 1]$$

$$y_i \frac{1}{\sqrt{n}} I_n \frac{1}{\sqrt{n}} I_n = \frac{x_{1i} + x_{2i} + \ldots + x_{ni}}{n} I_n = \overline{x_i} I_n$$

$$d_i = y_i - \overline{x_i} I_n = \begin{bmatrix} x_{1i} - \overline{x_i} \\ x_{2i} - \overline{x_i} \\ . \\ . \\ x_{2i} - \overline{x_i} \end{bmatrix}$$

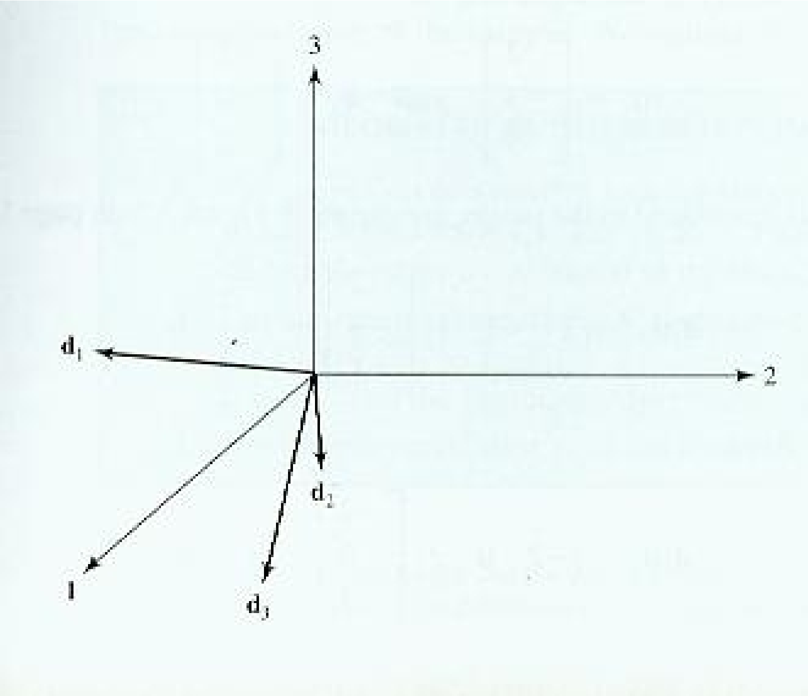


2.1. Multivariate Data

2.1.5. Sample Geometry

- Variance

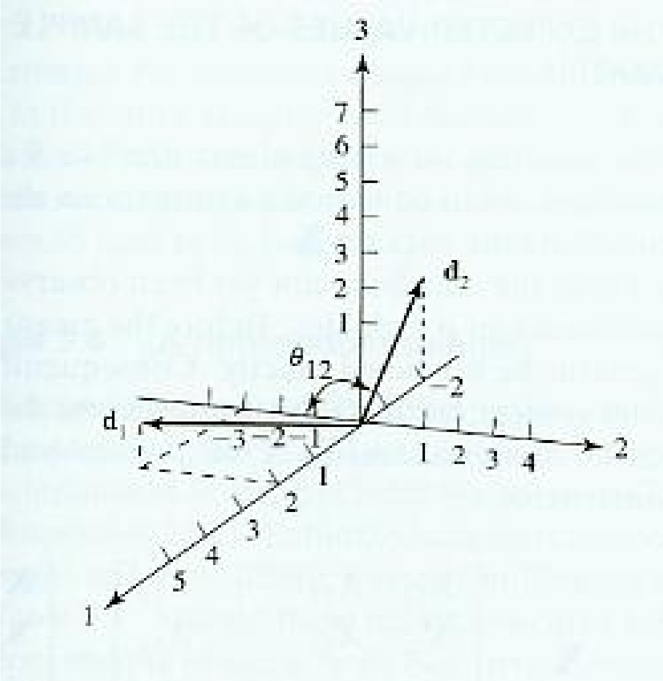
$$L_{d_i}^2 = d_i^{'} d_i = \sum_{j=1}^{n} (x_{ji} - \overline{x_i})^2$$



2.1. Multivariate Data

2.1.5. Sample Geometry

- Correlation Coefficient



$$\cos(\theta_{ik}) = \frac{d_i^{'} d_k}{L_{d_i} L_{d_k}} =$$

$$= \frac{\sum_{j=1}^{n} (x_{ji} - \overline{x}_i)(x_{jk} - \overline{x}_k)}{\sqrt{\sum_{j=1}^{n} (x_{ji} - \overline{x}_i)^2} \sqrt{\sum_{j=1}^{n} (x_{jk} - \overline{x}_k)^2}} = \frac{s_{ik}}{\sqrt{s_{ii}} \sqrt{s_{kk}}} = r_{ik}$$

2.1. Multivariate Data

2.1.5. Sample Geometry

- Generalized variance

$$S_{n-1} = \begin{bmatrix} s_{11} & s_{12} \ldots\ldots\ldots & s_{1p} \\ s_{21} & s_{22} \ldots\ldots\ldots & s_{2p} \\ . & . & . \\ . & . & . \\ s_{j1} & s_{j2} \ldots\ldots\ldots & s_{jp} \end{bmatrix} = \left\{ s_{ik} = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ji} - \overline{x_i})(x_{jk} - \overline{x_k}) \right\}$$

$$\text{Generalized sample variance} = \left| S_{n-1} \right| = (n-1)^{-p}\, volume^2$$

2.1. Multivariate Data

2.1.5. Sample Geometry

- Generalized variance

$$Generalized\ sample\ variance = |S_{n-1}| = (n-1)^{-p} volume^2$$

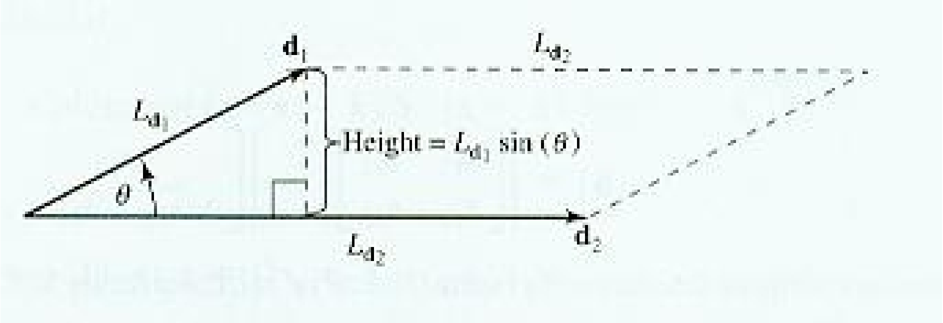
$$d_1 = y_1 - \overline{x_1} I_n, d_2 = y_2 - \overline{x_2} I_n, \ldots, d_p = y_p - \overline{x_p} I_n$$

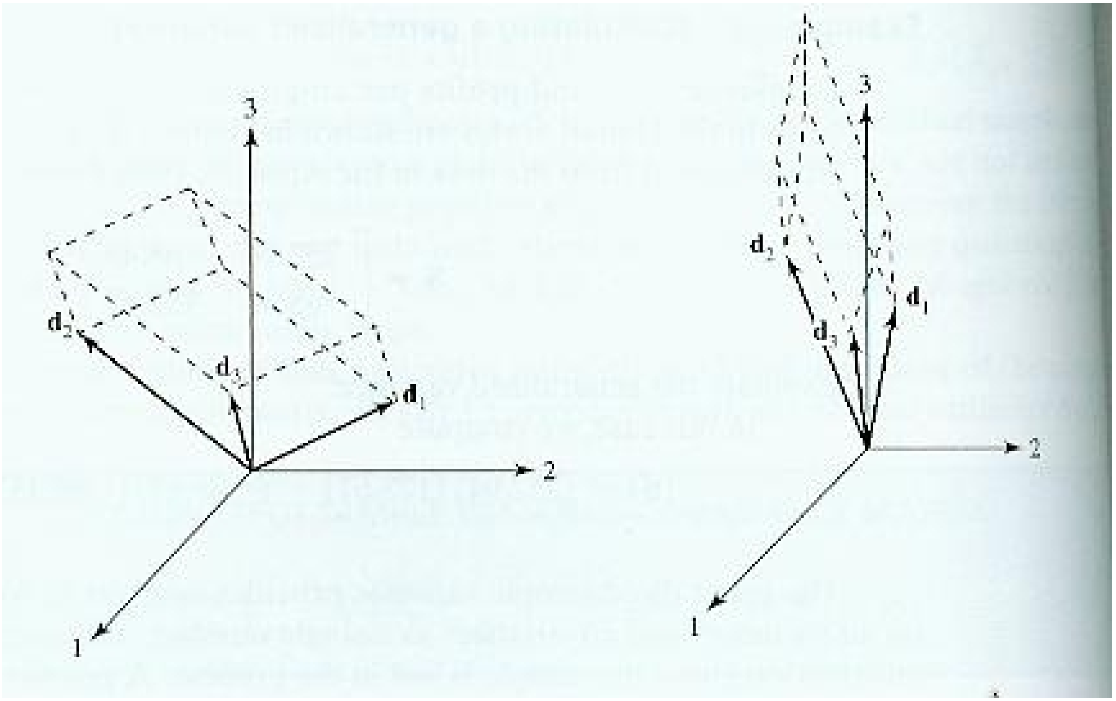
2.1. Multivariate Data

fit@hcmus

2.1.5. Sample Geometry

- Generalized variance





2.1. Multivariate Data

2.1.5. Sample Geometry

- Generalized variance

$$\text{Generalized sample variance of standardized variable} =$$

$$|R| = (n-1)^{-p} \, volume^2$$

$$d_1 = (y_1 - \overline{x_1} I_n) \Big/ \sqrt{s_{11}}, \; d_2 = (y_2 - \overline{x_2} I_n) \Big/ \sqrt{s_{22}}, \ldots,$$

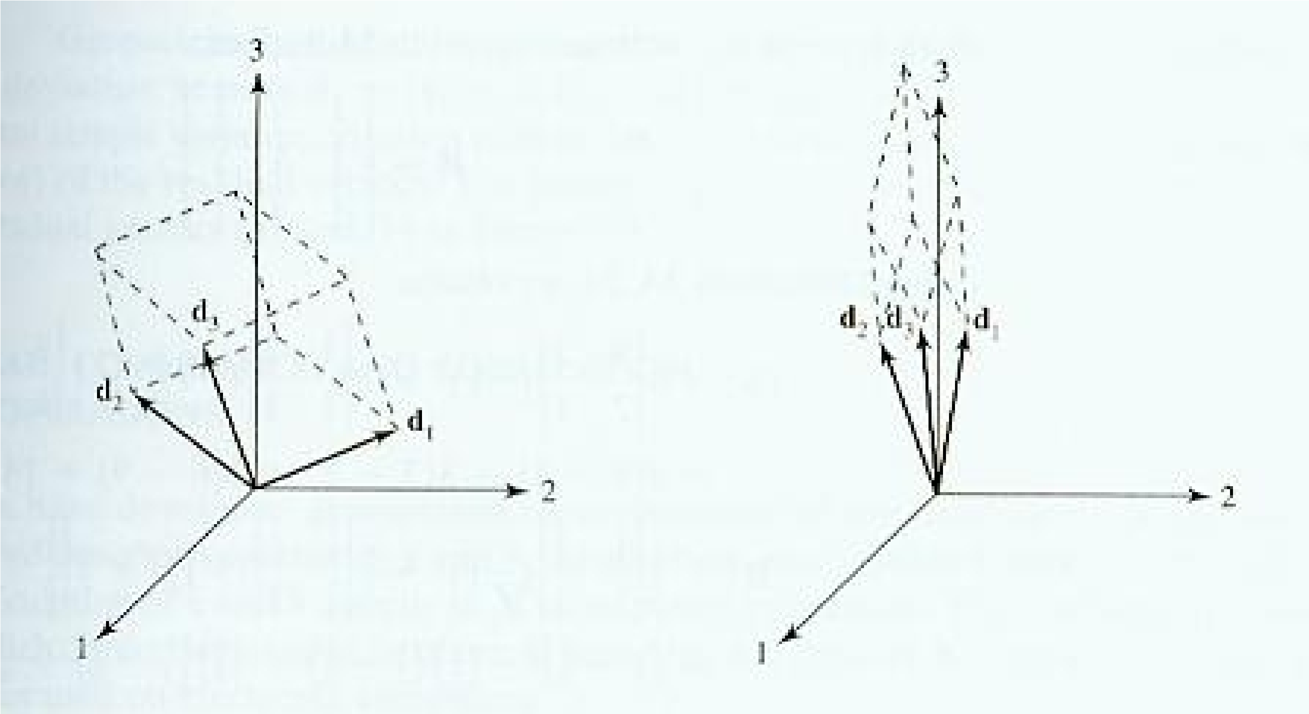
$$d_p = (y_p - \overline{x_p} I_n) \Big/ \sqrt{s_{pp}}$$

2.1. Multivariate Data

2.1.5. Sample Geometry

- Generalized variance

Generalized sample variance of standardized variable



Contents

2.2. Random Multivariate Data

2.2.1. Random Vectors and Matrices

- A random vector is a vector whose elements are random variables.
- A random matrix is a matrix whose elements are random variables

2.2. Random Multivariate Data

2.2.2. Expectation of Random Vectors and Matrices

$$X = \{X_{ij}\}, n \times p \; random\; matrix$$

$$E(X) = \begin{bmatrix} E(X_{11})\; E(X_{12}) \ldots E(X_{1p}) \\ E(X_{21})\; E(X_{22}) \ldots E(X_{2p}) \\ . \\ . \\ E(X_{n1})\; E(X_{n2}) \ldots E(X_{np}) \end{bmatrix}$$

$$E(X_{ij}) = \begin{Bmatrix} \int_{-\infty}^{\infty} x_{ij} f_{ij}(x_{ij}) dx_{ij} \;\text{ if } X_{ij} \text{ is continuous random variable with pdf } f_{ij}(x_{ij}) \\ \sum_{x_{ij}} x_{ij} p_{ij}(x_{ij}) \;\text{ if } X_{ij} \text{ is discrete random variable with probability function } p_{ij}(x_{ij}) \end{Bmatrix}$$

2.2. Random Multivariate Data

2.2.3. Mean Vector

$$X' = \{X_1, X_2, ..., X_p\}, p \times 1 \; random \text{ vector}$$

$$\mu_i = E(X_i), \mathrm{i} = 1, 2, ..., \mathrm{p}$$

$$\mu_i = \begin{Bmatrix} \int_{-\infty}^{\infty} x_i f_i(x_i) dx_i \;\; \text{if } X_i \text{ is continuous random variable with pdf } f_i(x_i) \\ \sum_{x_i} x_i p_i(x_i) \;\; \text{if } X_i \text{ is discrete random variable with probability function } p_i(x_i) \end{Bmatrix}$$

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

2.2. Random Multivariate Data

2.2.4. Variance and Covariance Matrices

$$X' = \{X_1, X_2, \ldots, X_p\}, \; p \times 1 \; random \text{ vector}$$

$$\sigma_{\mathrm{i}}^2 = E(X_i - \mu_i)^2 \;, \mathrm{i} = 1,2,\ldots,\mathrm{p}$$

$$\sigma_i^2 = \begin{cases} \int\limits_{-\infty}^{\infty} (x_i - \mu_i)^2 f_i(x_i) dx_i \;\; \text{if } X_{ij} \text{ is continuous random variable with pdf } f_i(x_i) \\ \sum\limits_{x_i} (x_i - \mu_i)^2 p_i(x_i) \;\; \text{if } X_i \text{ is discrete random variable with probability function } p_i(x_i) \end{cases}$$

$$\sigma_{ik} = E(X_i - \mu_i)(X_k - \mu_k) \;, \mathrm{i}, \mathrm{k} = 1,2,\ldots,\mathrm{p}$$

$$\sigma_{ik} = \begin{cases} \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} (x_i - \mu_i)(x_k - \mu_k) f_{ik}(x_i, x_k dx_i dx_k \;\; \text{if } X_i, X_k \text{ are continuous random variable} \\ \qquad\qquad\qquad \text{with jdf } f_{ik}(x_i, x_k) \\ \sum\limits_{x_i} \sum\limits_{x_k} (x_i - \mu_i)(x_k - \mu_k) p_{ik}(x_i, x_k) \;\; \text{if } X_i, X_k \text{ are discrete random variable} \\ \qquad\qquad\qquad \text{with probability function } p_{ik}(x_i, x_k) \end{cases}$$

2.2. Random Multivariate Data

2.2.4. Variance and Covariance Matrices

$$X' = \{X_1, X_2, ..., X_p\}, p \times 1 \, random \text{ vector}$$

$$\Sigma = E(X-\mu)(X-\mu)'$$

$$= E\begin{bmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ . \\ . \\ X_p-\mu_p \end{bmatrix}\left[X_1-\mu_1 \;\; X_2-\mu_2 \;\; ... X_p-\mu_p\right]$$

$$= \begin{bmatrix} E(X_1-\mu_1)^2 & E(X_1-\mu_1)(X_2-\mu_2) \; . \; . \; . & E(X_1-\mu_1)(X_p-\mu_p) \\ E(X_2-\mu_2)(X_1-\mu_1) & E(X_2-\mu_2)^2 \quad . \; . \; . & E(X_2-\mu_2)(X_p-\mu_p) \\ . & & \\ . & & \\ E(X_p-\mu_p)(X_1-\mu_1) & E(X_p-\mu_p)^2 \quad . \; . \; . & E(X_p-\mu_p)^2 \end{bmatrix}$$

fit@hcmus

2.2. Random Multivariate Data

2.2.4. Variance and Covariance Matrices

$X' = \{X_1, X_2, ..., X_p\}$, $p \times 1$ *random* vector

$$\Sigma = Cov(X) = \begin{bmatrix} \sigma_{11}\ \sigma_{12}\ \dots \sigma_{1p} \\ \sigma_{21}\ \sigma_{22}\ \dots \sigma_{2p} \\ . \\ . \\ \sigma_{p1}\ \sigma_{p2}\ \dots \sigma_{pp} \end{bmatrix}$$

2.2. Random Multivariate Data

2.2.5. Correlation Matrix

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}}$$

$$\rho = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{21}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{11}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} \\ . & & \\ . & & \\ \frac{\sigma_{p1}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{11}}} & \frac{\sigma_{p2}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{22}}} \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{bmatrix}$$

2.2. Random Multivariate Data

2.2.5. Correlation Matrix

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}}$$

$$\rho = \begin{bmatrix} 1 & \rho_{12} \; . \; . \; . & \rho_{1p} \\ \rho_{21} & 1 . \; . \; . & \rho_{2p} \\ . & & \\ . & & \\ \rho_{p1} & \rho_{p2} \; . \; . \; . & 1 \end{bmatrix}$$

Contents

2.3. The Multivariate Normal Distribution

2.3.1. Introduction

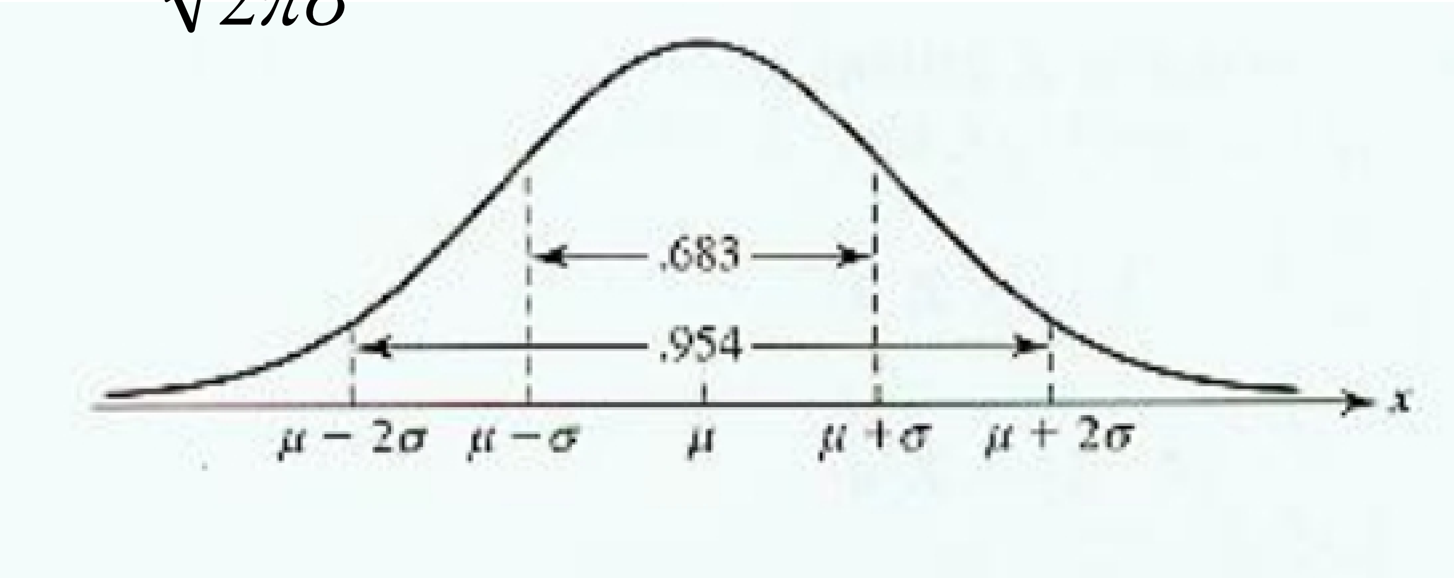
???

2.3. The Multivariate Normal Distribution

2.3.2. The Multivariate Normal Density Function

$p = 1, N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)/\sigma]^2/2}, -\infty < x < \infty$$



2.3. The Multivariate Normal Distribution

2.3.2. The Multivariate Normal Density Function

Consider random vector x on p different variables.

$$N_p(\mu,\Sigma)$$

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

$$-\infty < x_i < \infty, \quad \mathrm{i}=1,2,\ldots\mathrm{p}$$

2.3. The Multivariate Normal Distribution

2.3.2. The Multivariate Normal Density Function

Consider random vector x on 2 different variables.

$$N_2(\mu, \Sigma)$$

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}$$

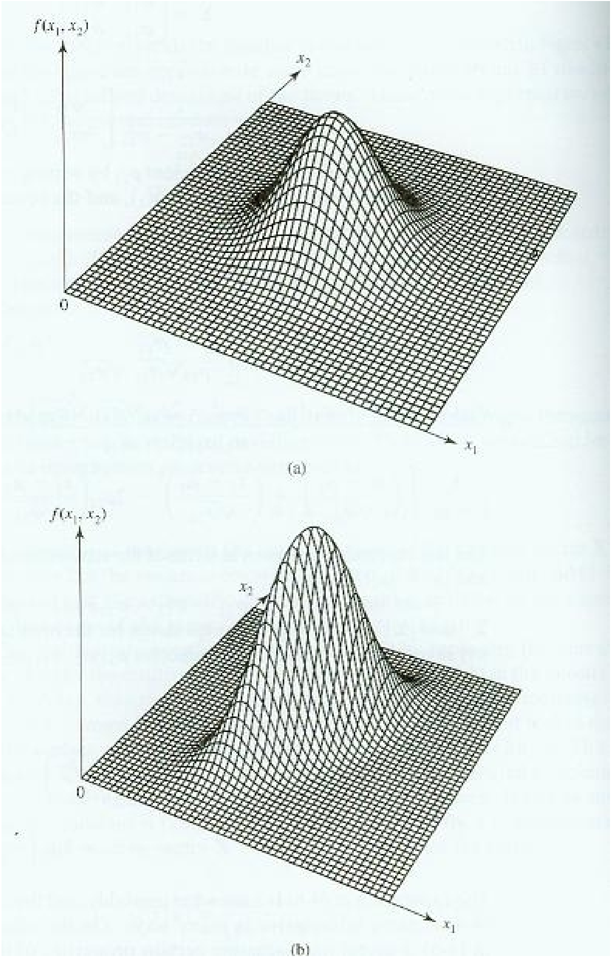
$$\times \exp\left\{ -\frac{1}{2(1-\rho_{12}^2)} \left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}} \right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}} \right)^2 - 2\rho_{12} \left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}} \right) \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}} \right) \right] \right\},$$

2.3. The Multivariate Normal Distribution

2.3.2. The Multivariate Normal Density Function

Consider random vector x on 2 different variables.

$$N_2(\mu, \Sigma)$$



2.4. Detecting Outliers and Cleaning Data

Steps for Detecting Outliers

1. Make a dot plot for each variable.

2. Make a scatter plot for each pairs of variables.

3. Calculate the standardized values. Examine these standardized values for large or small values.

$$z_{jk} = (x_{jk} - \bar{x}_k) \Big/ \sqrt{s_{k,k}},\ j = 1,2,\ldots n;\ k = 1,2,\ldots,p$$

4. Calculate the generalized squared distances. Examine these distances for unusually large values.

$$(x_j - \bar{x})' S^{-1} (x_j - \bar{x})$$

Contents

2.5. Some basic theorems

2.5.1. Maximum Likelihood Estimation

$$\begin{Bmatrix} \text{Joint density} \\ of\ X_1, X_2, \ldots, X_n \end{Bmatrix} = \prod_{j=1}^{n} \left\{ \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-(x_j-\mu)^T \Sigma^{-1} (x_j-\mu)/2} \right\}$$

$$= \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} e^{-\sum_{j=1}^{n} (x_j-\mu)^T \Sigma^{-1} (x_j-\mu)/2}$$

2.5. Some basic theorems

2.5.1. Maximum Likelihood Estimation

$$\left\{\begin{matrix}\text{Joint density} \\ of\ X_1, X_2, \ldots, X_n\end{matrix}\right\} = L(\mu, \Sigma)$$

$$= \frac{1}{(2\pi)^{np/2} \, |\Sigma|^{n/2}} \cdot$$

$$\exp\left\{-tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j - \bar{x})\cdot(x_j - \bar{x})^T + n.(\bar{x} - \mu)\cdot(\bar{x} - \mu)^T\right)/2\right]\right\}$$

2.5. Some basic theorems

2.5.1. Maximum Likelihood Estimation

$$Log(L(\mu,\Sigma)) = Log\left(\frac{1}{(2\pi)^{np/2}\,|\Sigma|^{n/2}}\right) -$$

$$-tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\overline{x})\cdot(x_j-\overline{x})^T + n.(\overline{x}-\mu)\cdot(\overline{x}-\mu)^T\right)\right]/2$$

$$= Log\left(\frac{1}{(2\pi)^{np/2}\,|\Sigma|^{n/2}}\right) - tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\overline{x})\cdot(x_j-\overline{x})^T\right)/2\right] -$$

$$n(\overline{x}-\mu)^T\,\Sigma^{-1}(\overline{x}-\mu)/2\cdot$$

fit@hcmus

2.5. Some basic theorems

2.5.1. Maximum Likelihood Estimation

$$\frac{\partial Log(L(\mu,\Sigma))}{\partial \mu} = \frac{1}{2}2.(\sum_{j=1}^{n}(x_j - \mu)^T)\Sigma^{-1} = 0$$

$$(\sum_{j=1}^{n} x_j - n\mu).\Sigma^{-1} = 0$$

$$\mu = \frac{1}{n}\sum_{j=1}^{n} x_j$$

2.5. Some basic theorems

2.5.1. Maximum Likelihood Estimation

$$\frac{\partial Log(L(\mu,\Sigma))}{\partial \Sigma^{-1}} = \frac{n}{2}(2M - DiagM) = 0, (M = \Sigma - S - (\bar{x} - \mu).(\bar{x} - \mu)^T)$$

$$\Rightarrow M = 0$$

$$\Rightarrow \Sigma = S + (\bar{x} - \mu).(\bar{x} - \mu)^T = S$$

2.5. Some basic theorems

2.5.2. Law of large numbers

Let $X_1, X_2, ..., X_n$ be independent observations from a population with mean $E(X_i) = \mu$. Then

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

We have: $p[-\varepsilon < \overline{X} - \mu < \varepsilon] \to 1 \; khi \; n \to \infty$

Result: $p[-\varepsilon < S - \sum < \varepsilon] \to 1 \; khi \; n \to \infty$

2.5. Some basic theorems

2.5.2. The central limit theorem

Let $X_1, X_2, ..., X_n$ be independent observations from a population with mean μ and finite covariance Σ. Then $\sqrt{n}(\overline{X} - \mu)$ has an approximate $N_p(0, \Sigma)$ distribution for large sample sizes.

$n(\overline{X} - \mu)^T S^{-1}(\overline{X} - \mu)$ is approximately χ_p^2 for n-p large.