KHOA CÔNG NGHỆ THÔNG TIN-ĐHKHTN

PHÂN TÍCH THỐNG KÊ DỮ LIỆU NHIỀU BIẾN

Giảng viên: PGS.TS. Lý Quốc Ngọc

TPHCM, 8-2020

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

fit@hcmus

PHÂN TÍCH THỐNG KÊ
DỮ LIỆU NHIỀU BIẾN

Bài giảng 2:
Các khái niệm cơ bản về PTTKDLNB

Giảng viên: PGS.TS. Lý Quốc Ngọc

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

fit@hcmus

Nội dung

2. Các khái niệm cơ bản trong PTTKDLNB

2.1. Dữ liệu nhiều chiều

2.2. Đại lượng ngẫu nhiên

2.3. Hàm phân bố chuẩn nhiều biến

2.4. Phát hiện dữ liệu kỳ dị và làm sạch dữ liệu

2.5. Một số định luật cơ bản

Nội dung dữ liệu nhiều chiều

2.1. Dữ liệu nhiều chiều

2.1.1. Cấu trúc mảng

2.1.2. Các đại lượng thống kê

2.1.3. PP thể hiện dữ liệu

2.1.4. Độ đo thống kê

2.1.5. Ý nghĩa hình học

2.1. Dữ liệu nhiều chiều

2.1.1. Cấu trúc mảng

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

2.1. Dữ liệu nhiều chiều

2.1.2. Các đại lượng thống kê

- Trung bình mẫu (Sample mean)
- Phương sai mẫu (Sample variance)
- Hiệp phương sai mẫu (Sample covariance)
- Hệ số tương quan mẫu (Sample correlation coefficient)

2.1. Dữ liệu nhiều chiều

2.1.2. Các đại lượng thống kê

- Trung bình trên tập mẫu (Sample mean)

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

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

2.1. Dữ liệu nhiều chiều

2.1.2. Các đại lượng thống kê

- Phương sai trên tập mẫu (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. Dữ liệu nhiều chiều

2.1.2. Các đại lượng thống kê

- Hiệp phương sai trên tập mẫu (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. Dữ liệu nhiều chiều

2.1.2. Các đại lượng thống kê

❖ Ma trận Phương sai và Hiệp phương sai trên tập mẫu (Sample variance & covariance)

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

2.1. Dữ liệu nhiều chiều

2.1.2. Các đại lượng thống kê

❖ Hệ số tương quan trên tập mẫu (Sample correlation coefficient)

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

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

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

2.1. Dữ liệu nhiều chiều

2.1.2. Các đại lượng thống kê

- Hệ số tương quan trên tập mẫu (Sample correlation coefficient)

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

2.1. Dữ liệu nhiều chiều

2.1.3. Phương pháp trực quan hóa dữ liệu

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

2.1. Dữ liệu nhiều chiều

2.1.4. Độ đo thống kê

Độ đo Euclide

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

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

2.1. Dữ liệu nhiều chiều

2.1.4. Độ đo thống kê

Độ đo thống kê (Statistical Distance)

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

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

2.1. Dữ liệu nhiều chiều

2.1.4. Độ đo thống kê

Độ đo thống kê (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. Dữ liệu nhiều chiều

2.1.5. Ý nghĩa hình học

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

2.1. Dữ liệu nhiều chiều

2.1.5. Ý nghĩa hình học

- Mean vector

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

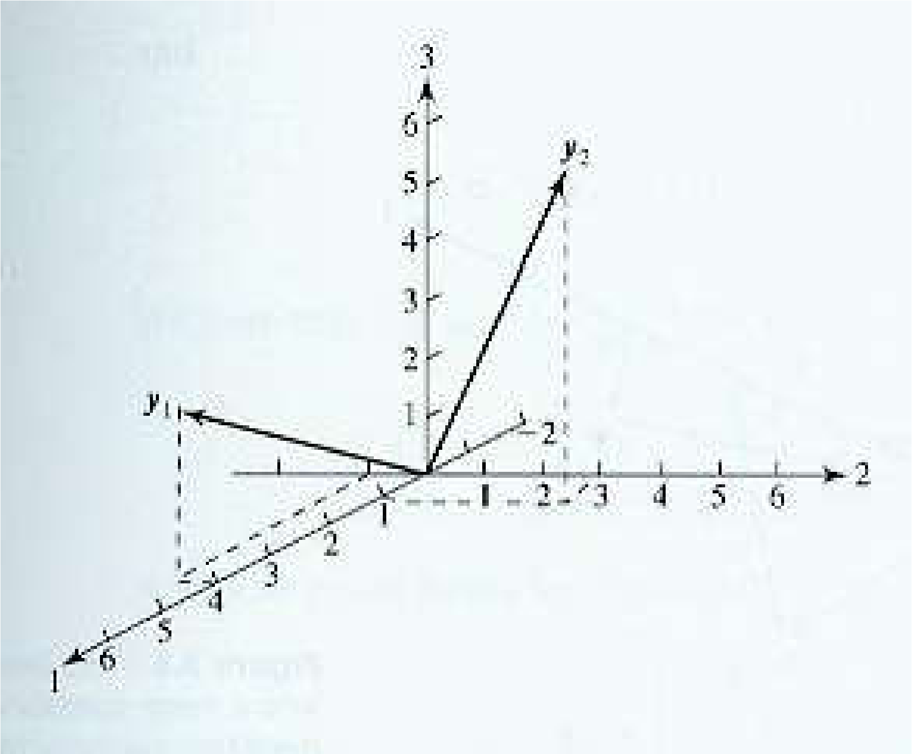
2.1. Dữ liệu nhiều chiều

2.1.5. Ý nghĩa hình học

- 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. Dữ liệu nhiều chiều

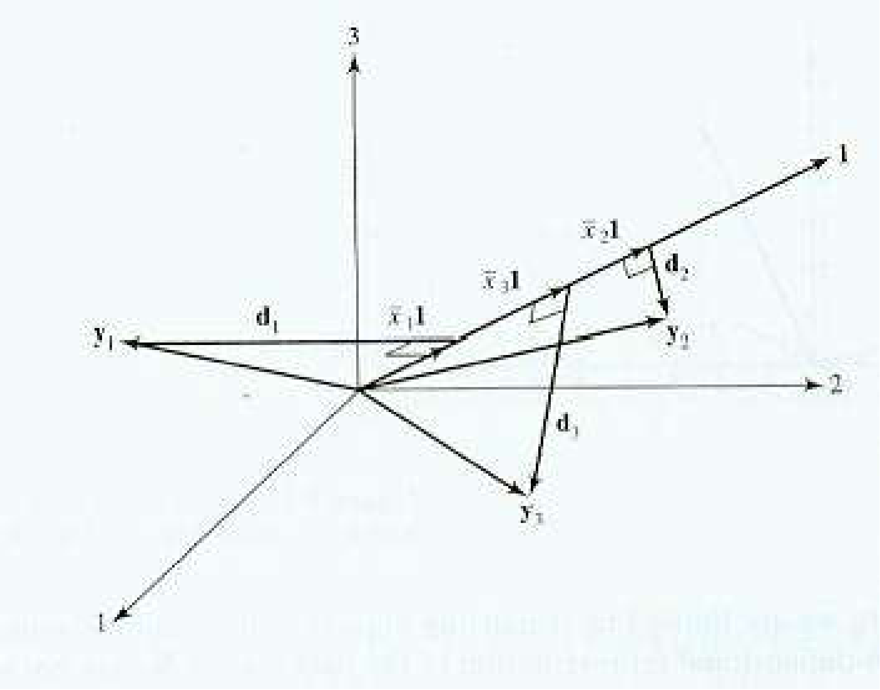
2.1.5. Ý nghĩa hình học

- Deviation vector

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

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

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

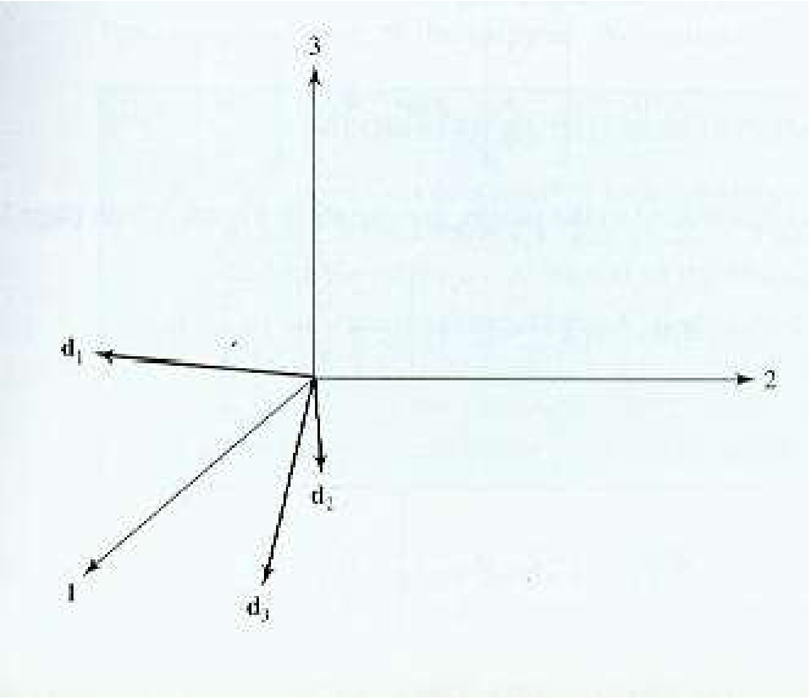


2.1. Dữ liệu nhiều chiều

2.1.5. Ý nghĩa hình học

- Variance

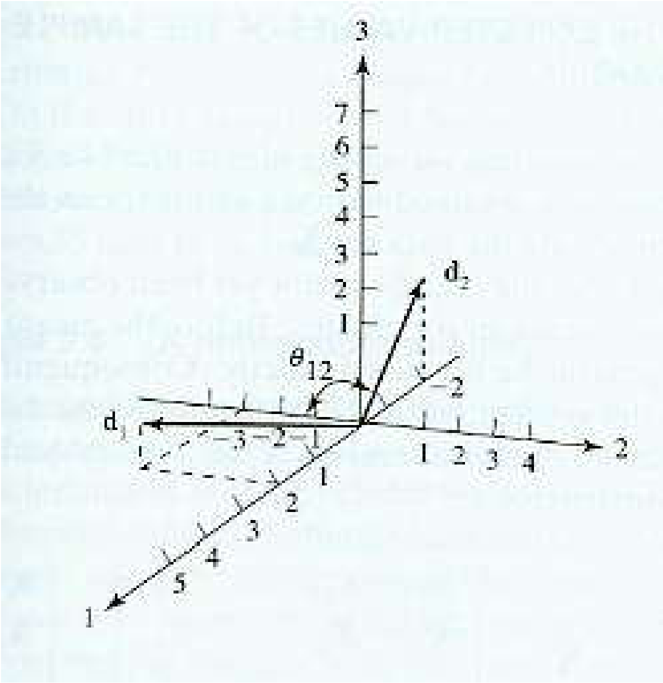
$$L_{d_i}^2 = d_i^{'} d_i = \sum_{j=1}^{n} (x_{ji} - \overline{x_i})^2$$



2.1. Dữ liệu nhiều chiều

2.1.5. Ý nghĩa hình học

- 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. Dữ liệu nhiều chiều

2.1.5. Ý nghĩa hình học

- Generalized variance

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

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

2.1. Dữ liệu nhiều chiều

2.1.5. Ý nghĩa hình học

- Generalized variance

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

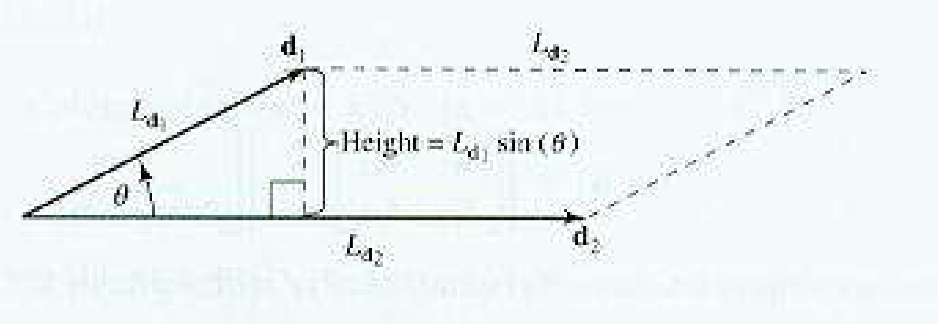
$$d_1 = y_1 - \overline{x_1} I_n, d_2 = y_2 - \overline{x_2} I_n, \dots, d_p = y_p - \overline{x_p} I_n$$

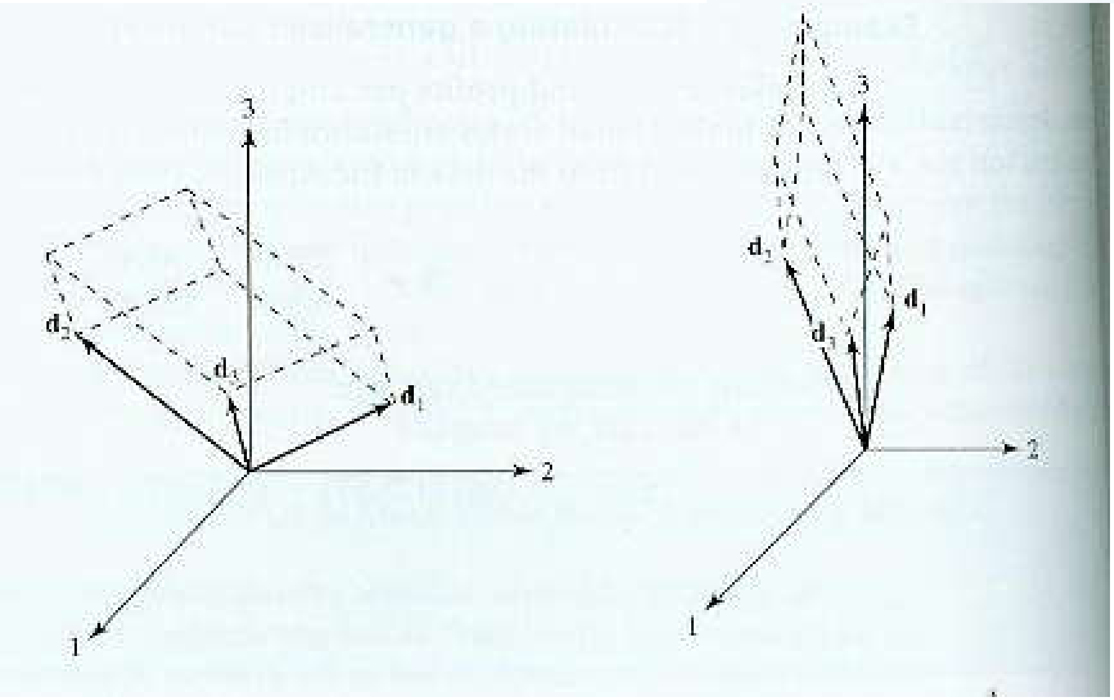
2.1. Dữ liệu nhiều chiều

fit@hcmus

2.1.5. Ý nghĩa hình học

- Generalized variance





2.1. Dữ liệu nhiều chiều

2.1.5. Ý nghĩa hình học

- Generalized variance

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

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

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

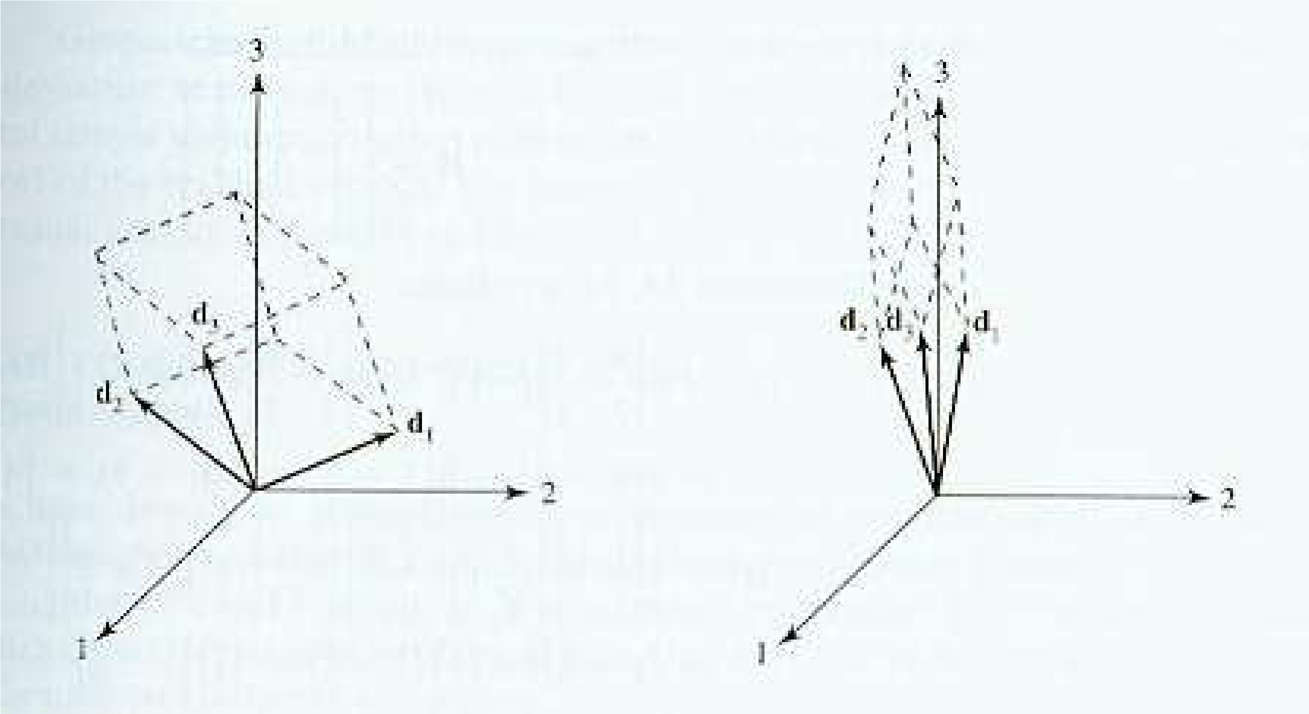
$$d_p = (y_p - \overline{x_p} I_n) \big/ \sqrt{s_{pp}}$$

2.1. Dữ liệu nhiều chiều

2.1.5. Ý nghĩa hình học

- Generalized variance

Generalized sample variance of standardized variable



Nội dung **đại lượng ngẫu nhiên**

2.2. Đại lượng ngẫu nhiên

2.2.1. Vector ngẫu nhiên

2.2.2. Kỳ vọng của đại lượng ngẫu nhiên

2.2.3. Vector trung bình

2.2.4. Ma trận phương sai và hiệp phương sai

2.2.5. Ma trận tương quan

2.2. Đại lượng ngẫu nhiên

2.2.1. Vector ngẫu nhiên

- Vector ngẫu nhiên là vector mà các phần tử của nó là các biến ngẫu nhiên.

- Ma trận ngẫu nhiên là ma trận mà các phần tử của nó là các biến ngẫu nhiên.

2.2. Đại lượng ngẫu nhiên

2.2.2. Kỳ vọng của đại lượng ngẫu nhiên

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

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

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

2.2. Đại lượng ngẫu nhiên

2.2.3. Vector trung bình (mean vector)

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

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

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

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

2.2. Đại lượng ngẫu nhiên

2.2.4. Ma trận phương sai và hiệp phương sai

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

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

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

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

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

2.2. Đại lượng ngẫu nhiên

2.2.4. Ma trận phương sai và hiệp phương sai

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

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

$$= E\begin{bmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ . \\ . \\ X_p-\mu_p \end{bmatrix}\begin{bmatrix} X_1-\mu_1 & X_2-\mu_2 & \ldots X_p-\mu_p \end{bmatrix}$$

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

2.2. Đại lượng ngẫu nhiên

2.2.4. Ma trận phương sai và hiệp phương sai

$X' = \{X_1, X_2, ..., X_p\}, p \times 1$ *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. Đại lượng ngẫu nhiên

2.2.5. Ma trận tương quan (population 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. Đại lượng ngẫu nhiên

2.2.5. Ma trận tương quan (population correlation matrix)

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

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

Nội dung **Hàm phân bố chuẩn nhiều biến**

2.3. Hàm phân bố chuẩn nhiều biến

2.3.1. Ý nghĩa

2.3.2. Dạng tổng quát

2.3.3. Tính chất

2.3. Hàm phân bố chuẩn nhiều biến

2.3.1. Ý nghĩa

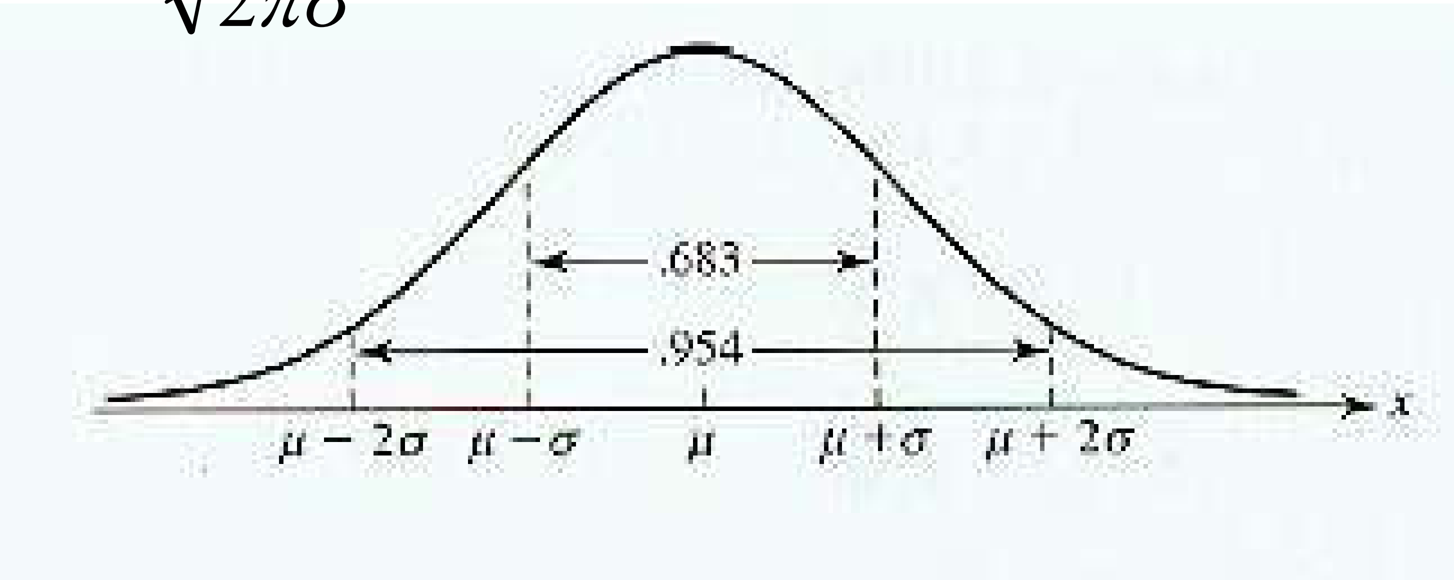
???

2.3. Hàm phân bố chuẩn nhiều biến

2.3.2. Dạng tổng quát

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

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



2.3. Hàm phân bố chuẩn nhiều biến

2.3.2. Dạng tổng quát

- Xét đại lượng ngẫu nhiên x có p biến

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

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

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

2.3. Hàm phân bố chuẩn nhiều biến

2.3.2. Dạng tổng quát

- Xét đại lượng ngẫu nhiên x có 2 biến

$N_2(\mu, \Sigma)$

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

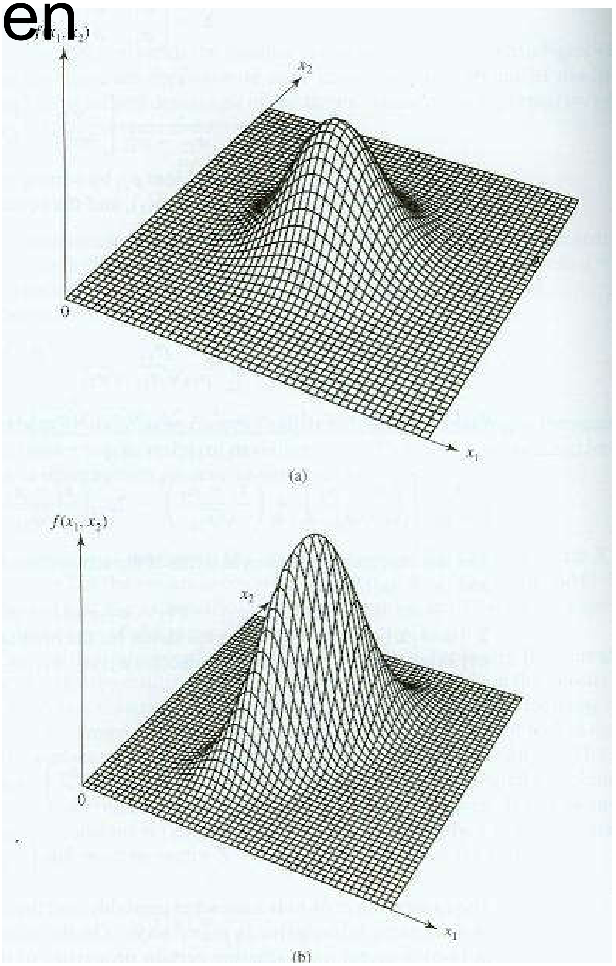
$$\times \exp\left\{ -\frac{1}{2(1-\rho_{12}^2)} \left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right) \right] \right\},$$

2.3. Hàm phân bố chuẩn nhiều biến

2.3.2. Dạng tổng quát

- Xét đại lượng ngẫu nhiên x có 2 biến

$N_2(\mu, \Sigma)$



2.4. Phát hiện dữ liệu kỳ dị và làm sạch dữ liệu

Các bước phát hiện mẫu ngoại lai

1. Tạo dot plot với mỗi biến
2. Tạo scatter plot với mỗi cặp biến
3. Tính các đại lượng chuẩn hóa.

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

Chú ý các đại lượng có giá trị lớn hoặc bé.

4. Tính khoảng cách thống kê từ mẫu đến trung bình mẫu.

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

Chú ý các khoảng cách có giá trị lớn bất thường.

Nội dung **một số định luật cơ bản**

2.5. Một số định luật cơ bản

2.5.1. Ước lượng triển vọng cực đại

2.5.2. Luật số lớn

2.5.3. Định lý giới hạn trung tâm

2.5. Một số định luật cơ bản

2.5.1. Ước lượng triển vọng cực đại

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

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

fit@hcmus

2.5. Một số định luật cơ bản

2.5.1. Ước lượng triển vọng cực đại

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

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

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

2.5. Một số định luật cơ bản

2.5.1. Ước lượng triển vọng cực đại

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

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

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

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

2.5. Một số định luật cơ bản

2.5.1. Ước lượng triển vọng cực đại

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

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

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

2.5. Một số định luật cơ bản

2.5.1. Ước lượng triển vọng cực đại

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

$$\Rightarrow M = 0$$

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

2.5. Một số định luật cơ bản

2.5.2. Luật số lớn

$X_1, X_2, ..., X_n$ là các mẫu khảo sát độc lập từ quần thể có $E(X_i) = \mu$

Đặt: $\overline{X} = \dfrac{X_1 + X_2 + ... + X_n}{n}$

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

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

2.5. Một số định luật cơ bản

2.5.3. Định lý giới hạn trung tâm

$X_1, X_2, ..., X_n$ là các mẫu khảo sát độc lập từ quần thể có μ, Σ

$\sqrt{n}(\overline{X} - \mu)$ có phân bố gần đúng với phân bố chuẩn $N_p(0, \Sigma)$

khi số mẫu đủ lớn.

$n(\overline{X} - \mu)^T S^{-1} (\overline{X} - \mu)$ xấp xỉ với χ_p^2

khi số mẫu đủ lớn.