BÀI GIẢNG PHÂN TÍCH THỐNG KÊ NHIỀU CHIỀU

Phạm Đình Tùng
VNU University of Science

Ngày 19 tháng 2 năm 2024

Giới thiệu về môn học

Nội dung môn học

1. Nhắc lại một số Kiến thức về Ma trận, mô hình hồi quy đơn, Phân bố chuẩn một chiều, hai chiều.
2. Vecto ngẫu nhiên, phân bố chuẩn nhiều chiều và Mẫu ngẫu nhiên
3. Phương pháp giảm số chiều dữ liệu: Phân tích thành phần chính, phân tích nhân tố
4. Mô hình hồi quy tuyến tính bội
5. Phân tích phân biệt và phân lớp

Giáo trình

- Nguyễn Văn Hữu, Nguyễn Hữu Dư, (2004). Phân tích thống kê và dự báo, NXB ĐHQGHN
- Bài giảng slide, tài liệu giảng viên cung cấp.
- Daniel Zelterman (2016). Applied multivariate statistics with

Vecto

Bộ n số thực $(x_1, \ldots, x_n)$ được gọi là vector n chiều

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, x' = (x_1, \ldots, x_n), cx = \begin{bmatrix} cx_1 \\ cx_2 \\ \vdots \\ cx_n \end{bmatrix}$$

$$x + y = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix};$$

$$|x| = (x, x)^{\frac{1}{2}} = (x_1^2 + x_2^2 + \cdots + x_n^2)^{\frac{1}{2}} = (\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}};$$

Ma trận

Ma trận A là một bảng số chữ nhật

$$A_{(n\times p)} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix}, A'_{(n\times p)} = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \dots & \dots & \dots & \dots \\ a_{p1} & a_{p2} & \dots & a_{pn} \end{bmatrix},$$

$$A_{(n\times p)} + B_{(n\times p)} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1p} \\ b_{21} & b_{22} & \dots & b_{2p} \\ \dots & \dots & \dots & \dots \\ b_{n1} & {}_{n2} & \dots & b_{np} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \dots & a_{1p}+b_{1p} \\ a_{21}+b_{21} & a_{22}+b_{22} & \dots & a_{2p}+b_{2p} \\ \dots & \dots & \dots & \dots \\ a_{n1}+b_{n1} & a_{n2}+b_{n2} & \dots & a_{np}+b_{np} \end{bmatrix}$$

Phép nhân hai ma trận

Cho $A = [a_{ij}]_{(n\times p)}, B = [b_{ij}]_{(p\times m)}$. Khi đó $C = [c_{ij}]_{(n\times m)}$ gọi là tích của A và B nếu $c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$ và kí hiệu $C = A \times B$.

Ví dụ: $A_{(2\times 3)} = \begin{bmatrix} 3 & -1 & 2 \\ 1 & 5 & 4 \end{bmatrix}, B_{(3\times 1)} = \begin{bmatrix} -2 \\ 7 \\ 9 \end{bmatrix}$. Khi đó

$A_{(2\times 3)} \times B_{(3\times 1)} = C_{(2\times 1)} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$ với

$c_1 = 3.(-2) + (-1).7 + 2.9 = 5; c_2 = 1.(-2) + 5.7 + 4.9 = 69.$

Vậy $A \times B = \begin{bmatrix} 5 \\ 69 \end{bmatrix}$.

Ma trận đối xứng

Ma trận A cấp $n \times n$ gọi là ma trận đối xứng nếu $A = (a_{ij})$ trong đó $a_{ij} = a_{ji}$ tức là $A = A'$.

Ma trận nghịch đảo

Cho ma trận vuông A cấp $n \times n$. Ma trận vuông cấp n A^{-1} gọi là ma trận nghịch đảo của A nếu nó tồn tại là ma trận sao cho

$$AA^{-1} = A^{-1}A = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix} := I_n \qquad (1)$$

trong đó I_n là ma trận vuông với các phần tử trên đường chéo chính bằng 1 , các phần tử còn lại đều bằng 0 gọi là ma trận đơn vị cấp n.

Với các ma trận đường chéo ta ký hiệu $A = diag(a_{11}, \cdots, a_{nn})$.

Ma trận trực giao

Ma trận vuông A cấp $n \times n$ gọi là ma trận trực giao nếu $A' = A^{-1}$
Hoặc

$$a_{i1}^2 + a_{i2}^2 + \cdots + a_{in}^2 = 1 \quad \forall i = 1 \div n \tag{2}$$

$$a_{i1}a_{k1} + \cdots + a_{in}a_{kn} = 0 \quad \forall k \neq i = 1 \div n \tag{3}$$

Giá trị riêng và vector riêng của ma trận vuông

Cho A là ma trận vuông cấp n . Khi đó số thực hoặc phức λ và vector tương ứng với nó là x được gọi là giá trị riêng và vector riêng của A nếu:

$$Ax = \lambda x, \tag{4}$$

trong đó các giá trị riêng $\lambda_1, \ldots, \lambda_n$ là nghiệm của phương trình:

$$|A - \lambda I_n| = 0.$$

Thông thường chọn vector riêng x sao cho $|x| = 1$.

Giá trị riêng của ma trận đối xứng xác định dương.

Ma trận A cấp $n \times n$ được gọi là ma trận đối xứng xác định không âm nếu $x'Ax \geq 0$ với mọi $x' = (x_1, \ldots, x_n)$. Nếu hơn nữa $x'Ax = 0$ khi và chỉ khi $x = 0 = (0, 0, \ldots, 0)$ thì A gọi là ma trận xác định dương.
Ký hiệu $A \geq 0$ nếu A là ma trận xác định không âm, ký hiệu $A > 0$ nếu A là ma trận xác định dương. Ngược lại, nếu $-A \geq 0$ (hoặc $-A > 0$ thì A được gọi là ma trận xác định không dương (hoặc ma trận xác định âm).

Khai triển phổ của ma trận đối xứng A

Nếu $\lambda_1, e_1; \ldots; \lambda_n, e_n$ là n cặp giá trị riêng và vector riêng của ma trận A thì A có thể viết dưới dạng:

$$A = \lambda_1 e_1 e_1' + \cdots + \lambda_n e_n e_n' \tag{5}$$

Bất đẳng thức Cauchy-Schwatz

Cho hai vecto b và d trong $\mathbb{R}^n$. Ta có

$$(b'd)^2 \leq (b'b)(d'd) \tag{6}$$

và dấu "=" xảy ra khi và chỉ khi $b = cd$ hoặc $d = cb$ với c là hằng số nào đấy.

Bất đẳng thức Cauchy-Schwatz mở rộng

Cho hai vecto b và d trong $\mathbb{R}^n$ và B là ma trận xác định dương cấp $n \times n$. Khi đó:

$$(b'd)^2 \leq (b'Bb)(d'B^{-1}d) \tag{7}$$

dấu "=" xảy ra khi và chỉ khi hoặc $b = cB^{-1}d$ hoặc $d = cBb$.

Bổ đề về maximum

Giả sử $B > 0$ cấp $n \times n, d \in \mathbb{R}^n$. Khi đó $\forall x \in \mathbb{R}^n$ ta có

$$\max_{x \neq 0} \frac{(x'd)^2}{x'Bx} = d'B^{-1}d. \tag{8}$$

với giá trị max đạt được khi $x = cB^{-1}d$ với bất kỳ hằng số $c \neq 0$.

Maximum của dạng thức toàn phương trên hình cầu đơn vị

Giả sử $B > 0$ cấp $n \times n$ với các giá trị riêng $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ và $e_1, e_2, \ldots, e_n$ là các vector riêng tương ứng đã chuẩn hoá. Khi đó:

$$\begin{aligned} \max_{x \neq 0} \frac{x'Bx}{x'x} &= \lambda_1 \quad (\text{đạt được khi } x = e_1) \\ \min_{x \neq 0} \frac{x'Bx}{x'x} &= \lambda_n \quad (\text{đạt được khi } x = e_n) \end{aligned} \tag{9}$$

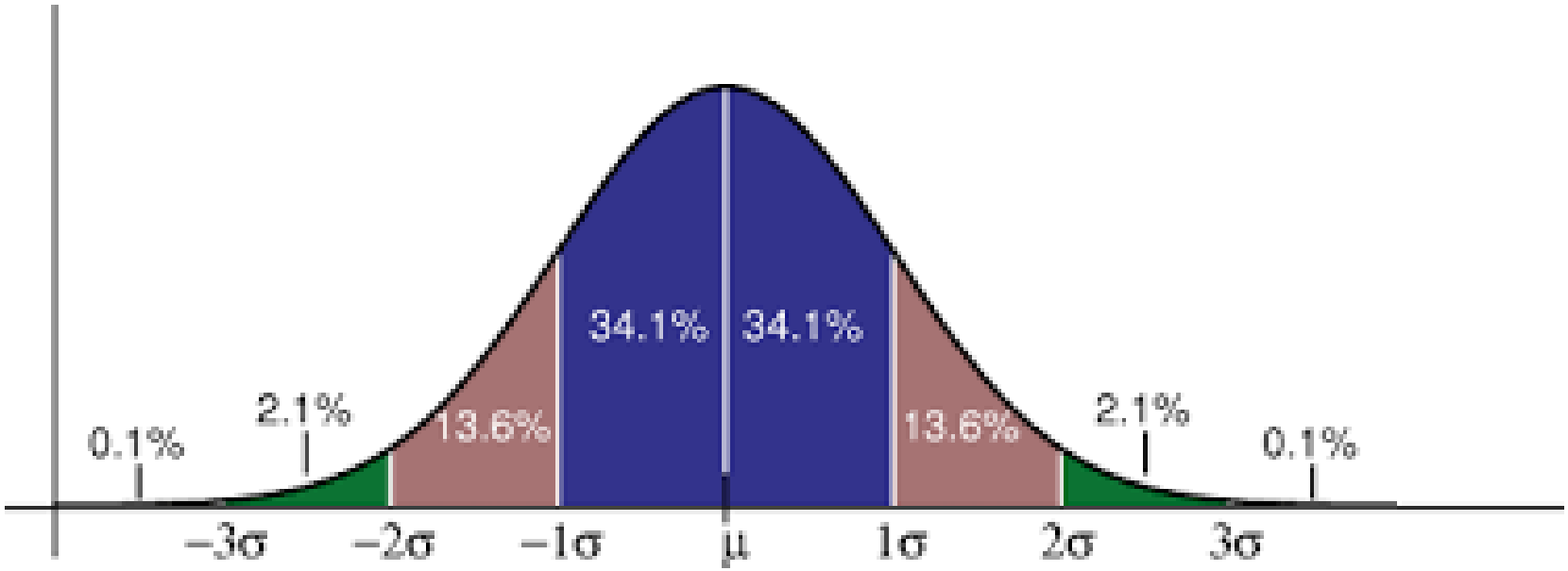
hơn nữa

$$\max_{x \perp e_1, \ldots, e_k} \frac{x'Bx}{x'x} = \lambda_{k+1} \quad (\text{đạt được khi } x = e_{k+1}),\ k = 1, 2, \ldots, n-1. \tag{10}$$

trong đó $x \perp y$ nếu $x'y = (x, y) = 0$.

Mật độ của phân bố chuẩn một chiều

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}, -\infty < x < +\infty \qquad (11)$$

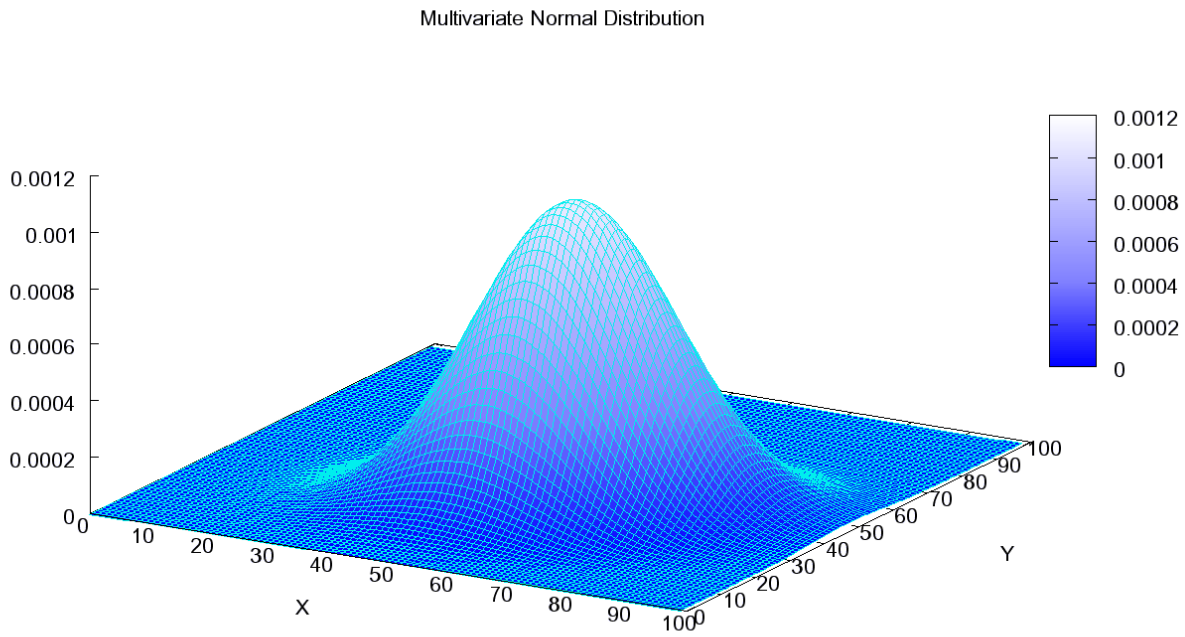


Hàm mật độ phân bố chuẩn hai chiều

$$f_{XY}(x, y, \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$
$$\times exp\left\{ \frac{-1}{2(1-\rho^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(x-\mu_Y)^2}{\sigma_Y^2} \right] \right\} \quad (12)$$

với $-\infty < x < \infty, -\infty < y < \infty$, tham số
$\sigma_X > 0, \sigma_Y > 0, -\infty < \mu_X < \infty, -\infty < \mu_X < \infty$ và $-1 < \rho < 1$.

Phân bố chuẩn hai chiều



Vecto ngẫu nhiên

Vecto ngẫu nhiên n chiều $X = [X_1, \cdots, X_n]$ nhận giá trị trong $\mathbb{R}^n$ là một vecto mà mỗi thành phần $X_1, \cdots, X_n$ là một biến ngẫu nhiên nhận giá trị trong $\mathbb{R}$.

Phân bố xác suất của vecto ngẫu nhiên n chiều X được xác định bởi

$$F(x_1, \cdots, x_n) = P(X_1 < x_1, \cdots, X_n < x_n). \quad (13)$$

Hàm không âm $f(x_1, \cdots, x_n)$ gọi là hàm mật độ của vecto ngẫu nhiên X nếu :

$$F(x_1, \cdots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(y_1, \cdots, y_n) dy_1 \cdots dy_n \quad (14)$$

Cụ thể : Với n=2, ta gặp trường hợp quen thuộc $X = [X_1, X_2]$ với hàm phân bố xác suất đã gặp $F(x_1, x_2) = P(X_1 < x_1, X_2 < x_2)$.

Mật độ của phân bố chuẩn nhiều chiều

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\} \qquad (15)$$

trong đó $\mu = (\mu_1, \ldots, \mu_n)' \in \mathbb{R}^n$; Σ là ma trận đối xứng xác định dương là các tham số của phân bố chuẩn n chiều, $x \in \mathbb{R}^n$. Kí hiệu $N(\mu, \Sigma)$

Dữ liệu mô phỏng phân phối chuẩn nhiều chiều

```
> library(mvtnorm)
> rmvnorm(5, mean = c(0, 0), sigma = matrix(c(1,.8, .8, 1), 2, 2))
           [,1]       [,2]
[1,]  1.03550508  0.06044561
[2,]  0.53386104  1.03063539
[3,] -0.06674766 -0.41792785
[4,] -0.59569721 -0.54805093
[5,]  0.96581969  0.61702999
```

Tính chất phân bố chuẩn nhiều chiều

Nếu X có phân bố chuẩn $N(\mu, \Sigma)$ thì:

1. $EX = \mu,\ var(X) = \Sigma.$

Tính chất phân bố chuẩn nhiều chiều

Nếu X có phân bố chuẩn $N(\mu, \Sigma)$ thì:

1. $EX = \mu,\ var(X) = \Sigma$.
2. Mọi tổ hợp tuyến tính $c'X = c_1X_1 + \cdots + c_nX_n$ có phân bố chuẩn $N(c'\mu, c'\sum c)$. (c là vector)

Tính chất phân bố chuẩn nhiều chiều

Nếu X có phân bố chuẩn $N(\mu, \Sigma)$ thì:

1. $EX = \mu,\ var(X) = \Sigma$.
2. Mọi tổ hợp tuyến tính $c'X = c_1X_1 + \cdots + c_nX_n$ có phân bố chuẩn $N(c'\mu, c'\sum c)$. (c là vector)
3. Mọi tập con các thành phần của X cũng có phân bố chuẩn.

Tính chất phân bố chuẩn nhiều chiều

Nếu X có phân bố chuẩn $N(\mu, \Sigma)$ thì:

1. $EX = \mu,\ var(X) = \Sigma$.
2. Mọi tổ hợp tuyến tính $c'X = c_1X_1 + \cdots + c_nX_n$ có phân bố chuẩn $N(c'\mu, c'\sum c)$. (c là vector)
3. Mọi tập con các thành phần của X cũng có phân bố chuẩn.
4. Các phân bố có điều kiện của một số các thành phần X khi biết trước các thành phần khác cũng là phân bố chuẩn.

Tính chất phân bố chuẩn nhiều chiều

Nếu X có phân bố chuẩn $N(\mu, \Sigma)$ thì:

1. $EX = \mu,\ var(X) = \Sigma$.
2. Mọi tổ hợp tuyến tính $c'X = c_1X_1 + \cdots + c_nX_n$ có phân bố chuẩn $N(c'\mu, c'\sum c)$. (c là vector)
3. Mọi tập con các thành phần của X cũng có phân bố chuẩn.
4. Các phân bố có điều kiện của một số các thành phần X khi biết trước các thành phần khác cũng là phân bố chuẩn.
5. Nếu Σ có dạng $\Sigma = diag(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{nn})$ thì các thành phần $X_1, X_2, \ldots, X_n$ là độc lập.

Tính chất phân bố chuẩn nhiều chiều

Nếu X có phân bố chuẩn $N(\mu, \Sigma)$ thì:

1. $EX = \mu,\ var(X) = \Sigma$.
2. Mọi tổ hợp tuyến tính $c'X = c_1X_1 + \cdots + c_nX_n$ có phân bố chuẩn $N(c'\mu, c'\sum c)$. (c là vector)
3. Mọi tập con các thành phần của X cũng có phân bố chuẩn.
4. Các phân bố có điều kiện của một số các thành phần X khi biết trước các thành phần khác cũng là phân bố chuẩn.
5. Nếu Σ có dạng $\Sigma = diag(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{nn})$ thì các thành phần $X_1, X_2, \ldots, X_n$ là độc lập.
6. $Y = \Sigma^{-1/2}(X - \mu)$ có phân bố chuẩn $N(0, I_n)$;
   $\sum_1^n Y_i^2 = Y'Y = [X - \mu]'\Sigma^{-1}[X - \mu]$ có phân bố χ^2 với n bậc tự do.

Đặc trưng của Vecto ngẫu nhiên

- Trung bình:

$$EX = (EX_1, \cdots, EX_n)^T = (\mu_1, \cdots, \mu_n)^T$$

được gọi là vecto giá trị trung bình của X.

- Ma trận hiệp phương sai:

$$\Sigma = cov(X) = E(X - \mu)(X - \mu)^T = \left[\sigma_{ij}\right]_{n \times n}$$

gọi là ma trận hiệp phương sai của vecto X, trong đó

$$\sigma_{ii} = E(X_i - \mu_i)^2;\ \sigma_{ij} = E(X_i - \mu_i)(X_j - \mu_j)$$

lần lượt là phương sai của X_i với $i = 1, \cdots, n$ và hiệp phương sai của X_i và X_j với $i, j = 1, \cdots, n$

Đặc trưng của Vecto ngẫu nhiên

- Ma trận tương quan:

$$\rho = [\rho_{ij}]$$

gọi là ma trận tương quan của X, trong đó

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

là hệ số tương quan của X_i và X_j.

- Độc lập và không tương quan:
  - Nếu $\sigma_{ij} = 0 \ \forall \ i \neq j$ thì X_i và X_j gọi là không tương quan.
  - X_i và X_j gọi là độc lập nếu

$$P(X_i < x_i, X_j < x_j) = P(X_i < x_i)P(X_j < x_j) \ \forall \ x_i, x_j \in \mathbb{R}^1$$

  - Nếu X_i và X_j độc lập thì $\sigma_{ij} = 0$, điều ngược lại nói chung không đúng.

Đặc trưng của Vecto ngẫu nhiên

- Tổ hợp tuyến tính $C'X = c_1X_1 + c_2X_2 + \cdots + c_nX_n$ có

$$E(C'X) = C'E(X), \ var(C'X) = C'\Sigma C \tag{16}$$

Đặc trưng của Vecto ngẫu nhiên

- Tổ hợp tuyến tính $C'X = c_1X_1 + c_2X_2 + \cdots + c_nX_n$ có

$$E(C'X) = C'E(X),\ var(C'X) = C'\Sigma C \tag{16}$$

- Với q tổ hợp tuyến tính

$$z_1 = c_{11}X_1 + \cdots + c_{1n}X_n$$
$$\vdots$$
$$z_q = c_{q1}X_1 + \cdots + c_{qn}X_n$$

Đặc trưng của Vecto ngẫu nhiên

- Tổ hợp tuyến tính $C'X = c_1X_1 + c_2X_2 + \cdots + c_nX_n$ có

$$E(C'X) = C'E(X),\ var(C'X) = C'\Sigma C \tag{16}$$

- Với q tổ hợp tuyến tính

$$z_1 = c_{11}X_1 + \cdots + c_{1n}X_n$$
$$\vdots$$
$$z_q = c_{q1}X_1 + \cdots + c_{qn}X_n$$

ta cũng nhận được phương trình (16) với

$$Z = \begin{bmatrix} z_1 \\ \vdots \\ z_q \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{21} & c_{q2} & \cdots & c_{qn} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = C'X$$

Ví dụ 1

Trong trường hợp $n = 2$ chiều, vecto $X = \begin{bmatrix} X_1 \\ X2 \end{bmatrix}$ có phân bố chuẩn. Biến ngẫu nhiên X_1 độc lập với X_2 và cùng có phân bố chuẩn với trung bình 0 và phương sai 1. Hãy xác định vecto trung bình, ma trận hiệp phương sai và ma trận tương quan của X.

Ví dụ 1

Trong trường hợp $n = 2$ chiều, vecto $X = \begin{bmatrix} X_1 \\ X2 \end{bmatrix}$ có phân bố chuẩn. Biến ngẫu nhiên X_1 độc lập với X_2 và cùng có phân bố chuẩn với trung bình 0 và phương sai 1. Hãy xác định vecto trung bình, ma trận hiệp phương sai và ma trận tương quan của X.

Ví dụ 2

Trong trường hợp $n = 2$ chiều, vecto $X = \begin{bmatrix} X_1 \\ X2 \end{bmatrix}$ có phân bố chuẩn với hàm mật độ xác suất

$$f(x) = f(x_1, x_2) = \frac{1}{\sqrt{3}\pi} e^{-\frac{2}{3}[x_1^2 - x_1 x_2 + x_2^2]}$$

Hãy xác định vecto trung bình, ma trận hiệp phương sai và ma trận tương quan của X.

Ví dụ 3

Cho vecto $c = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, hãy tính trung bình và phương sai của $c'X$.

Vecto $d = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$, hãy tính trung bình và ma trận hiệp phương sai của vecto $Y = \begin{pmatrix} c'X \\ d'X \end{pmatrix}$.

Mẫu ngẫu nhiên nhiều chiều

Cho vecto ngẫu nhiên n chiều $X = [X_1, \cdots, X_n]^T$. Thực hiện N quan sát độc lập về X. Giả sử lần quan sát thứ nhất thu được $X^1 = (x_{11}, x_{12}, \cdots, x_{1n})^T$; quan sát lần 2 thu được $X^2 = (x_{21}, x_{22}, \cdots, x_{2n})^T$ .... và quan sát lần thứ N được $X^N = (x_{N1}, x_{N2}, \cdots, x_{Nn})^T$. Khi đó

$$\begin{bmatrix} X^1 \\ X^2 \\ \vdots \\ X^N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{bmatrix}$$

là mẫu cỡ N quan sát vecto n chiều X.

Dữ liệu về Giá trị dinh dưỡng của các thương hiệu kẹo nổi tiếng (nguồn: Table 7.4, Daniel Zelterman (2016)).

Name	Calories	Fat	Satfat	Carbs	Sugar	Sodium
100 Grand	190	8	5	30	22	90
3 Musketeers	240	7	5	42	36	90
5th Avenue	260	12	5	38	29	120
Almond Joy	220	13	8	26	20	50
Andes Mints	200	13	11	22	20	20
Baby Ruth	275	13	7	39	32	138
Butterfinger	270	11	6	43	29	135
Cadbury Dairy Milk	260	15	9	28	28	0
Charleston Chew	230	6	5	43	30	30
Dove Smooth Milk Choc.	220	13	8	24	22	25
Goobers	200	13	5	20	17	15
Heath Toffee	210	13	7	24	23	135
Hershey's bar	210	13	8	26	24	35
Hershey's Skor	200	12	7	25	24	130
Junior Mints	220	4	3	45	42	35
Kit Kat	207	10	7	26	20	22
M&M's, peanut	250	13	5	30	25	25
M&M's, plain	230	9	6	34	31	35
Milk Duds	230	8	5	38	27	135
Milky Way	240	9	7	37	31	75
Mounds	240	13	10	29	21	55
Mr Goodbar	250	16	7	25	22	50
Nestle Crunch	220	11	7	30	24	60
Oh Henry!	280	17	7	36	32	65
Payday	240	13	3	27	21	120
Raisinets	190	8	5	32	28	15
Reese's Fast Break	260	12	5	35	30	190
Reese's Nutrageous	260	16	5	28	22	100
Reese's Peanut Butter cups	210	13	5	24	21	150
Reese's Pieces	200	9	8	25	21	55
Reese's Sticks	220	13	5	23	17	130
Rolo	220	10	7	33	29	80
Snickers	230	11	4	32	27	115
Symphony	223	13	8	24	23	42
Twix	250	12	7	33	24	100
Whatchamacalit	237	11	8	30	23	144
Whoppers	190	7	7	31	24	100
Zero Candy Bar	200	7	5	34	29	105

Values collected from `www.fatsecret.com`. Used with permission

Dữ liệu mô phỏng phân phối chuẩn nhiều chiều

```
> library(mvtnorm)
> rmvnorm(5, mean = c(0, 0), sigma = matrix(c(1,.8, .8, 1), 2, 2))
           [,1]       [,2]
[1,]  1.03550508  0.06044561
[2,]  0.53386104  1.03063539
[3,] -0.06674766 -0.41792785
[4,] -0.59569721 -0.54805093
[5,]  0.96581969  0.61702999
```

Đặc trưng của Mẫu ngẫu nhiên

- Vecto trung bình mẫu:

$$\overline{X'} = [\overline{X_1}, \ldots, \overline{X_n}],$$

trong đó $\overline{X_1} = \frac{1}{N}\sum_{j=1}^{N} x_{j1}, \ldots, \overline{X_n} = \frac{1}{N}\sum_{j=1}^{N} x_{jn},.$

Đặc trưng của Mẫu ngẫu nhiên

- Vecto trung bình mẫu:

$$\overline{X'} = [\overline{X_1}, \ldots, \overline{X_n}],$$

trong đó $\overline{X_1} = \frac{1}{N}\sum_{j=1}^{N} x_{j1}, \ldots, \overline{X_n} = \frac{1}{N}\sum_{j=1}^{N} x_{jn},$.

- Ma trận hiệp phương sai mẫu:

$$S_n = \begin{bmatrix} s_{11} & \ldots & s_{1n} \\ \ldots & \ldots & \ldots \\ s_{n1} & \ldots & s_{nn} \end{bmatrix},$$

trong đó

$$s_{ij} = \frac{1}{N}\sum_{k=1}^{N}(x_{ki} - \overline{X_i})(x_{kj} - \overline{X_j}) = \frac{1}{N}\sum_{k=1}^{N} x_{ki}x_{kj} - \bar{X}_i \times \bar{X}_j \quad (17)$$

là hiệp phương sai mẫu.

Đặc trưng của Mẫu ngẫu nhiên

- Ma trận tương quan mẫu:

$$R = [r_{ij}]$$

với $r_{ij} = \dfrac{s_{ij}}{(s_{ii}s_{jj})^{\frac{1}{2}}}$ là ma trận hệ số tương quan mẫu.

Đặc trưng của Mẫu ngẫu nhiên

- Ma trận tương quan mẫu:

$$R = [r_{ij}]$$

với $r_{ij} = \dfrac{s_{ij}}{(s_{ii}s_{jj})^{\frac{1}{2}}}$ là ma trận hệ số tương quan mẫu.

Tính không chệch của đặc trưng mẫu

Ước lượng không chệch cho vecto trung bình và ma trận hiệp phương sai:

- $\overline{X}$ là ước lượng không chệch (ULKC) của $\mu = E(X)$;

Đặc trưng của Mẫu ngẫu nhiên

- Ma trận tương quan mẫu:

$$R = [r_{ij}]$$

với $r_{ij} = \dfrac{s_{ij}}{(s_{ii}s_{jj})^{\frac{1}{2}}}$ là ma trận hệ số tương quan mẫu.

Tính không chệch của đặc trưng mẫu

Ước lượng không chệch cho vecto trung bình và ma trận hiệp phương sai:

- $\overline{X}$ là ước lượng không chệch (ULKC) của $\mu = E(X)$;
- $S_n^* = \dfrac{nS_n}{n-1}$ là ULKC của Σ.

Một số biểu đồ mô tả dữ liệu nhiều chiều

Dữ liệu về nhiệt độ cao nhất trong tháng giêng ở các thành phố lớn nhất của Mỹ với toạ độ (nguồn: Table 3.1, Daniel Zelterman (2016)).

Table 3.1: Maximum January temperature (T, in degrees Fahrenheit), latitude, longitude, and altitude in feet above sea level, for some of the largest US cities

T	Lat	Long	Alt	Name	T	Lat	Long	Alt	Name
61	30	88	5	Mobile, AL	59	32	86	160	Montgomery, AL
30	58	134	50	Juneau, AK	64	33	112	1090	Phoenix, AZ
51	34	92	286	Little Rock, AR	65	34	118	340	Los Angeles, CA
55	37	122	65	San Francisco	42	39	104	5280	Denver, CO
37	41	72	40	New Haven, CT	41	39	75	135	Wilmington, DE
44	38	77	25	Washington, DC	67	30	81	20	Jacksonville, FL
74	24	81	5	Key West, FL	76	25	80	10	Miami, FL
52	33	84	1050	Atlanta, GA	79	21	157	21	Honolulu, HI
36	43	116	2704	Boise, ID	33	41	87	595	Chicago, IL
37	39	86	710	Indianapolis, IN	29	41	93	805	Des Moines, IA
27	42	90	620	Dubuque, IA	42	37	97	1290	Wichita, KS
44	38	85	450	Louisville, KY	64	29	90	5	New Orleans, LA
32	43	70	25	Portland, ME	44	39	76	20	Baltimore, MD
37	42	71	21	Boston, MA	33	42	83	585	Detroit, MI
23	46	84	650	Sault Ste Marie	22	44	93	815	Minneapolis, MN
40	38	90	455	St Louis, MO	29	46	112	4155	Helena, MT
32	41	95	1040	Omaha, NE	32	43	71	290	Concord, NH
43	39	74	10	Atlantic City, NJ	46	35	106	4945	Albuquerque, NM
31	42	73	20	Albany, NY	40	40	73	55	New York, NY
51	35	80	720	Charlotte, NC	52	35	78	365	Raleigh, NC
20	46	100	1674	Bismark, ND	41	39	84	550	Cincinnati, OH
35	41	81	660	Cleveland, OH	46	35	97	1195	Oklahoma City
44	45	122	77	Portland, OR	39	40	76	365	Harrisburg, PA
40	39	75	100	Philadelphia, PA	61	32	79	9	Charleston, SC
34	44	103	3230	Rapid City, SD	49	36	86	450	Nashville, TN
50	35	101	3685	Amarillo, TX	61	29	94	5	Galveston, TX
37	40	111	4390	Salt Lake City	25	44	73	110	Burlington, VT
50	36	76	10	Norfolk, VA	44	47	122	10	Seattle, WA
31	47	117	1890	Spokane, WA	26	43	89	860	Madison, WI

Một số biểu đồ mô tả dữ liệu nhiều chiều

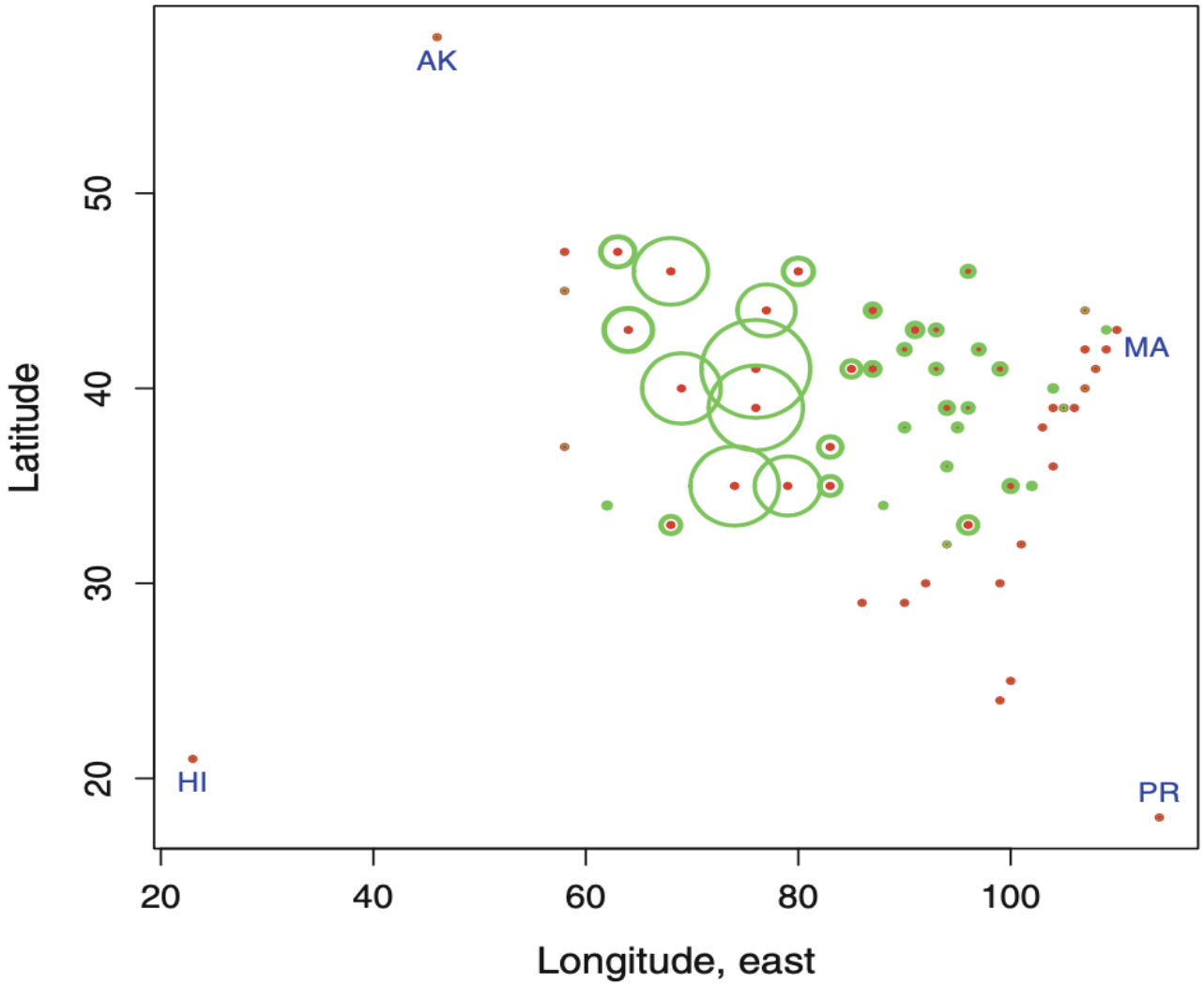


Figure 3.13: Bubble plot of altitudes of US cities

Một số biểu đồ mô tả dữ liệu nhiều chiều

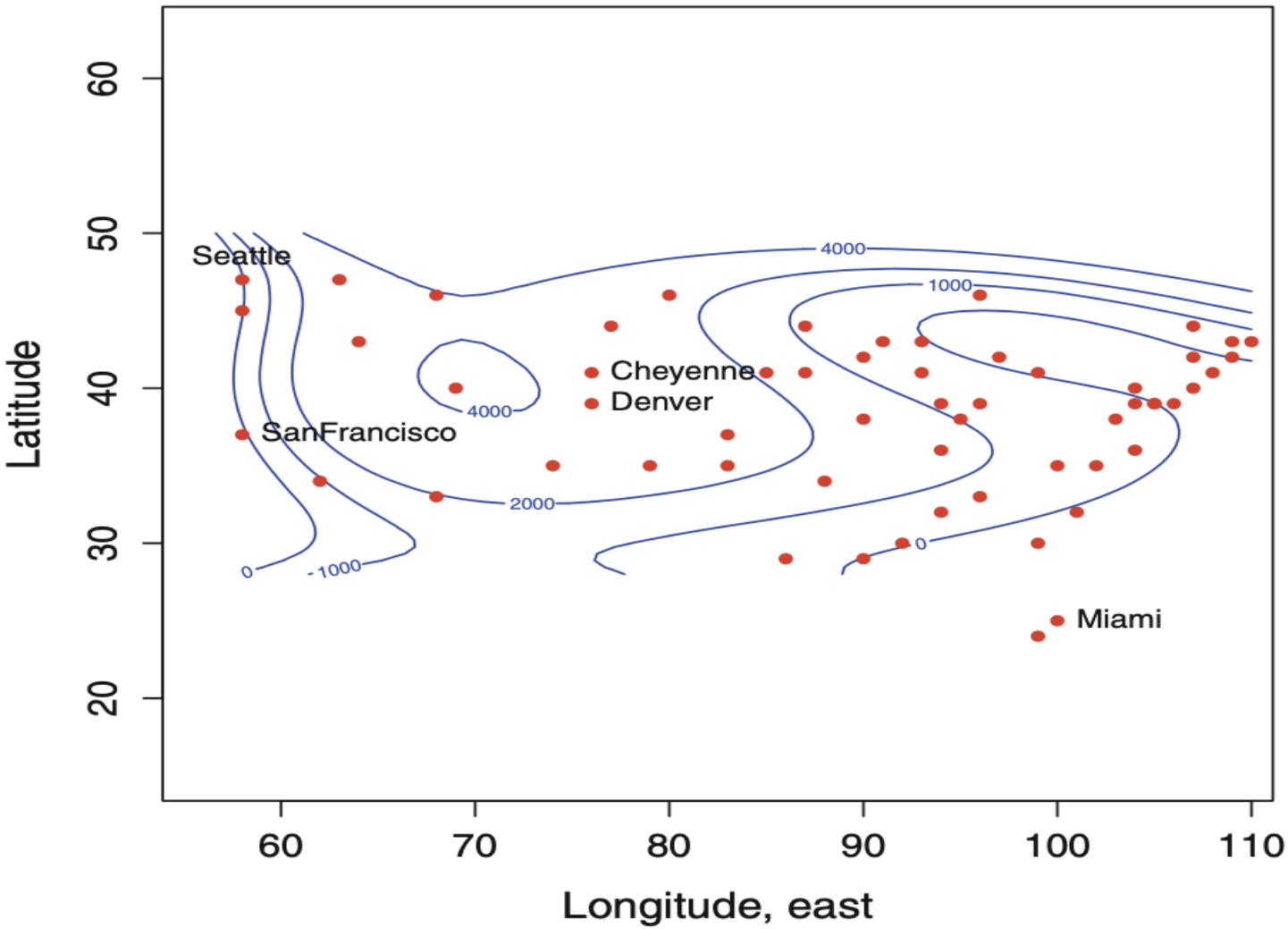
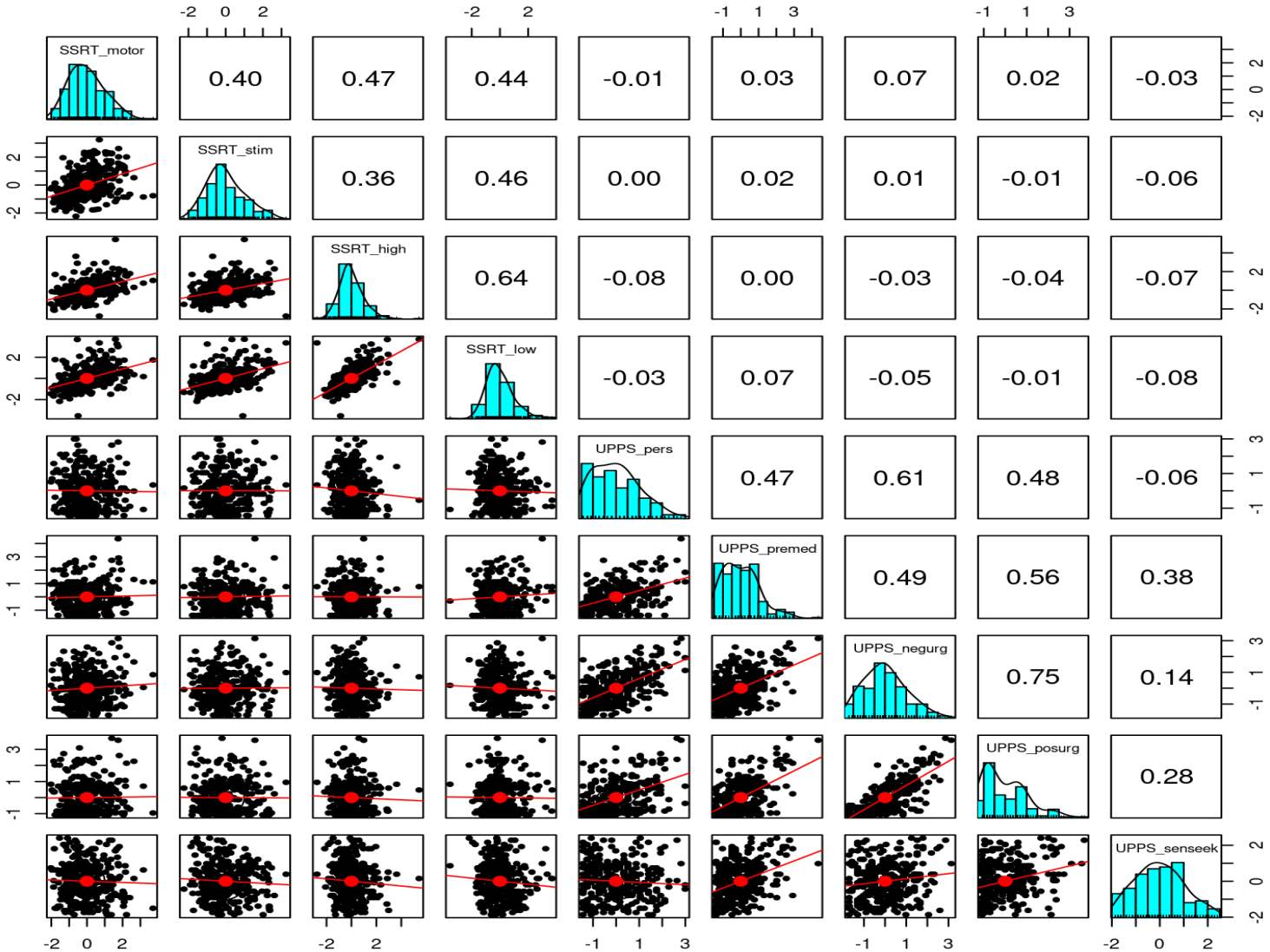


Figure 3.14: Kriging plot of altitudes of US cities

Một số biểu đồ mô tả dữ liệu nhiều chiều



Một số biểu đồ mô tả dữ liệu nhiều chiều

```
library(psych)
pairs.panels(attitude)    #see the graphics window
data(iris)
pairs.panels(iris[1:4],bg=c("red","yellow","blue")[
          pch=21,main="Fisher Iris data by Species")

pairs.panels(iris[1:4],bg=c("red","yellow","blue")[
   pch=21+as.numeric(iris$Species),main="Fisher Iris
    #to show changing the diagonal

#to show 'significance'
    pairs.panels(iris[1:4],bg=c("red","yellow","blue
   pch=21+as.numeric(iris$Species),main="Fisher Iris
```

Một số biểu đồ mô tả dữ liệu nhiều chiều

```
#demonstrate not showing the data points
data(sat.act)
pairs.panels(sat.act,show.points=FALSE)
#better yet is to show the points as a period
pairs.panels(sat.act,pch=".")
#show many variables with 0 gap between scatterplot
# data(bfi)
# pairs.panels(psychTools::bfi,show.points=FALSE,ga

#plot raw data points and then the weighted correla
#output from statsBy
sb <- statsBy(sat.act,"education")
pairs.panels(sb$mean,wt=sb$n) #report the weighted
#compare with
pairs.panels(sb$mean) #unweighted correlations
```

Phân bố mẫu ngẫu nhiên

Cho $\left[X^1, X^2, \cdots, X^N\right]$ là mẫu ngẫu nhiên quan sát $X = (X_1, \ldots, X_n)'$ có phân bố chuẩn $N(\mu, \Sigma)$. Khi đó $X^1, X^2, \cdots, X^N$ là N vector độc lập có phân bố chuẩn $N(\mu, \Sigma)$

Phân bố mẫu ngẫu nhiên

Cho $\left[X^1, X^2, \cdots, X^N\right]$ là mẫu ngẫu nhiên quan sát $X = (X_1, \ldots, X_n)'$ có phân bố chuẩn $N(\mu, \Sigma)$. Khi đó $X^1, X^2, \cdots, X^N$ là N vector độc lập có phân bố chuẩn $N(\mu, \Sigma)$ và hàm mật độ đồng thời của $X^1, X^2, \cdots, X^N$ là

$$\prod_{j=1}^{N} \frac{1}{2\pi^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x^j - \mu)'\Sigma^{-1}(x^j - \mu)\right\} =$$

$$= (2\pi)^{-Nn/2}|\Sigma|^{-N/2} \exp\left\{-\frac{1}{2}\sum_{j=1}^{N}(x^j - \mu)'\Sigma^{-1}(x^j - \mu)\right\} \quad (18)$$

Ước lượng hợp lý cực đại

Ước lượng $\widehat{\mu},\ \widehat{\Sigma} = [\widehat{\sigma}_{ij}]$ cấp $n \times n$ được gọi là ước lượng hợp lý cực đại của vector trung bình μ và ma trận hiệp phương sai Σ của phân bố chuẩn nếu

$$L(\widehat{\mu}, \widehat{\Sigma}|X^1, \ldots, X^n) = \max_{\mu,\Sigma} L(\mu, \Sigma|X^1, \ldots, X^n), \qquad (19)$$

trong đó hàm $L(\mu, \Sigma|X^1, \ldots, X^n)$ được cho bởi (18) khi coi $X^1, \ldots, X^n$ là cố định được gọi là hàm hợp lý.

Tính chất của $\bar{X}$ và S

Với mẫu ngẫu nhiên quan sát từ phân bố chuẩn nhiều chiều,

- $(\bar{X}, S)$ là ước lượng hợp lý cực đại của (μ, Σ)

Tính chất của $\bar{X}$ và S

Với mẫu ngẫu nhiên quan sát từ phân bố chuẩn nhiều chiều,

- $(\bar{X}, S)$ là ước lượng hợp lý cực đại của (μ, Σ)
- $\bar{X}$ có phân bố chuẩn $N(\mu, \frac{1}{n}\Sigma)$

Tính chất của $\bar{X}$ và S

Với mẫu ngẫu nhiên quan sát từ phân bố chuẩn nhiều chiều,

- $(\bar{X}, S)$ là ước lượng hợp lý cực đại của (μ, Σ)
- $\bar{X}$ có phân bố chuẩn $N(\mu, \frac{1}{n}\Sigma)$
- nS có phân bố Wishart với $n-1$ bậc tự do. (*xem* `https://en.wikipedia.org/wiki/Wishart_distribution`)

Tính chất của $\bar{X}$ và S

Với mẫu ngẫu nhiên quan sát từ phân bố chuẩn nhiều chiều,

- $(\bar{X}, S)$ là ước lượng hợp lý cực đại của (μ, Σ)
- $\bar{X}$ có phân bố chuẩn $N(\mu, \frac{1}{n}\Sigma)$
- nS có phân bố Wishart với $n-1$ bậc tự do. (*xem* `https://en.wikipedia.org/wiki/Wishart_distribution`)
- $\bar{X}$ và S là độc lập.

Tính chất của $\bar{X}$ và S

Với mẫu ngẫu nhiên quan sát từ phân bố chuẩn nhiều chiều,

- $(\bar{X}, S)$ là ước lượng hợp lý cực đại của (μ, Σ)
- $\bar{X}$ có phân bố chuẩn $N(\mu, \frac{1}{n}\Sigma)$
- nS có phân bố Wishart với $n-1$ bậc tự do. (*xem* `https://en.wikipedia.org/wiki/Wishart_distribution`)
- $\bar{X}$ và S là độc lập.
- Khi N đủ lớn, $\sqrt{N}(\bar{X} - \mu)$ có phân bố gần chuẩn $N(0, \Sigma)$.

Nhận dạng phân phối chuẩn nhiều chiều

Cho vecto ngẫu nhiên m chiều $X = [X_1, \cdots, X_m]^T$ và mẫu ngẫu nhiên quan sát X. Khi đó việc nhận dạng X có phân bố chuẩn hay không được thực hiện qua hai bước sau:

- Bước 1: Xem từng thành phần $X_i, i = 1, \cdots, m$ có phân bố chuẩn hay không?
- Bước 2: Kiểm tra tính phân bố chuẩn đồng thời của X ?

Bước 2 được thực hiện thông qua tính toán khoảng cách Mahalanobis.

- Xét m giá trị $1 > \alpha_1 > \cdots > \alpha_k > 0$.
- Ký hiệu $V(\alpha) = \left\{x : (x - \overline{X})^T S^{-1}(x - \overline{X}) \leq \chi^2_m(\alpha)\right\}$ là một ellipsoid.
- $p_1 = 1 - \alpha_1; p_2 = \alpha_1 - \alpha_2; \cdots; p_k = \alpha_{k-1} - \alpha_k; p_{k+1} = \alpha_k$.
- $d^2(x) = (x - \overline{X})^T S^{-1}(x - \overline{X})$ là khoảng cách Mahalanobis.
- $X_1, \cdots, X_n$ là n quan sát vecto X. Tính giá trị các $d^2(X_i)$ và

$$\eta_1 = \#\left\{X_i : d^2(X_i) \leq \chi^2_m(\alpha_1)\right\}$$
$$\eta_2 = \#\left\{X_i : \chi^2_m(\alpha_1) < d^2(X_i) \leq \chi^2_m(\alpha_2)\right\}$$
$$\cdots \quad \cdots$$
$$\eta_k = \#\left\{X_i : \chi^2_m(\alpha_{k-1}) < d^2(X_i) \leq \chi^2_m(\alpha_k)\right\}$$
$$\eta_{k+1} = \#\left\{X_i : \chi^2_m(\alpha_k) < d^2(X_i)\right\}$$

Yêu cầu : $\eta_i \geq 5$. Do đó việc tìm dãy $\alpha_1; \cdots; \alpha_k$ dựa trên việc phân chia lực lượng của tập các $d^2(X_i)$.

- Tính tổng Khi bình phương :

$$\chi^2 = \sum_{i=1}^{k+1} \frac{\eta_i^2}{np_i} - n.$$

- Miền tiêu chuẩn : $S = \{\chi^2 \geq \chi_k^2(\alpha)\}$

Trong phần mềm R ta có thể tính toán đơn giản hơn bằng tính phân vị chuẩn từ phân phối khi bình phương của các khoảng cách Mahalanobis. Cụ thể R code:
mah= mahalanobis(candy, colMeans(candy), var(candy))
shapiro.test(qnorm(pchisq(mah, 6)))

So sánh hai vecto trung bình

Cho $X^{(1)}$ và $X^{(2)}$ có phân phối chuẩn $N(\mu^{(1)}, \Sigma), N(\mu^{(2)}, \Sigma)$. Xét bài toán kiểm định giả thiết sau:

Giả thiết $H_0 : \mu^{(1)} = \mu^{(2)}$; Đối thiết $\mu^{(1)} \neq \mu^{(2)}$.

So sánh hai vecto trung bình

Cho $X^{(1)}$ và $X^{(2)}$ có phân phối chuẩn $N(\mu^{(1)}, \Sigma), N(\mu^{(2)}, \Sigma)$. Xét bài toán kiểm định giả thiết sau:

$$\text{Giả thiết } H_0 : \mu^{(1)} = \mu^{(2)}; \qquad \text{Đối thiết } \mu^{(1)} \neq \mu^{(2)}.$$

Với hai mẫu thu thập độc lập, H_0 bị bác bỏ nếu thống kê

$$T_0^2 = (\overline{X^{(1)}} - \overline{X^{(2)}})' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S \right]^{-1} (\overline{X^{(1)}} - \overline{X^{(2)}}) > \frac{(n_1 + n_2 - 2)n}{n_1 + n_2 - n - 1} F_{n, n_1+n_2-n-1}(\alpha),$$

trong đó $\overline{X^{(1)}} = \frac{1}{n_1} \sum_{j=1}^{n_1} X_j^{(1)}, \overline{X^{(2)}} = \frac{1}{n_2} \sum_{j=1}^{n_2} X_j^{(2)}, S^{(1)}, S^{(2)}$ lần lượt là trung bình mẫu và phương sai mẫu khi quan sát hai vecto $X^{(1)}, X^{(2)}$; $S = \frac{1}{n_1+n_2-2}[n_1 S^{(1)} + n_2 S^{(2)}]$. $F_{k_1,k_2}(\alpha)$ là phân vị trên mức α của phân bố F với các bậc tự do là k_1, k_2 sẽ là tiêu chuẩn

So sánh hai vecto trung bình

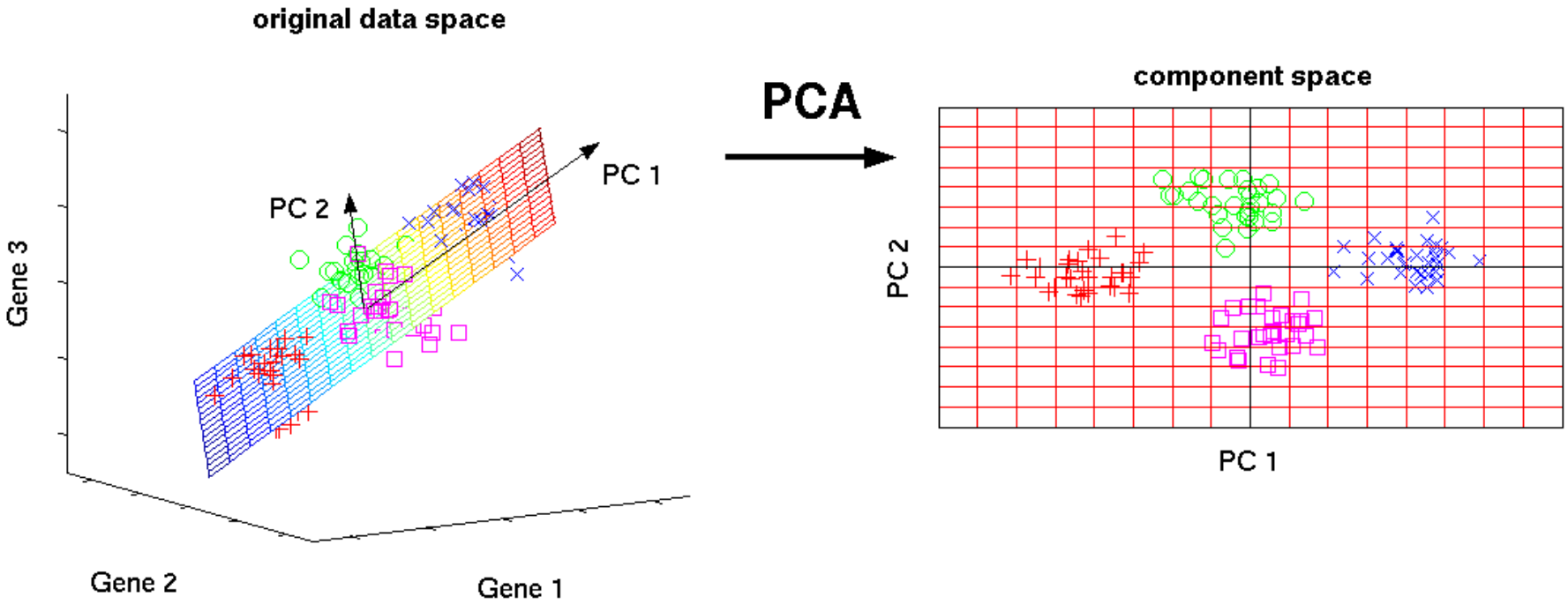
Với mẫu thu thập dạng cặp, giả thiết H_0 bị bác bỏ nếu thống kê

$$T_0^2 = n(\overline{D})' S_d^{-1} \overline{D} > F_{p,n-p}(\alpha)$$

trong đó $D_j = X_j^{(1)} - X_j^{(2)}$ và $\overline{D} = \frac{1}{n}\sum_{j=1}^{n} D_j$, S_d là vector trung bình mẫu và ma trận phương sai mẫu.

Phân tích thành phần chính

Phân tích thành phần chính



Phân tích thành phần chính

Cho vector ngẫu nhiên k chiều $X = (X_1, \ldots, X_k)' \in \mathbb{R}^k$ có ma trận hiệp phương sai $cov(X) = \Sigma$.

1.Cấu trúc của các thành phần chính

Mô hình

$$\begin{aligned} Y_1 &= a_1'X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1k}X_k \\ Y_2 &= a_2'X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2k}X_k \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ Y_k &= a_k'X = a_{k1}X_1 + a_{k2}X_2 + \cdots + a_{kk}X_k \end{aligned} \tag{20}$$

Khi đó

$$D(Y_i) = a_i'\Sigma a_i, \tag{21}$$

$$cov(Y_i, Y_j) = a_i'\Sigma a_j. \tag{22}$$

Bài toán: Các biến $Y_1, \cdots, Y_k$ thỏa mãn

$$
\begin{aligned}
D(Y_1) &= \max_{a_1' a_1 = 1} D(a_1' X), \\
D(Y_2) &= \max_{a_2' a_2 = 1;\ cov(Y_1, a_2' X) = 0} D(a_2' X), \\
&\cdots\cdots \\
D(Y_k) &= \max_{a_k' a_k = 1;\ cov(Y_i, a_k' X) = 0, i = 1 \div k-1} D(a_k' X).
\end{aligned}
$$

được gọi là các thành phần chính của vecto X.

Thành phần chính từ ma trận hiệp phương sai

Cho vector X có $cov(X) = \Sigma$. Giả sử $(\lambda_1, e_1), \ldots, (\lambda_k, e_k)$ là k cặp giá trị riêng và vector riêng của Σ sao cho $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. Khi đó thành phần chính thứ i của vector X được xác định bởi:

$$Y_i = e_i' X, i = 1 \div k, \tag{23}$$

và với việc chọn như vậy, ta có

$$D(Y_i) = \lambda_i; cov(Y_i, Y_j) = 0, \quad \forall i, j = 1 \div k. \tag{24}$$

Ví dụ

Giả sử $X = (X_1, X_2, X_3)'$ là ma trận có hiệp phương sai là:

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Tìm các thành phần chính của X.

Ví dụ

Giả sử $X = (X_1, X_2, X_3)'$ là ma trận có hiệp phương sai là:

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Tìm các thành phần chính của X.
Ta có 3 cặp giá trị riêng và vector riêng của Σ là

$$\lambda_1 = 5,83, e_1 = [0,383; -0,924; 0]',$$
$$\lambda_2 = 2,00, e_2 = [0,0,1]',$$
$$\lambda_3 = 0,17, e_3 = [0,924; 0,383; 0]'.$$

Tính chất

- $\sum_{i=1}^{k} D(X_i) = \sum_{i=1}^{k} D(Y_i) = \lambda_1 + \cdots + \lambda_k$

Tính chất

- $\sum_{i=1}^{k} D(X_i) = \sum_{i=1}^{k} D(Y_i) = \lambda_1 + \cdots + \lambda_k$
- $\dfrac{\lambda_i}{\lambda_1 + \cdots + \lambda_k}$ là tỷ lệ của biến sai tổng cộng của các thành phần của X do thành phần chính thứ i gây ra

Tính chất

- $\sum_{i=1}^{k} D(X_i) = \sum_{i=1}^{k} D(Y_i) = \lambda_1 + \cdots + \lambda_k$
- $\dfrac{\lambda_i}{\lambda_1 + \cdots + \lambda_k}$ là tỷ lệ của biến sai tổng cộng của các thành phần của X do thành phần chính thứ i gây ra
- Hệ số tương quan và hiệp phương sai giữa thành phần chính Y_i và thành phần X_j của vector X là:

$$cov(Y_i, X_j) = \lambda_i e_{ij}, \tag{25}$$

$$\rho_{Y_i, X_j} = \frac{\sqrt{\lambda_i} e_{ij}}{\sqrt{\sigma_{jj}}}. \tag{26}$$

- Ký hiệu

$$\widehat{X}_j = \beta_{1j} Y_1 + \beta_{2j} Y_2 + \cdots + \beta_{pj} Y_p, j = 1 \div k$$

là các dự báo tuyến tính tốt nhất của thành phần thứ j theo p thành phần đầu tiên của vector thành phần chính $Y = (Y_1, \ldots, Y_k)' \quad (1 \leq p \leq k)$.

- Ký hiệu

$$\widehat{X}_j = \beta_{1j} Y_1 + \beta_{2j} Y_2 + \cdots + \beta_{pj} Y_p, j = 1 \div k$$

là các dự báo tuyến tính tốt nhất của thành phần thứ j theo p thành phần đầu tiên của vector thành phần chính $Y = (Y_1, \ldots, Y_k)' \quad (1 \leq p \leq k)$. Khi đó tỷ lệ của tổng bình phương của các phương sai của các phần dư $\sum_{j=1}^{k} E(\widehat{\varepsilon}_j)^2$ với $\sum_{j=1}^{k} D(X_j)$ được cho bởi

$$\frac{\sum_{j=1}^{k} E(\widehat{\varepsilon}_j)^2}{\sum_{1}^{k} \sigma_{jj}} = \frac{(\lambda_{p+1} + \cdots + \lambda_k)}{(\lambda_1 + \cdots + \lambda_k)}, \tag{27}$$

(tỉ số đó bằng 0 khi $p = k$), trong đó $\widehat{\varepsilon}_j = X_j - \widehat{X}_j$.

Biến chuẩn hóa

Cho $X = (X_1, \ldots, X_k)' \in \mathbb{R}^k$ có ma trận hiệp phương sai $cov(X) = \Sigma$, các biến chuẩn hóa nhận được như sau:

$$\begin{aligned} z_1 &= (X_1 - \mu_1)/\sigma_{11}^{1/2} \\ z_2 &= (X_2 - \mu_2)/\sigma_{22}^{1/2} \\ &\ldots \\ z_k &= (X_k - \mu_k)/\sigma_{kk}^{1/2} \end{aligned} \tag{28}$$

Biến chuẩn hóa

Cho $X = (X_1, \ldots, X_k)' \in \mathbb{R}^k$ có ma trận hiệp phương sai $cov(X) = \Sigma$, các biến chuẩn hóa nhận được như sau:

$$\begin{aligned} z_1 &= (X_1 - \mu_1)/\sigma_{11}^{1/2} \\ z_2 &= (X_2 - \mu_2)/\sigma_{22}^{1/2} \\ &\ldots \\ z_k &= (X_k - \mu_k)/\sigma_{kk}^{1/2} \end{aligned} \tag{28}$$

Dạng ma trận

$$Z = V^{-1/2}(X - \mu), \tag{29}$$

trong đó $V = diag(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{kk})$.

Biến chuẩn hóa

Cho $X = (X_1, \ldots, X_k)' \in \mathbb{R}^k$ có ma trận hiệp phương sai $cov(X) = \Sigma$, các biến chuẩn hóa nhận được như sau:

$$\begin{aligned} z_1 &= (X_1 - \mu_1)/\sigma_{11}^{1/2} \\ z_2 &= (X_2 - \mu_2)/\sigma_{22}^{1/2} \\ &\ldots \\ z_k &= (X_k - \mu_k)/\sigma_{kk}^{1/2} \end{aligned} \tag{28}$$

Dạng ma trận

$$Z = V^{-1/2}(X - \mu), \tag{29}$$

trong đó $V = diag(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{kk})$. Tính chất :

$$E(Z) = 0 \; ; \; var(Z) = V^{-1/2}\Sigma V^{-1/2} = \rho,$$

Thành phần chính từ ma trận tương quan

Thành phần chính thứ i của vector đã được chuẩn hoá $Z = (Z_1, \ldots, Z_k)$ với ma trận ρ được xác định bởi:

$$Y_i^0 = e_i' Z = e_i' V^{-1/2}(X - \mu), i = 1 \div k. \tag{30}$$

Thành phần chính từ ma trận tương quan

Thành phần chính thứ i của vector đã được chuẩn hoá $Z = (Z_1, \ldots, Z_k)$ với ma trận ρ được xác định bởi:

$$Y_i^0 = e_i' Z = e_i' V^{-1/2}(X - \mu), i = 1 \div k. \tag{30}$$

Hơn nữa

$$\sum_1^k D(Y_i^0) = \sum_{i=1}^k D(Z_i) = k, \tag{31}$$

và

$$\rho_{Y_i^0, Z_j} = e_{ij}\sqrt{\lambda_i}; i, j = 1 \div k,$$

trong đó $(\lambda_1, e_1), \ldots, (\lambda_k, e_k)$ là các cặp giá trị riêng và vector riêng của ρ với $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$.

Ví dụ

Giả sử $X = [X_1, X_2]$ có ma trận hiệp phương sai

$$\Sigma = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$$

và ma trận tương quan tương ứng là

$$\begin{bmatrix} 1 & 0,4 \\ 0,4 & 1 \end{bmatrix}.$$

Tìm các thành phần chính của X và của biến chuẩn hóa X

Phân tích nhân tố

Cho vector ngẫu nhiên có thể quan sát được $X' = (X_1, \ldots, X_k)$ có $E(X) = \mu$ và $cov(X) = \Sigma$.

1. Mô hình nhân tố

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2 \\ &\cdots \\ X_k - \mu_k &= l_{k1}F_1 + l_{k2}F_2 + \cdots + l_{km}F_m + \varepsilon_k \end{aligned} \tag{32}$$

Cho vector ngẫu nhiên có thể quan sát được $X' = (X_1, \ldots, X_k)$ có $E(X) = \mu$ và $cov(X) = \Sigma$.

1. Mô hình nhân tố

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2 \\ &\cdots \\ X_k - \mu_k &= l_{k1}F_1 + l_{k2}F_2 + \cdots + l_{km}F_m + \varepsilon_k \end{aligned} \tag{32}$$

trong đó các biến ngẫu nhiên không quan sát được $F_1, \ldots, F_m$ $(m < k)$ gọi là các nhân tố chung và k biến cộng thêm $\varepsilon_1, \ldots, \varepsilon_k$ được gọi là các sai số hoặc các nhân tố xác định.

Cho vector ngẫu nhiên có thể quan sát được $X' = (X_1, \ldots, X_k)$ có $E(X) = \mu$ và $cov(X) = \Sigma$.

1. Mô hình nhân tố

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2 \\ &\cdots \\ X_k - \mu_k &= l_{k1}F_1 + l_{k2}F_2 + \cdots + l_{km}F_m + \varepsilon_k \end{aligned} \tag{32}$$

trong đó các biến ngẫu nhiên không quan sát được $F_1, \ldots, F_m$ $(m < k)$ gọi là các nhân tố chung và k biến cộng thêm $\varepsilon_1, \ldots, \varepsilon_k$ được gọi là các sai số hoặc các nhân tố xác định.

Mô hình nhân tố dạng ma trận

$$\underset{(k\times 1)}{X - \mu} = \underset{(k\times m)}{L} \times \underset{(m\times 1)}{F} + \underset{(k\times 1)}{\varepsilon} \tag{33}$$

Mô hình nhân tố trực giao

$$\underset{(k\times 1)}{X-\mu} = \underset{(k\times m)}{L} \times \underset{(m\times 1)}{F} + \underset{(k\times 1)}{\varepsilon} \tag{34}$$

trong đó

$$E(F) = 0, cov(F) = E(FF') = I$$
$$E(\varepsilon) = 0; cov(\varepsilon) = E(\varepsilon\varepsilon') = \psi = diag(\psi_1, \ldots, \psi_k)$$
$$cov(\varepsilon, F) = E(\varepsilon F') = \underset{(k\times m)}{0}$$

Mô hình nhân tố trực giao

$$\underset{(k\times 1)}{X-\mu} = \underset{(k\times m)}{L} \times \underset{(m\times 1)}{F} + \underset{(k\times 1)}{\varepsilon} \tag{34}$$

trong đó

$$E(F) = 0, cov(F) = E(FF') = I$$
$$E(\varepsilon) = 0; cov(\varepsilon) = E(\varepsilon\varepsilon') = \psi = diag(\psi_1, \ldots, \psi_k)$$
$$cov(\varepsilon, F) = E(\varepsilon F') = \underset{(k\times m)}{0}$$

Tính chất

$$\begin{aligned} cov(X) &= \Sigma = LL' + \psi. \\ cov(X, F) &= L \end{aligned} \tag{35}$$

Ví dụ.

Xét ma trận hiệp phương sai sau:

$$\Sigma = \begin{bmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{bmatrix}$$

Tìm phân tích dạng $\Sigma = LL' + \psi$.

Dễ thấy rằng :

$$\Sigma = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} 4 & 7 & -1 & 1 \\ 1 & 2 & 6 & 8 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$
$$= LL' + \psi$$

$h_1^2 = 4^2 + 1^2 = 17, \psi_1 = 2$. Như vậy: $\sigma_{11} = 19 = 17 + 2 = h_1^2 + \psi_1$.

Ước lượng mô hình

Cho N quan sát độc lập $X^1, \ldots, X^N$ về vector ngẫu nhiên X. Chúng ta cần phải trả lời các câu hỏi sau:

1. Mô hình nhân tố với m nhỏ có phù hợp với thực tế không?
2. Mô hình phân tích ma trận hiệp phương sai có thích hợp không?

Ước lượng mô hình

Cho N quan sát độc lập $X^1, \ldots, X^N$ về vector ngẫu nhiên X. Chúng ta cần phải trả lời các câu hỏi sau:

1. Mô hình nhân tố với m nhỏ có phù hợp với thực tế không?
2. Mô hình phân tích ma trận hiệp phương sai có thích hợp không?

Các bước thực hiện

1. Exploratory factor analysis : Ví dụ ta ước lượng ma trận $\Sigma = S$ và $\mu = \bar{X}$ và sử dụng phương pháp thành phần chính để ước lượng ma trận L, ψ.
2. Confirmatory factor analysis: Thực hiện các kiểm định thống kê để kiểm tra sự phù hợp của mô hình.

Phương pháp phân tích thành phần chính

Giả sử $(\lambda_1, e_1), \ldots, (\lambda_k, e_k)$ là k cặp giá trị riêng và vector riêng của ma trận $cov(X) = \Sigma$. Khi đó:

$$\begin{aligned}\Sigma &= \lambda_1 e_1 e_1' + \cdots + \lambda_k e_k e_k' \\ &= [\sqrt{\lambda_1} e_1 \vdots \cdots \vdots \sqrt{\lambda_k} e_k][\sqrt{\lambda_1} e_1 \vdots \cdots \vdots \sqrt{\lambda_k} e_k]'.\end{aligned}$$

Phương pháp phân tích thành phần chính

Giả sử $(\lambda_1, e_1), \ldots, (\lambda_k, e_k)$ là k cặp giá trị riêng và vector riêng của ma trận $cov(X) = \Sigma$. Khi đó:

$$\begin{aligned}\Sigma &= \lambda_1 e_1 e_1' + \cdots + \lambda_k e_k e_k' \\ &= [\sqrt{\lambda_1} e_1 \vdots \cdots \vdots \sqrt{\lambda_k} e_k][\sqrt{\lambda_1} e_1 \vdots \cdots \vdots \sqrt{\lambda_k} e_k]'.\end{aligned}$$

Với $m = k$ nhân tố thì

$$\underset{(k\times k)}{\Sigma} = \underset{(k\times k)}{L}\underset{(k\times k)}{L}' + 0 = LL', \tag{36}$$

trong đó $L = [\sqrt{\lambda_1} e_1 \vdots \cdots \vdots \sqrt{\lambda_k} e_k]$.

Với $m < k$,

$$\Sigma = \underbrace{[\sqrt{\lambda_1}e_1 \vdots \cdots \vdots \sqrt{\lambda_m}e_m]}_{L}\underbrace{[\sqrt{\lambda_1}e_1 \vdots \cdots \vdots \sqrt{\lambda_m}e_m]'}_{L'}$$
$$+ \lambda_{m+1}e_{m+1}e'_{m+1} + \cdots + \lambda_k e_k e'_k$$

Nếu $k - m$ giá trị riêng $\lambda_{m+1}, \ldots, \lambda_k$ là nhỏ thì đại lượng $\lambda_{m+1}e_{m+1}e'_{m+1} + \cdots + \lambda_k e_k e'_k$ có thể bỏ qua. Khi đó,

Với $m < k$,

$$\Sigma = \underbrace{[\sqrt{\lambda_1}e_1 \vdots \cdots \vdots \sqrt{\lambda_m}e_m]}_{L}\underbrace{[\sqrt{\lambda_1}e_1 \vdots \cdots \vdots \sqrt{\lambda_m}e_m]'}_{L'} + \lambda_{m+1}e_{m+1}e'_{m+1} + \cdots + \lambda_k e_k e'_k$$

Nếu $k - m$ giá trị riêng $\lambda_{m+1}, \ldots, \lambda_k$ là nhỏ thì đại lượng $\lambda_{m+1}e_{m+1}e'_{m+1} + \cdots + \lambda_k e_k e'_k$ có thể bỏ qua. Khi đó,

$$\Sigma \approx \underbrace{[\sqrt{\lambda_1}e_1 \vdots \cdots \vdots \sqrt{\lambda_m}e_m]}_{L}\underbrace{[\sqrt{\lambda_1}e_1 \vdots \cdots \vdots \sqrt{\lambda_m}e_m]'}_{L'} + diag(\psi_1, \ldots, \psi_k) \tag{37}$$

trong đó $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2; i = 1 \div k$.

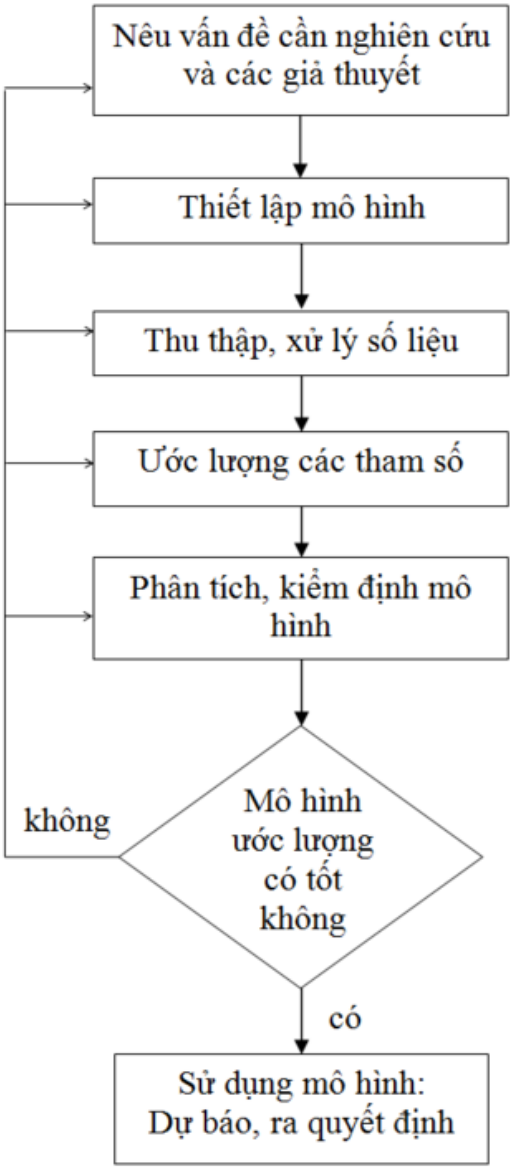
Ví dụ

Ma trận của 5 biến

$$\begin{bmatrix} 1,00 & 0,02 & 0,96 & 0,42 & 0,01 \\ . & 1,00 & 0,13 & 0,71 & 0,85 \\ . & . & 1,00 & 0,50 & 0,11 \\ . & . & . & 1,00 & 0,79 \\ . & . & . & . & 1,00 \end{bmatrix}$$

Tìm ma trận tải trọng của hai nhân tố theo phương pháp thành phần chính.

Chương 4. Mô hình hồi quy tuyến tính bội


Nêu vấn đề cần nghiên cứu và các giả thuyết
Thiết lập mô hình
Thu thập, xử lý số liệu
Ước lượng các tham số
Phân tích, kiểm định mô hình
Mô hình ước lượng có tốt không
không
có
Sử dụng mô hình: Dự báo, ra quyết định

1.Mô hình hồi quy tuyến tính cổ điển

Mô hình

Mô hình quy hồi tuyến tính cổ điển khẳng định rằng Y phụ thuộc tuyến tính vào k biến độc lập $X_1, \ldots, X_k$ và sai số ngẫu nhiên ε

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon \tag{38}$$

trong đó $\beta_i, i = 0 \div k$ là các hệ số chưa biết.

1.Mô hình hồi quy tuyến tính cổ điển

Mô hình

Mô hình quy hồi tuyến tính cổ điển khẳng định rằng Y phụ thuộc tuyến tính vào k biến độc lập $X_1, \ldots, X_k$ và sai số ngẫu nhiên ε

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon \tag{38}$$

trong đó $\beta_i, i = 0 \div k$ là các hệ số chưa biết. **Dữ liệu**

$$\begin{array}{cccc} x_{11} & x_{12} & \cdots & x_{1k} & y_1 \\ x_{21} & x_{22} & \cdots & x_{2k} & y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{n2} & \cdots & x_{Nk} & y_N \end{array}$$

Khi đó mô hình các số liệu quan sát sẽ là:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_N &= \beta_0 + \beta_1 x_{N1} + \cdots + \beta_k x_{Nk} + \varepsilon_N \end{aligned} \tag{39}$$

trong đó các sai số $\varepsilon_1, \ldots, \varepsilon_N$ thoả mãn 3 điều kiện sau:

1. $E(\varepsilon_j) = 0$
2. $D(\varepsilon_j) = \sigma^2$ (phương sai không đổi)
3. $cov(\varepsilon_i, \varepsilon_j) = 0$ với mọi $i \neq j = 1 \div N$

Dạng ma trận

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{121} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

Dạng ma trận

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{121} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

hoặc

$$\underset{(N\times 1)}{Y} = \underset{(N\times(k+1))}{X}\underset{(k+1\times 1)}{\beta} + \underset{(N\times 1)}{\varepsilon} \qquad (3.1.2')$$

với

1. $E(\varepsilon) = 0$
2. $cov(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I_N$

2. Ước lượng bình phương tối thiểu

Hàm tổn thất

$$S(b) = \sum_{j=1}^{n}(y_j - b_0 - b_1 x_{j1} - \cdots - b_k x_{jk})^2$$
$$= (Y - Xb)'(Y - Xb) \to \min \tag{40}$$

2. Ước lượng bình phương tối thiểu

Hàm tổn thất

$$S(b) = \sum_{j=1}^{n}(y_j - b_0 - b_1 x_{j1} - \cdots - b_k x_{jk})^2$$
$$= (Y - Xb)'(Y - Xb) \to \min \quad (40)$$

Ước lượng

- Đại lượng $\widehat{\beta} = (X'X)^{-1}X'Y$ cực tiểu hoá $S(b)$ được gọi là ước lượng bình phương cực tiểu của β

2. Ước lượng bình phương tối thiểu

Hàm tổn thất

$$S(b) = \sum_{j=1}^{n}(y_j - b_0 - b_1 x_{j1} - \cdots - b_k x_{jk})^2$$
$$= (Y - Xb)'(Y - Xb) \to \min \quad (40)$$

Ước lượng

- Đại lượng $\widehat{\beta} = (X'X)^{-1}X'Y$ cực tiểu hoá $S(b)$ được gọi là ước lượng bình phương cực tiểu của β
- $\widehat{\varepsilon_j} = y_j - \widehat{\beta_0} - \widehat{\beta_1} x_{j1} - \cdots - \widehat{\beta_k} x_{jk}, j = 1 \div n$ là phần dư của phép hồi quy tuyến tính

2. Ước lượng bình phương tối thiểu

Hàm tổn thất

$$S(b) = \sum_{j=1}^{n}(y_j - b_0 - b_1 x_{j1} - \cdots - b_k x_{jk})^2$$
$$= (Y - Xb)'(Y - Xb) \to \min \quad (40)$$

Ước lượng

- Đại lượng $\widehat{\beta} = (X'X)^{-1}X'Y$ cực tiểu hoá $S(b)$ được gọi là ước lượng bình phương cực tiểu của β
- $\widehat{\varepsilon_j} = y_j - \widehat{\beta_0} - \widehat{\beta_1} x_{j1} - \cdots - \widehat{\beta_k} x_{jk}, j = 1 \div n$ là phần dư của phép hồi quy tuyến tính
- Phương trình

$$\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_k} X_k$$

được gọi là *phương trình hồi quy tuyến tính mẫu.*

Tính chất ước lượng bằng phương pháp bình phương cực tiểu

1. Ước lượng $\widehat{\beta}$ là ước lượng không chệch với

$$E\widehat{\beta} = \beta; cov(\widehat{\beta}) = \sigma^2(X'X)^{-1} \tag{41}$$

Tính chất ước lượng bằng phương pháp bình phương cực tiểu

1. Ước lượng $\widehat{\beta}$ là ước lượng không chệch với

$$E\widehat{\beta} = \beta; cov(\widehat{\beta}) = \sigma^2(X'X)^{-1} \tag{41}$$

2. Phần dư $\widehat{\varepsilon}$ có tính chất: $\bar{\widehat{\varepsilon}} = 0 \Leftrightarrow \bar{Y} = \bar{\widehat{Y}}$

$$E(\widehat{\varepsilon}) = 0; cov(\widehat{\varepsilon}) = \sigma^2(I - H), \tag{42}$$

trong đó $H = X(X'X)^{-1}X'$

Tính chất ước lượng bằng phương pháp bình phương cực tiểu

1. Ước lượng $\widehat{\beta}$ là ước lượng không chệch với

$$E\widehat{\beta} = \beta; cov(\widehat{\beta}) = \sigma^2(X'X)^{-1} \tag{41}$$

2. Phần dư $\widehat{\varepsilon}$ có tính chất: $\bar{\widehat{\varepsilon}} = 0 \Leftrightarrow \bar{Y} = \bar{\widehat{Y}}$

$$E(\widehat{\varepsilon}) = 0; cov(\widehat{\varepsilon}) = \sigma^2(I - H), \tag{42}$$

trong đó $H = X(X'X)^{-1}X'$

3. $\widehat{\sigma}^2 = \widehat{\varepsilon}'\widehat{\varepsilon}/(N-k-1) = \sum_1^N \widehat{\varepsilon}_j^2/(N-k-1)$ là ước lượng không chệch của σ^2, tức là $E(\widehat{\sigma}^2) = \sigma^2$.

Tính chất ước lượng bằng phương pháp bình phương cực tiểu

1. Ước lượng $\widehat{\beta}$ là ước lượng không chệch với

$$E\widehat{\beta} = \beta; cov(\widehat{\beta}) = \sigma^2(X'X)^{-1} \tag{41}$$

2. Phần dư $\widehat{\varepsilon}$ có tính chất: $\bar{\widehat{\varepsilon}} = 0 \Leftrightarrow \bar{Y} = \bar{\widehat{Y}}$

$$E(\widehat{\varepsilon}) = 0; cov(\widehat{\varepsilon}) = \sigma^2(I - H), \tag{42}$$

trong đó $H = X(X'X)^{-1}X'$

3. $\widehat{\sigma}^2 = \widehat{\varepsilon}'\widehat{\varepsilon}/(N-k-1) = \sum_1^N \widehat{\varepsilon}_j^2/(N-k-1)$ là ước lượng không chệch của σ^2, tức là $E(\widehat{\sigma}^2) = \sigma^2$.
4. $\widehat{\beta}$ và $\widehat{\varepsilon}$ là không tương quan

$$cov(\widehat{\beta}, \widehat{\varepsilon}) = 0; cov(\widehat{\beta}, \widehat{\sigma}^2) = 0; \tag{43}$$

Hệ số xác định R

Đại lượng

$$R^2 = \frac{\widehat{Y'}\widehat{Y} - N(\bar{y})^2}{Y'Y - N(\bar{y})^2} = \frac{\sum_1^N \hat{y}_j^2 - N(\bar{y})^2}{\sum_1^N y_j^2 - N(\bar{y})^2} \tag{44}$$

là bình phương của hệ số xác định, đó là tỷ lệ biến thiên của các biến y_j được giải thích bởi các biến $x_{j1}, \ldots, x_{jk}$.

Miền tin cậy đồng thời của β

Miền tin cậy đồng thời của β

Miền tin cậy đồng thời mức $1-\alpha$ của β xác định bởi

$$(\beta-\widehat{\beta})'X'X(\beta-\widehat{\beta}) \leq (k+1)\widehat{\sigma}^2 F_{k+1,N-k-1}(\alpha) \qquad (45)$$

trong đó $F_{k+1,N-k-1}(\alpha)$ là phân vị trên mức α của phân bố F với $k+1$ và $N-k-1$ bậc tự do

Miền tin cậy đồng thời của β

Miền tin cậy đồng thời mức $1-\alpha$ của β xác định bởi

$$(\beta-\widehat{\beta})'X'X(\beta-\widehat{\beta}) \leq (k+1)\widehat{\sigma}^2 F_{k+1,N-k-1}(\alpha) \qquad (45)$$

trong đó $F_{k+1,N-k-1}(\alpha)$ là phân vị trên mức α của phân bố F với $k+1$ và $N-k-1$ bậc tự do

Khoảng tin cậy của β_i với mức tin cậy $(1-\alpha)$

$$\beta_i \in \left(\widehat{\beta}_i - t_{n-k-1}\left(\frac{\alpha}{2(k+1)}\right)\sqrt{\widehat{D}(\widehat{\beta}_i)}; \widehat{\beta}_i + t_{n-k-1}\left(\frac{\alpha}{2(k+1)}\right)\sqrt{\widehat{D}(\widehat{\beta}_i)}\right) \qquad (46)$$

Chú ý: Nếu 0 không thuộc vào khoảng tin cậy này thì ta coi $\beta_i \neq 0$.

Khảo sát phần dư

Giả thiết $H_0 : \varepsilon$ có phân phối chuẩn $N(0, \sigma^2 I)$; Đối thiết H_1 : ngược lại.

Khảo sát phần dư

Giả thiết H_0 : ε có phân phối chuẩn $N(0, \sigma^2 I)$; Đối thiết H_1 : ngược lại.

Miền bác bỏ giả thiết H_0 là

$$|T| > t_{N-k-2}(\frac{\alpha}{2}),$$

trong đó

$$T = \frac{(N-k-1)^{1/2}\widetilde{\varepsilon}^*}{\left[\sum_{j=1}^{N-k-1}(\varepsilon_j^* - \widetilde{\varepsilon}^*)^2/(N-k-2)\right]^{1/2}}$$

với $\bar{\varepsilon}^* = \sum_1^N \varepsilon_j^*/N; \widetilde{\varepsilon}^* = \sum_{j=1}^{N-k-1} \varepsilon_j^*/(N-k-1)$

Tiêu chuẩn F

Giả thiết $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$; Đối thiết H_1 : ngược lại.

Tiêu chuẩn F

Giả thiết $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$; Đối thiết H_1 : ngược lại.

Miền bác bỏ giả thiết H_0 là

$$F > F_{k,N-k-1}(\alpha),$$

trong đó $F = \frac{R^2(N-k-1)}{k(1-R^2)}$

Ước lượng khoảng cho Y khi $X = X^0$

$$X^{0\prime}\widehat{\beta} \pm \widehat{\sigma} t_{N-k-1}\left(\frac{\alpha}{2}\right)\sqrt{X^{0\prime}(X'X)^{-1}X^0 + 1}.$$

Ví dụ Để nghiên cứu về sự phụ thuộc giữa doanh thu (Y) và chi phí quảng cáo (X_1), chi phí tiếp thị (X_2) người ta điều tra ngẫu nhiên doanh thu của 12 công ty trong 12 thời kỳ, kết quả ta có bảng sau

STT	x_0	x_1	x_2	y
1	1	18	10	127
2	1	25	11	149
3	1	19	6	106
4	1	24	16	163
5	1	15	7	102
6	1	26	17	180
7	1	25	14	161
8	1	16	12	128
9	1	17	12	139
10	1	23	12	144
11	1	22	14	159
12	1	15	15	138

Hãy ước lượng mô hình biểu diễn sự phụ thuộc của doanh thu

Mô hình

Giả sử m biến phụ thuộc (biến đáp ứng) $Y_1, Y_2, \ldots, Y_m$ biểu diễn qua một bộ các biến dự báo $X_1, \ldots, X_k$ theo mô hình sau:

$$\begin{array}{ll} Y_1 & = \beta_{01} + \beta_{11}X_1 + \cdots + \beta_{k1}X_k + \varepsilon_1 \\ Y_2 & = \beta_{02} + \beta_{12}X_1 + \cdots + \beta_{k2}X_k + \varepsilon_2 \\ \cdots & \qquad\qquad \cdots \\ Y_m & = \beta_{0m} + \beta_{1m}X_1 + \cdots + \beta_{km}X_k + \varepsilon_m \end{array} \tag{47}$$

trong đó vector sai số $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_m]'$ có $E(\varepsilon) = 0, cov(\varepsilon) = \Sigma$.

Mô hình

Giả sử m biến phụ thuộc (biến đáp ứng) $Y_1, Y_2, \ldots, Y_m$ biểu diễn qua một bộ các biến dự báo $X_1, \ldots, X_k$ theo mô hình sau:

$$\begin{array}{ll} Y_1 & = \beta_{01} + \beta_{11}X_1 + \cdots + \beta_{k1}X_k + \varepsilon_1 \\ Y_2 & = \beta_{02} + \beta_{12}X_1 + \cdots + \beta_{k2}X_k + \varepsilon_2 \\ \cdots & \qquad\qquad \cdots \\ Y_m & = \beta_{0m} + \beta_{1m}X_1 + \cdots + \beta_{km}X_k + \varepsilon_m \end{array} \tag{47}$$

trong đó vector sai số $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_m]'$ có $E(\varepsilon) = 0, cov(\varepsilon) = \Sigma$.

Dữ liệu

$$\begin{array}{cccccccc} x_{11} & x_{12} & \cdots & x_{1k} & y_{11} & \cdots & y_{1m} \\ x_{21} & x_{22} & \cdots & x_{2k} & y_{21} & \cdots & y_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nk} & y_{N1} & \cdots & y_{Nm} \end{array}$$

Ký hiệu:

$$\underset{(N\times m)}{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ y_{N1} & y_{N2} & \cdots & y_{Nm} \end{bmatrix} = [Y_{(1)} \vdots Y_{(2)} \vdots \cdots \vdots Y_{(m)}]$$

$$\underset{(N\times k+1)}{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{N1} & \cdots & x_{Nk} \end{bmatrix}; \underset{(k+1\times m)}{\beta} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{k1} & \beta_{k2} & \cdots & \beta_{km} \end{bmatrix}$$

$$\underset{(N\times m)}{\varepsilon} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \varepsilon_{N1} & \varepsilon_{N2} & \cdots & \varepsilon_{Nm} \end{bmatrix} = [\varepsilon_{(1)} \vdots \varepsilon_{(2)} \vdots \cdots \vdots \varepsilon_{(m)}]$$

Mô hình dạng ma trận:

$$Y = X\beta + \varepsilon \tag{48}$$

với giả thiết sau đây về ma trận sai số:

$$E(\varepsilon_{(j)}) = 0,\ cov(\varepsilon_{(i)}, \varepsilon_{(j)}) = \sigma_{ij} I_N \ \forall\ j, i = 1, 2, \ldots, m, \tag{49}$$

Mô hình dạng ma trận:

$$Y = X\beta + \varepsilon \tag{48}$$

với giả thiết sau đây về ma trận sai số:

$$E(\varepsilon_{(j)}) = 0,\ cov(\varepsilon_{(i)}, \varepsilon_{(j)}) = \sigma_{ij} I_N \ \forall\ j, i = 1, 2, \ldots, m, \tag{49}$$

Ước lượng hệ số bình phương tối thiểu của mô hình

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{50}$$

Ví dụ

xét hệ hai mô hình tuyến tính:

$$y_{j1} = \beta_{01} + \beta_{11}x_{j1} + \varepsilon_{j1}$$
$$y_{j2} = \beta_{02} + \beta_{12}x_{j1} + \varepsilon_{j2}, j = 1, 2, \ldots, 5$$

Với số liệu được cho trong bảng dưới đây:

x_{j1}	0	1	2	3	4
y_{j1}	1	4	3	8	9
y_{j2}	-1	-1	2	3	2

Hãy ước lượng hệ số và phần dư của mô hình.

Mô hình

Giả sử biến đáp ứng Y và các biến giải thích $X = (X_1, \ldots, X_k)'$ đều là các biến ngẫu nhiên. Vector $(Y, X_1, \ldots, X_k)'$ có vector trung bình và ma trận hiệp phương sai là

$$\mu = \begin{bmatrix} \mu_Y \\ \ldots \\ \mu_X \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{YY} & \vdots & \Sigma_{YX} \\ \ldots\ldots\ldots & & \\ \Sigma_{XY} & \vdots & \Sigma_{XX} \end{bmatrix}.$$

Mô hình

Giả sử biến đáp ứng Y và các biến giải thích $X = (X_1, \ldots, X_k)'$ đều là các biến ngẫu nhiên. Vector $(Y, X_1, \ldots, X_k)'$ có vector trung bình và ma trận hiệp phương sai là

$$\mu = \begin{bmatrix} \mu_Y \\ \ldots \\ \mu_X \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{YY} & \vdots & \Sigma_{YX} \\ \ldots\ldots & \ldots & \ldots\ldots \\ \Sigma_{XY} & \vdots & \Sigma_{XX} \end{bmatrix}.$$

Khi đó bài toán dự báo tuyến tính biến Y theo $X_1, \ldots, X_k$ theo mô hình

$$\widetilde{Y} = b_0 + b_1 X_1 + \cdots + b_k X_k = b_0 + b'X \tag{51}$$

sao cho $E(Y - b_0 - b'X)^2$ đạt cực tiểu.

Nghiệm của bài toán là :

$$b = \beta := \Sigma_{XX}^{-1}\Sigma_{XY}, b_0 = \beta_0 = \mu_Y - \beta'\mu_X \tag{52}$$

với sai số bình phương trung bình cực tiểu là

$$\begin{aligned} E\{(Y - \beta_0 - \beta' X)^2 &= \Sigma_{YY} - \Sigma'_{XY}\Sigma_{XX}^{-1}\Sigma_{XY} \\ &= \Sigma_{YY} - \Sigma'_{XY}\beta \end{aligned} \tag{53}$$

Phương trình (51) được gọi là phương trình hồi quy tuyến tính lý thuyết.
Hệ số tương quan bội giữa Y và X là

$$\rho_{Y,X} = [\frac{\Sigma'_{XY}\Sigma_{XX}^{-1}\Sigma_{XY}}{\Sigma_{YY}}]^{1/2} \tag{54}$$

Ví dụ

Cho Y, X_1, X_2 có ma trận giá trị trung bình và hiệp phương sai như sau

$$\mu = \begin{bmatrix} \mu_Y \\ \mu_X \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix} = \begin{bmatrix} 10 & 1 & -1 \\ 1 & 7 & 3 \\ -1 & 3 & 2 \end{bmatrix}$$

Hãy xác định phương trình HQTT của Y theo X_1, X_2 và sai số bình phương trung bình $E(\widehat{\varepsilon})^2$ và hệ số tương quan tuyến tính bội $\rho_{Y,X}$.

1. Bài toán phân biệt và phân lớp

Dựa trên quan sát p dấu hiệu $U_1, U_2, \ldots, U_p$ của một cá thể hoặc một đối tượng, bài toán phân biệt là cần phải xác định xem đối tượng hoặc cá thể đó thuộc vào 1 trong k nhóm đã xác định trước.

1. Bài toán phân biệt và phân lớp

Dựa trên quan sát p dấu hiệu $U_1, U_2, \ldots, U_p$ của một cá thể hoặc một đối tượng, bài toán phân biệt là cần phải xác định xem đối tượng hoặc cá thể đó thuộc vào 1 trong k nhóm đã xác định trước. Trong trường hợp k nhóm chưa biết, việc phân chia thành các nhóm được gọi là bài toán phân lớp

1. Bài toán phân biệt và phân lớp

Dựa trên quan sát p dấu hiệu $U_1, U_2, \ldots, U_p$ của một cá thể hoặc một đối tượng, bài toán phân biệt là cần phải xác định xem đối tượng hoặc cá thể đó thuộc vào 1 trong k nhóm đã xác định trước. Trong trường hợp k nhóm chưa biết, việc phân chia thành các nhóm được gọi là bài toán phân lớp

Ví dụ: Một nhà sinh học quan sát p dấu hiệu của một đối tượng sinh học, ông ta cần phải liệt đối tượng đó thuộc vào 1 trong k loại sinh vật nào.

2. Bài toán phân biệt

Cho U là số đo p dấu hiệu của một đối tượng và S là tập hợp các giá trị có thể có của U.

1. **Quy tắc phân biệt không ngẫu nhiên**: Chia không gian S thành k miền rời nhau $W_1, \cdots, W_k$. Khi đó

   Đối tượng thuộc nhóm i nếu $U \in W_i$, $i = 1, \cdots, k$

2. Bài toán phân biệt

Cho U là số đo p dấu hiệu của một đối tượng và S là tập hợp các giá trị có thể có của U.

1. **Quy tắc phân biệt không ngẫu nhiên**: Chia không gian S thành k miền rời nhau $W_1, \cdots, W_k$. Khi đó

$$\textit{Đối tượng thuộc nhóm } i \textit{ nếu } U \in W_i,\ i = 1, \cdots, k$$

$$\text{hoặc } \delta(U) = i, \text{ trong đó } \delta(U) = \sum_{i=1}^{k} i I_{W_i}(U)$$

2. Bài toán phân biệt

Cho U là số đo p dấu hiệu của một đối tượng và S là tập hợp các giá trị có thể có của U.

1. **Quy tắc phân biệt không ngẫu nhiên**: Chia không gian S thành k miền rời nhau $W_1, \cdots, W_k$. Khi đó

   Đối tượng thuộc nhóm i nếu $U \in W_i$, $i = 1, \cdots, k$

   hoặc $\delta(U) = i$, trong đó $\delta(U) = \sum_{i=1}^{k} i I_{W_i}(U)$

2. **Quy tắc phân biệt ngẫu nhiên**: Xác định k hàm không âm $\lambda_1(U), \ldots, \lambda_k(U)$ sao cho $\sum_{i=1}^{k} \lambda_i(U) = 1$. Khi đó

   Đối tượng thuộc nhóm i với xác suất là $\lambda_i(U)$

Hàm tổn thất

- $P_1(u), \ldots, P_k(u)$ là mật độ xác suất của U (trường hợp liên tục) hoặc là xác suất để U nhận giá trị u (rời rạc) trong nhóm $1, \ldots, k$.

Hàm tổn thất

- $P_1(u), \ldots, P_k(u)$ là mật độ xác suất của U (trường hợp liên tục) hoặc là xác suất để U nhận giá trị u (rời rạc) trong nhóm $1, \ldots, k$.
- r_{ij} là tổn thất gây ra khi liệt cá thể thuộc nhóm i vào nhóm j.

Hàm tổn thất

- $P_1(u), \ldots, P_k(u)$ là mật độ xác suất của U (trường hợp liên tục) hoặc là xác suất để U nhận giá trị u (rời rạc) trong nhóm $1, \ldots, k$.
- r_{ij} là tổn thất gây ra khi liệt cá thể thuộc nhóm i vào nhóm j.

Hàm tổn thất

- $P_1(u), \ldots, P_k(u)$ là mật độ xác suất của U (trường hợp liên tục) hoặc là xác suất để U nhận giá trị u (rời rạc) trong nhóm $1, \ldots, k$.
- r_{ij} là tổn thất gây ra khi liệt cá thể thuộc nhóm i vào nhóm j.

Khi đó, nếu đối tượng thuộc nhóm i thì

$$L_i^{\delta} = Er_{i\delta(U)} = \sum_{j=1}^{k} r_{ij} P_i(W_j)$$

$$L_i^{\lambda} = E[r_{i1}\lambda_1(U) + \cdots + r_{ik}\lambda_k(U)] = \int_S [r_{i1}\lambda_1(u) + \cdots + r_{ik}\lambda_k(u)]P_i(u)du$$

được gọi là hàm tổn thất của quy tắc δ hoặc λ đối với nhóm i.

So sánh hai quy tắc

- δ_1 là tốt hơn δ_2 nếu

$$L_i^{\delta_1} \leq L_i^{\delta_2}, \ \forall i = 1, \ldots, k \tag{55}$$

và với ít nhất một i : $L_i^{\delta_1} < L_i^{\delta_2}$.

- δ_1 tương đương với δ_2 nếu

$$L_i^{\delta_1} = L_i^{\delta_2}, \ \forall i = 1, \ldots, k \tag{56}$$

So sánh hai quy tắc

- δ_1 là tốt hơn δ_2 nếu

$$L_i^{\delta_1} \leq L_i^{\delta_2}, \ \forall i = 1, \ldots, k \tag{55}$$

và với ít nhất một $i : L_i^{\delta_1} < L_i^{\delta_2}$.

- δ_1 tương đương với δ_2 nếu

$$L_i^{\delta_1} = L_i^{\delta_2}, \ \forall i = 1, \ldots, k \tag{56}$$

Quy tắc chấp nhận được

Quy tắc phân biệt δ hoặc λ gọi là quy tắc chấp nhận được nếu không tồn tại quy tắc nào khác tốt hơn nó

Quy tắc Bayes

Giả sử k nhóm được phân chia theo tỷ lệ $\pi_1, \ldots, \pi_k$. Khi đó

$$\begin{aligned} L^\delta &= \pi_1 L_1^\delta + \pi_2 L_2^\delta + \cdots + \pi_k L_k^\delta \\ &= \sum_{j=1}^{k} [\pi_1 r_{1j} P_1(W_j) + \pi_2 r_{2j} P_2(W_j) + \cdots + \pi_k r_{kj} P_k(W_j)] \\ &= -S_1(W_1) - S_2(W_2) - \cdots - S_k(W_k) \end{aligned}$$

$$L^\lambda = -\int_S \left[\sum_{i=1}^{k} \lambda_i(u) S_i(u) \right] du$$

trong đó, $S_j(u) = -(\pi_1 r_{1j} P_1 + \cdots + \pi_k r_{kj} P_k)(u)$ được gọi là hàm thông tin phân biệt thứ j.

Quy tắc Bayes chấp nhận được

1. Quy tắc δ chấp nhận được :

$$u \in W_i^* \Leftrightarrow S_i(u) = \max_{1 \le j \le k} S_j(u)$$

2. Quy tắc λ chấp nhận được :
   - $\lambda_i^*(u) = 1, \lambda_j^*(u) = 0 \; \forall j \neq i$ nếu $S_i(u) > S_j(u) \; \forall j \neq i$.
   - Nếu $S_{i_1}(u) = \cdots = S_{i_r} > S_{i_{r+1}}(u) \ge \cdots \ge S_{i_k}(u)$, ta đặt $\lambda_{i_{r+1}}^* = \cdots = \lambda_{i_k}^* = 0$ còn $\lambda_{i_1}^*, \ldots, \lambda_{i_r}^*$ có thể chọn tuỳ ý sao cho $\sum_{j=1}^{r} \lambda_{i_j}^*(u) = 1$.

Ví dụ: Quy tắc phân biệt Gauss

Giả sử phân bố trong mỗi nhóm là phân bố chuẩn $N_p(\mu_i, A_i)$ với $A_i = A \; \forall i = 1, \cdots, k$. Tổn thất $r_{ii} = 0, r_{ij} = 1$ với $i \neq j$. Khi đó, *cá thể được xếp vào nhóm thứ i nếu* $\bar{S}_i(U)$ *lớn nhất,* trong đó

$$\bar{S}_i(U) = \mu_i' A^{-1} U - \frac{1}{2} \mu_i' A^{-1} \mu_i + \ln \pi_i, i = 1 \div k$$

Ví dụ

(Rao và Slater-1949). Xét việc phân biệt trạng thái thần kinh của một người dựa trên các số đo về 3 dấu hiệu tâm thần U_1, U_2, U_3. Sau đây là số liệu thống kê dựa trên việc đo 3 dấu hiệu trên 256 người dưới dạng các trung bình mẫu và ma trận hiệp phương sai mẫu. Giả sử các phân bố xác suất trong mỗi nhóm là phân bố chuẩn cùng ma trận hiệp phương sai. Các đại lượng π_i, μ_i, A được ước lượng từ mẫu. Hỏi rằng với quan sát của một người là $U_1 = 0,8201; U_2 = 1,6; U_3 = 0,68$, ta sẽ xếp người này vào nhóm nào?

Các nhóm	Cỡ mẫu	Trung bình mẫu của các nhóm		
		$\overline{u}_1^{(1)}$	$\overline{u}_1^{(1)}$	$\overline{u}_1^{(1)}$
1-Tâm thần bất an	114	2.9298	1.667	0,7281
2-Bị điên	33	3.0303	1.2424	0.5455
3-Bệnh thái nhân cách	32	3.8125	1.8438	0.8125
4-Bệnh hoang tưởng	17	4.7059	1.5882	1.1176
5-Thay đổi cá tính	5	1.4000	0.2000	0.0000
6-Trạng thái bình thường	55	0.6000	0.1455	0.2182

Ma trận hiệp phương sai mẫu $S = (s_{ij})$				Nghịch đảo của ma trận hiệp phương sai mẫu $S^{-1} = (s^{ij})$			
$N^o\_$	1	2	3	$N^o\_$	1	2	3
1	2,3008	0,2516	0,4742	1	0,5432	-0,2002	-0,4208
2	0,2516	0,6075	0,0358	2	-0,2002	1,7258	0,0558
3	0,4742	0,0358	0,5951	3	-0,4208	0,0558	2,0123

3. Bài toán phân lớp

Với dữ liệu quan sát về N đối tượng, ta cần phân chia chúng thành các nhóm khác nhau.

Khoảng cách giữa hai phần tử

Cho hai phần tử $x = (x_1, \ldots, x_k),\ y = (y_1, \ldots, y_k)$,

- Khoảng cách Euclide

$$d_1^2(x, y) = \sum_{i=1}^{k} (x_i - y_i)^2 = (x - y)(x - y)'.$$

- Khoảng cách thống kê

$$d_2^2(x, y) = (x - y)A(x - y)',$$

trong đó A là ma trận đối xứng xác định dương.

- Khoảng cách Minkowski

$$d_3(x,y) = \left(\sum_{i=1}^{k} |x_i - y_i|^m \right)^{\frac{1}{m}}, m = 1, 2, 3 \ldots$$

- Khoảng cách Canberra

$$d_4(x,y) = \sum_{i=1}^{k} |x_i - y_i| / (x_i + y_i)$$

(chỉ xác định cho các $x_i, y_i > 0$.)

- e) Hệ số Czekanowski

$$d_5(x,y) = 1 - 2 \sum_{i=1}^{k} \min(x_i, y_i) / \sum_{i=1}^{k} (x_i + y_i)$$

(chỉ xác định cho các $x_i, y_i > 0$.)

Phương pháp kết nối đơn cho N phần tử

1. Bắt đầu với N cụm, mỗi cụm chứa 1 phần tử và lập ma trận các khoảng cách cấp N là $D = \{d_{ik}\}$.
2. Tìm một ma trận khoảng cách của các cặp các cụm gần nhất. Giả sử khoảng cách giữa hai cụm gần nhất U, V là d_{UV}.
3. Gộp cụm U với V. Ký hiệu cụm mới là (UV). Lập các phần tử của ma trận khoảng cách mới bằng cách
   1. loại các hàng và các cột tương ứng với cụm U, V;
   2. thêm vào một hàng và một cột gồm các khoảng cách từ cụm (UV) đến các cụm còn lại.
4. Lặp lại bước 2-3 $N - 1$ lần. Tất cả các phần tử sẽ tạo thành một cụm duy nhất sau khi kết thúc thuật toán. Ghi lại sự nhận dạng của các cụm đã được kết hợp và mức độ (khoảng cách hoặc sự tương tự) mà ở đó việc kết hợp các cụm đã được thực hiện.

Ví dụ

Xét ma trận khoảng cách của 5 cá thể:

$$D = [d_{ik}] = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{pmatrix} \end{array}$$

Hãy tiến hành phân cụm (lớp) cho 5 cá thể trên.

Phương pháp $K-$ trung bình

1. Phân chia ngẫu nhiên các đối tượng vào K cụm ban đầu.
2. Từ toàn bộ danh sách các đối tượng, phân phối từng đối tượng cho cụm có trung tâm (TB) gần nó nhất (thường theo khoảng cách Euclide, có thể chuẩn hoá hoặc không chuẩn hoá các quan sát). Tính toán lại trung tâm cho cụm nhận được đối tượng mới và cho cụm mất đối tượng.
3. Lặp lại bước 2 cho đến khi không có sự phân phối lại.

ví dụ

Ta có bảng số liệu sau đây

Đối tượng	Quan sát	
	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Mục tiêu là phân chia các đối tượng vào $K = 2$ cụm mà thoả mãn mỗi đối tượng gần tâm của cụm chứa nó nhất.

THANK YOU