ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA TOÁN - CƠ - TIN HỌC

PHÂN TÍCH THỐNG KÊ NHIỀU CHIỀU

Multivariate Statistical Analysis

CHUYÊN ĐỀ:

CÔNG THỨC THEO CÁC DẠNG BÀI TẬP

Mã lớp học phần: MAT3452

Sinh viên: Tạ Quang Tùng

Lớp: K66A2 Toán Tin

Hà Nội, 2024

Mục lục

1 Phân bố chuẩn nhiều chiều **4**

1.1 Hàm mật độ đồng thời: 4

1.2 Ma trận tương quan mẫu: 4

1.3 Tính chất phân bố chuẩn nhiều chiều: 5

1.4 Ví dụ minh họa: 5

2 Phân bố chuẩn hai chiều **6**

2.1 Hàm mật độ: 6

2.2 Công thức chuyển đổi: 6

2.3 Chuyển 2 chiều về 1 chiều: 6

2.4 Ví dụ minh họa: 6

3 Đặc trưng của mẫu ngẫu nhiên **8**

3.1 Vector trung bình mẫu: 8

3.2 Ma trận hiệp phương sai mẫu: 8

3.3 Phân bố mẫu ngẫu nhiên nhiều chiều: 8

4 Phân tích thành phần chính của X **9**

4.1 Tính chất thành phần chính: 9

4.2 Dựa trên ma trận hiệp phương sai: 9

4.3 Dựa trên ma trận tương quan: 9

4.4 Ví dụ minh họa: 10

5 Phân tích nhân tố về dạng $\Sigma = LL' + \psi$ **12**

5.1 Phân tích nhân tố: 12

5.2 Phương pháp giải: 12

5.2.1 TH1: $m = k => \psi = 0$ 12

5.2.2 TH2: $m < k$ 12

5.3 Ví dụ minh họa: 13

5.3.1 Ví dụ 1: 13

5.3.2 Ví dụ 2: 13

6 Mô hình hồi quy tuyến tính bội **15**

6.1 Ước lượng mô hình: 15

6.2 Kiểm định mô hình có ý nghĩa thống kê: 15

6.3 Kiểm định các hệ số: 16

6.4 Ví dụ minh họa: 16

7 Mô hình hồi quy tuyến tính nhiều biến phụ thuộc **18**

7.1 Mô hình hồi quy: 18

7.2 Ước lượng hệ số bình phương tối thiểu: 18

7.3 Ví dụ minh họa: 18

8 Hồi quy theo các biến ngẫu nhiên **20**
8.1 Dự báo tuyến tính Y theo X_i: 20
8.2 Sai số bình phương cực tiểu: 20
8.3 Hệ số tương quan bội giữa Y và X: 20
8.4 Ví dụ minh họa: 20

9 Phân biệt lớp bằng quy tắc Bayes chấp nhận được **22**
9.1 Quy tắc phân biệt Gauss thuộc nhóm i: 22
9.2 Ví dụ minh họa: 23

10 Các phương pháp phân cụm trong bài toán phân lớp **25**
10.1 Khoảng các giữa 2 phần tử: 25
10.1.1 Khoảng cách Euclide: 25
10.1.2 Khoảng cách thống kê: 25
10.1.3 Khoảng cách Minkowski: 25
10.1.4 Khoảng cách Canberra: 25
10.1.5 Hệ số Czekanowski: 25
10.2 Phân cụm bằng phương pháp kết nối đơn: 26
10.2.1 Ma trận khoảng cách đối xứng: 26
10.2.2 Tìm cặp có khoảng cách bé nhất: 26
10.2.3 Ví dụ phương pháp kết nối đơn: 26
10.3 Phân cụm bằng phương pháp K-means: 28
10.3.1 Triển khai thuật toán: 28
10.3.2 Ví dụ minh họa: 28

11 Hướng dẫn sử dụng RStudio **30**
11.1 Hồi quy tuyến tính đơn: 30
11.1.1 Phương trình đường thẳng HQTT: 30
11.1.2 Lấy hệ số a, b từ Y = a + bX: 30
11.1.3 Kiểm định phần dư có phân bố chuẩn với GTTB = 0 30
11.1.4 Khoảng tin cậy $\alpha\%$ cho hàm số hồi quy: 30
11.1.5 Với Y = newValue, đưa ra dự đoán về giá trị của X với khoảng tin cậy $\alpha\%$ cho GTTB của X . . . 30
11.1.6 Bài toán kiểm định phần dư: 31
11.1.7 Kẻ đường tuyến tính: 31
11.1.8 Kiểm định phương sai không đổi: 31
11.2 Phân bố chuẩn: 31
11.2.1 Các hàm trong phân phối chuẩn $X \sim N(mean, sd^2)$: 31
11.2.2 Các đặc trưng cơ bản: 31
11.3 Phân phối chuân nhiều chiều: 32
11.3.1 Kiểm định GTTB của 1 biến có bằng value không? 32
11.3.2 Kiểm định GTTB 2 mẫu có sự khác biệt: 32
11.3.3 Kiểm định nhiều chiều với phân phối chuẩn: . . . 32

11.3.4 Kiểm định nhiều chiều: 32
11.3.5 Kiểm tra phân phối chuẩn nhiều chiều: 33
11.4 Phân tích thành phần chính: 33
11.4.1 Phân tích TPC trên ma trận hiệp phương sai: . . 33
11.4.2 Phân tích TPC trên ma trận tương quan: 33
11.5 Phân tích nhân tố: 34
11.6 Mô hình hồi quy tuyến tính bội: 34
11.6.1 Kiểm định phần dư có tương quan không? 34
11.6.2 Phần dư có tuân theo phân phối chuẩn với GTTB = 0 không? 34
11.6.3 Các hệ số trong mô hình có thực sự khác 0 không? 35
11.6.4 Phân tích theo phương pháp forward/backward/both: 35
11.6.5 Kiểm tra sự phục thuộc của từng biến: 35
11.6.6 Ước lượng khoảng tin cậy $\alpha\%$ cho các hệ số (β_n và KTC tương ứng) 36
11.6.7 Bài toán dự đoán mô hình: 36
11.7 Các phép toán với ma trận: 36
11.8 Một số loại biểu đồ: 36

1 Phân bố chuẩn nhiều chiều

$$N(\mu, \Sigma) \sim X$$

- Cho $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ ... \\ \mu_m \end{pmatrix}$; $\Sigma = (\sigma_{ij})_{mm} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & ... & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & ... & \sigma_{2m} \\ ... & ... & ... & ... \\ \sigma_{m1} & \sigma_{m2} & ... & \sigma_{mm} \end{pmatrix}$
- $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ ... \\ X_m \end{pmatrix}$ có phân bố chuẩn

1.1 Hàm mật độ đồng thời:

$$f(x_1, x_2, ..., x_m) = \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma|^{\frac{1}{2}}} * exp(\frac{-1}{2}(x-\mu)^T * \Sigma^{-1} * (x-\mu))$$

1.2 Ma trận tương quan mẫu:

$$\rho = V^{\frac{-1}{2}} * \Sigma * V^{\frac{-1}{2}} \text{ } với \text{ } V = \begin{pmatrix} c_{11} & ... & 0 \\ ... & ... & ... \\ 0 & ... & c_{mm} \end{pmatrix}$$

Chú ý: $\sigma_{ij} = cov(X_i, X_j)$

- $\begin{cases} \sigma_{11} = cov(X_1, X_1) = DX_1 = \sigma_1^2 \\ \sigma_{22} = cov(X_2, X_2) = DX_2 = \sigma_2^2 \\ ... \\ \sigma_{mm} = cov(X_m, X_m) = DX_m = \sigma_m^2 \end{cases}$
- $\sigma_{12} = \frac{cov(X_1,X_2)}{\sqrt{DX_1 * DX_2}} = \frac{E(X_1X_2) - EX_1 * EX_2}{\sqrt{DX_1 * DX2}}$
- $\sigma_{ij} = 0 \ \forall \ i \neq j$ tức là $\Sigma = \begin{pmatrix} \sigma_1^2 & ... & 0 \\ ... & ... & ... \\ 0 & ... & \sigma_m^2 \end{pmatrix}$

=> Nếu các biến độc lập tức $\rho = 0$ thì $\rho = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii} * \sigma jj}}$

1.3 Tính chất phân bố chuẩn nhiều chiều:

1. $EX = \mu; var(X) = \Sigma$
2. Mọi tổ hợp tuyến tính $c'X = c_1X_1 + ... + c_mX_m$ có phân bố chuẩn $N(c'\mu, c'\Sigma c)$ với c là vector
3. Mọi tập con các thành phần của X có phân bố chuẩn
4. Các phân bố có điều kiện của 1 số thành phần X khi biết trước các thành phần khác cũng là phân bố chuẩn
5. Nếu Σ có dạng $\Sigma = diag(\sigma_{11}, \sigma_{22}, ..., \sigma_{mm})$ thì các thành phần $X_1, X_2, ..., X_m$ là độc lập
6. $\begin{cases} Y = \Sigma^{\frac{-1}{2}}(X-\mu) \sim N(0, I_m) \\ \sum_1^m Y_i^2 = Y'Y = [X-\mu]' \sum^{-1} [X-\mu] \sim \chi^2 \end{cases}$ với m bậc tự do

1.4 Ví dụ minh họa:

Cho X là vector ngẫu nhiên có phân phối chuẩn 3 chiều với $\mu = (1, 2, 3)$ và ma trận hiệp phương sai là ma trận đơn vị. Tìm tọa độ các điểm trong mặt mức với $c^2 = 9$?

Bài làm

- Đặt $X = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix}^T$
- Phương trình mặt mức là $(X-\mu)^T * \Sigma^{-1} * (X-\mu) = c^2$
  $\Leftrightarrow \begin{pmatrix} x_1 - 1 & x_2 - 2 & x_3 - 3 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} * \begin{pmatrix} x_1 - 1 \\ x_2 - 2 \\ x_3 - 3 \end{pmatrix} = 9$
  $\Leftrightarrow \begin{pmatrix} x_1 - 1 & x_2 - 2 & x_3 - 3 \end{pmatrix} * \begin{pmatrix} x_1 - 1 \\ x_2 - 2 \\ x_3 - 3 \end{pmatrix} = 9$
  $\Leftrightarrow (x_1 - 1)^2 + (x_2 - 2)^2 + (x_3 - 3)^2 = 9$
- Tọa độ các điểm trong mặt mức thỏa mãn:

$$(x_1 - 1)^2 + (x_2 - 2)^2 + (x_3 - 3)^2 < 9$$

2 Phân bố chuẩn hai chiều

$$T = \begin{bmatrix} X \\ Y \end{bmatrix} \sim N(\mu, \Sigma) \ với \ \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

2.1 Hàm mật độ:

$$f(t) = f(x, y) = \frac{1}{(2\pi)^{\frac{2}{2}}|\Sigma|^{\frac{1}{2}}} * exp(\frac{-1}{2}(T - \mu)^T * \Sigma^{-1} * (T - \mu))$$

2.2 Công thức chuyển đổi:

- $cov(T) = \Sigma$
- $DX = \sigma_1^2; DY = \sigma_2^2$
- $\rho = \rho(X, Y) = \sqrt{\lambda}e$
- $cov(X, Y) = \rho\sigma_1\sigma_2$

2.3 Chuyển 2 chiều về 1 chiều:

Lưu ý: Nếu $det|\Sigma| = 0$ thì đưa về dạng $Z_1 = aZ_2 + b$

Ta thay $\begin{cases} Z_1 \sim N(0, 1) \\ Z_2 \sim N(0, 1) \end{cases}$ vào phương trình ban đầu để tìm a và b

$=> Z_1 \sim kZ_2$

2.4 Ví dụ minh họa:

Cho X = $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$. Trong đó $\begin{cases} X_1 \sim N(0, 1) \\ X_2 \sim N(0, 1) \end{cases}$ $; \rho = cov(X_1, X_2)$

a) $\rho = 0$. Viết hàm mật độ của $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$?

b) $\rho = 1$. Viết hàm mật độ của $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$?

Bài làm

a) Viết hàm mật độ của X với $\rho = 0$

+ Cách 1: Sử dụng công thức

Ta có: $\begin{cases} \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} => |\sum| = 1 \\ => \sum^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{cases}$

Ta có công thức hàm mật độ:

$$\begin{aligned} f(x,y) &= \frac{1}{(2\pi)^{\frac{2}{2}}|\Sigma|^{\frac{1}{2}}} * e^{\frac{-1}{2}(x-\mu)' * \sum^{-1} * (x-\mu)} \\ &= \frac{1}{2\pi} * e^{\frac{-1}{2}\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}} \\ &= \frac{1}{2\pi} * e^{\frac{-1}{2}\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}} \\ &= \frac{1}{2\pi} * e^{\frac{-1}{2}(x^2+y^2)} \end{aligned}$$

+ Cách 2: Sử dụng tính chất hàm mật độ

Do $\rho = 0 \longrightarrow X_1, X_2$ độc lập

$$\begin{aligned} f(z_1, z_2) &= f(z_1) * f(z_2) \\ &= \frac{1}{2\pi} * e^{\frac{-z_1}{2}} * e^{\frac{-z_2}{2}} = \frac{1}{2\pi} * e^{-\frac{z_1^2+z_2^2}{2}} \end{aligned}$$

b) Viết hàm mật độ của X với $\rho = 1$

Do $\rho = 1 \longrightarrow X_1 = aX_2 + b$ nên $\begin{cases} X_1 \sim N(0,1) \\ X_2 \sim N(0,1) \end{cases} => a = 1, b = 0$

$=> X_1 = X_2 = X \sim N(0,1)$

3 Đặc trưng của mẫu ngẫu nhiên

3.1 Vector trung bình mẫu:

$$\overline{X'} = [\overline{X_1}, \overline{X_2}, ..., \overline{X_n}]$$

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

3.2 Ma trận hiệp phương sai mẫu:

$$S_n = \begin{bmatrix} S_{11} & ... & S_{1n} \\ ... & ... & ... \\ S_{n1} & ... & S_{nn} \end{bmatrix}$$

trong đó: S_{ij} là ma trận hiệp phương sai mẫu

$$\begin{aligned} S_{ij} &= \tfrac{1}{N}\sum_{k=1}^{N}(x_{ki} - \overline{X_i}) * (x_{kj} - \overline{X_j}) \\ &= \tfrac{1}{N-1}(\sum_{k=1}^{N} x_{ki}x_{kj} - n\overline{X_i X_j}) \end{aligned}$$

3.3 Phân bố mẫu ngẫu nhiên nhiều chiều:

$$X \sim N(\mu, \Sigma) \textit{ hoặc } \overline{X} \sim N(\mu, \frac{1}{n}\Sigma)$$

- S = $[\overline{\sigma_{ij}}]$ với $\overline{\sigma_{ij}} = \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \overline{X_i}) * (X_{kj} - \overline{X_k})$

  => $\overline{X}$ và S độc lập với nhau

4 Phân tích thành phần chính của X

Có $X = (X_1, X_2)$; $\Sigma = \begin{pmatrix} \sigma_{11} & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_{22} \end{pmatrix}$ và ma trận tương quan tương ứng $\begin{bmatrix} ... & ... \\ ... & ... \end{bmatrix}$

4.1 Tính chất thành phần chính:

- Ta có $\sum_{i=1}^{n} D(Y_i) = \lambda_1 + ... + \lambda_k$
- Ta có $\frac{\lambda_i}{\lambda_1+...+\lambda_k}$ là tỷ lệ biến sai tổng cộng của các thành phần của k do thành phần chính thứ i gây ra

4.2 Dựa trên ma trận hiệp phương sai:

1. Xét đa thức đặc trưng sau $\longrightarrow det(\Sigma - I_n\lambda) = 0$
2. Tìm giá trị riêng từ lớn đến nhỏ $\begin{cases} \lambda_1 = ... \\ \lambda_2 = ... \\ \lambda_3 = ... \end{cases}$ với $(\lambda_1 > \lambda_2 > \lambda_3)$
3. Tìm vector riêng ứng với mỗi giá trị riêng e_1, e_2, e_3
4. Phân tích thành phần chính $\begin{cases} Y_1 = e_{11}X_1 + e_{21}X_2 \longrightarrow DY_1 = \lambda_1 \\ Y_2 = e_{21}X_1 + e_{22}X_2 \longrightarrow DY_2 = \lambda_2 \\ ... \end{cases}$

4.3 Dựa trên ma trận tương quan:

Thành phần chính của biến chuẩn hóa X dựa trên ma trận tương quan $=> Z = (Z_1, Z_2)$

$$\lambda_1 \longrightarrow e_1 = \begin{pmatrix} a_1 \\ b_1 \end{pmatrix} \text{ và } \lambda_2 \longrightarrow e_2 = \begin{pmatrix} a_2 \\ b_2 \end{pmatrix}$$

=> Phân tích thành phần chính: $\begin{cases} Y_1 = e_1^T Z = a_1 Z_1 + b_1 Z_2 \\ Y_2 = e_2^T Z = a_2 Z_1 + b_2 Z_2 \end{cases}$

4.4 Ví dụ minh họa:

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

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

Tìm các thành phần chính của X?

Bài làm

1. Xét đa thức đặc trưng:

$$det|\Sigma - I_n\lambda| = \begin{vmatrix} 1-\lambda & -2 & 0 \\ -2 & 5-\lambda & 0 \\ 0 & 0 & 2-\lambda \end{vmatrix} = 0$$

$\Leftrightarrow (1-\lambda)(5-\lambda)(2-\lambda) - 4(2-\lambda) = 0$
$\Leftrightarrow (2-\lambda)[(\lambda-1)(\lambda-5) - 4] = 0$
$\Leftrightarrow \left[\begin{array}{l} \lambda = 2 \\ \lambda^2 - 6\lambda + 1 = 0 \end{array}\right. \Leftrightarrow \left[\begin{array}{l} \lambda = 2 \\ \lambda^2 - 2*\lambda*3 + 9 - 9 + 1 = 0 \end{array}\right.$
$\Leftrightarrow \left[\begin{array}{l} \lambda = 2 \\ (\lambda-3)^2 = 18 \end{array}\right. \Leftrightarrow \left[\begin{array}{l} \lambda = 2 \\ \lambda = \pm\sqrt{8} + 3 \end{array}\right.$
$\Leftrightarrow \left[\begin{array}{ll} \lambda_1 = 3 + 2\sqrt{2} & \longrightarrow e_1 = ... \\ \lambda_2 = 2 & \longrightarrow e_2 = ... \\ \lambda_3 = 3 - 2\sqrt{2} & \longrightarrow e_3 = ... \end{array}\right.$
Lưu ý: Giá trị riêng được xếp theo thứ tự giảm dần

2. Tìm vector riêng tương ứng với từng giá trị riêng:

$$\text{Ta giải } (\Sigma - \lambda_i I) * e_i = 0$$

- Tìm e_1 với $\lambda_1 = 3 + 2\sqrt{2}$

$$\left(\begin{array}{ccc|c} -2-2\sqrt{2} & -2 & 0 & 0 \\ -2 & 2-2\sqrt{2} & 0 & 0 \\ 0 & 0 & -1-2\sqrt{2} & 0 \end{array}\right)$$

$$\Leftrightarrow \begin{cases} (-2-2\sqrt{2})x_1 - 2x_2 = 0 \\ -2x_1 + (2-2\sqrt{2})x_2 = 0 \\ (-1-2\sqrt{2})x_3 = 0 \end{cases}$$

$$=> e_1 = \begin{pmatrix} (1-\sqrt{2})x_2 \\ x_2 \\ 0 \end{pmatrix} = \begin{pmatrix} 1-\sqrt{2} \\ 1 \\ 0 \end{pmatrix}$$

- Tìm e_2 với $\lambda_2 = 2$

$$\left(\begin{array}{ccc|c} -1 & -2 & 0 & 0 \\ -2 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

$$\Leftrightarrow \begin{cases} -x_1 - 2x_2 = 0 \\ -2x_1 + 3x_2 = 0 \\ x_3 \textit{ tự do} \end{cases} \quad => e_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

- Tìm e_3 với $\lambda_3 = 3 + 2\sqrt{2}$

$$\left(\begin{array}{ccc|c} -2+2\sqrt{2} & -2 & 0 & 0 \\ -2 & 2+2\sqrt{2} & 0 & 0 \\ 0 & 0 & -1+2\sqrt{2} & 0 \end{array}\right)$$

$$\Leftrightarrow \begin{cases} (-2+2\sqrt{2})x_1 - 2x_2 = 0 \\ -2x_1 + (2+2\sqrt{2})x_2 = 0 \\ (-1+2\sqrt{2})x_3 = 0 \end{cases}$$

$$=> e_3 = \begin{pmatrix} (1+\sqrt{2})x_2 \\ x_2 \\ 0 \end{pmatrix} = \begin{pmatrix} 1+\sqrt{2} \\ 1 \\ 0 \end{pmatrix}$$

3. Phân tích thành phần chính

$$=> \begin{cases} Y_1 = (1-\sqrt{2})X_1 + X_2 => DY_1 = 3 + 2\sqrt{2} \\ Y_2 = X_3 \qquad\qquad\qquad\quad => DY_2 = \lambda_2 = 2 \\ Y_3 = (1+\sqrt{2})X_1 + X_2 => DY_3 = \lambda_3 = 3 - 2\sqrt{2} \end{cases}$$

5 Phân tích nhân tố về dạng $\Sigma = LL' + \psi$

$$\Sigma = LL' + \psi = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + ... + \lambda_k e_k e_k'$$

5.1 Phân tích nhân tố:

Biểu diễn X_i theo F' = $(F_1, F_2, ..., F_m)$ với $m \leq k; i \in \overline{1,k}$
=> Cần tím giá trị m phù hợp

Mô hình nhân tố dạng ma trận:

$$X - \mu = L * F + \epsilon = \begin{cases} X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + ... + l_{1m}F_m + \epsilon_1 \\ X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + ... + l_{2m}F_m + \epsilon_2 \\ ... \\ X_k - \mu_k = l_{k1}F_1 + l_{k2}F_2 + ... + l_{km}F_m + \epsilon_k \end{cases}$$

với các $F_1, F_2, ..., F_m$ là các nhân tố chung

- $(X - \mu)$ cỡ k*1
- L cỡ k*n
- F cỡ m*1
- ϵ cỡ k*1

5.2 Phương pháp giải:

5.2.1 TH1: $m = k => \psi = 0$

$$\Sigma = LL'$$

$$\text{với} \begin{cases} L = [\sqrt{\lambda_1}e_1, \sqrt{\lambda_2}e_2, ..., \sqrt{\lambda_k}e_k] \\ L' = [\sqrt{\lambda_1}e_1, \sqrt{\lambda_2}e_2, ..., \sqrt{\lambda_k}e_k]' \end{cases}$$

5.2.2 TH2: $m < k$

$$\Sigma = LL' + \psi = LL' + diag(\psi_1, \psi_2, ..., \psi_k)$$

$$\begin{aligned} \Sigma &= LL' + \psi \\ &= [\sqrt{\lambda_1}e_1, \sqrt{\lambda_2}e_2, ..., \sqrt{\lambda_k}e_k] * [\sqrt{\lambda_1}e_1, \sqrt{\lambda_2}e_2, ..., \sqrt{\lambda_k}e_k]' + \psi \end{aligned}$$

biết rằng: $\psi = \lambda_{m+1}e_{m+1}e_{m+1}' + ... + \lambda_k e_k e_k'$

5.3 Ví dụ minh họa:

5.3.1 Ví dụ 1:

Phân tích nhân tố $\Sigma = LL' + \psi$ thành 2 thành phần chính với ma trận hiệp phương sai:

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

Bài làm

- Tính toán $\left[\begin{array}{lll} \lambda_1 = 116.269 & => & e_1 = (0.324; 0.607; 0.648; 1) \\ \lambda_2 = 61.416 & => & e_2 = (-1.35; -2.401; 1.379; 1) \\ \lambda_3 = 2.517 & => & e_3 = (5.844; -3.534; -1.156; 1) \\ \lambda_4 = 1.798 & => & e_4 = (-0.455; -0.055; -1.266; 1) \end{array}\right.$

- Phân tích thành 2 TPC $\begin{cases} L = [\sqrt{116.269}e_1; \sqrt{61.416}e_2] \\ L' = [\sqrt{116.269}e_1; \sqrt{61.416}e_2]' \\ \psi = 2.517e_3e_3' + 1.798e_4e_4' \end{cases}$

5.3.2 Ví dụ 2:

Phân tích nhân tố với tất cả thành phần chính sau:

$$\begin{bmatrix} 1 & 0.2 & 0.96 & 0.32 & 0.01 \\ ... & 1 & 0.13 & 0.71 & 0.85 \\ ... & ... & 1 & 0.5 & 0.11 \\ ... & ... & ... & 1 & 0.79 \\ ... & ... & ... & ... & 1 \end{bmatrix}$$

Bài làm

- Tính $\left[\begin{array}{lll} \lambda_1 = 2.853 & => & e_1 = (0.701; 0.974; 0.808; 1.177; 1) \\ \lambda_2 = 1.806 & => & e_2 = (-1.502; 0.965; -1.377; 0.193; 1) \\ \lambda_3 = 0.204 & => & e_3 = (-0.477; -3.367; -0.769; 2.728; 1) \\ \lambda_4 = 0.102 & => & e_4 = (0.185; -0.375; 0.156; -0.756; 1) \\ \lambda_5 = 0.034 & => & v_5 = (77.867; 7.953; -78.636; 0.184; 1) \end{array}\right.$

- Ta có Y = $\begin{cases} 0.701X_1 + 0.974X_2 + 0.808X_3 + 1.177X_4 + X_5 \\ -1.502X_1 + 0.965X_2 - 1.377X_3 + 0.193X_4 + X_5 \\ -0.477X_1 - 3.356X_2 - 0.764X_3 + 2.728X_4 + X_5 \\ 0.185X_1 - 0.374X_2 + 0.156X_3 - 0.756X_4 + X_5 \\ 77.867X_1 + 7.953X_2 - 78.636X_3 + 0.184X_4 + X_5 \end{cases}$

- Phân tích $\begin{cases} L = [\sqrt{2.853}e_1; \sqrt{1.806}e_2; \sqrt{0.204}e_3; \sqrt{0.102}e_4; \sqrt{0.034}e_5] \\ L' = [\sqrt{2.853}e_1; \sqrt{1.806}e_2; \sqrt{0.204}e_3; \sqrt{0.102}e_4; \sqrt{0.034}e_5]' \\ \psi = 0 \end{cases}$

  => Σ = L L'

6 Mô hình hồi quy tuyến tính bội

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \epsilon$$

- X $= \begin{pmatrix} X_1 \\ X_2 \\ ... \\ X_k \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & ... & x_{k1} \\ ... & ... & ... & ... & ... \\ 1 & x_{1n} & x_{2n} & ... & x_{kn} \end{pmatrix}$

- $Y = X\beta + \epsilon = \begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + ... + \beta_k x_{k1} + \epsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_k x_{k2} + \epsilon_2 \\ ... \\ y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + ... + \beta_k x_{kn} + \epsilon_n \end{cases}$

6.1 Ước lượng mô hình:

1. $\hat{\beta} = (X^T X)^{-1} X^T Y$
2. Phần dư của HQTT $\longrightarrow \hat{\epsilon_j} = y_j - \hat{y_j} = y_j - \hat{\beta}_0 - \hat{\beta}_1 x_{j1} - ... - \hat{\beta}_k x_{jk}$
3. $cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = D(\beta)$
4. $\hat{\sigma}^2 = \sum_1^N \frac{\epsilon_j^2}{N-k-1}$
5. Hệ số xác định R:

$$R^2 = \frac{Y\ dự\ kiến}{Y\ ban\ đầu} = \frac{\sum_1^N \hat{y_j}^2 - N(\overline{y}^2)}{\sum_1^N y_j^2 - N(\overline{y}^2)} \in [0, 1]$$

6.2 Kiểm định mô hình có ý nghĩa thống kê:

Kiểm định $\begin{cases} H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \\ H_1 : ngược\ lại \end{cases}$

$$S = \{F \geq F_{k,N-k-1}(\alpha)\}$$

1. Xét F, có $\begin{cases} F = \frac{R^2(N-k-1)}{k(1-R^2)} \\ F_{k,N-k-1}(\alpha) \end{cases}$

2. Nếu $F \geq F_{k,N-k-1}(\alpha)$ thì ta bác bỏ H_0

=> Vậy mô hình có ý nghĩa thống kê

6.3 Kiểm định các hệ số:

Kiểm tra $\begin{cases} \beta_0 = 0 \\ \beta_0 \neq 0 \end{cases}$ tương tự với β_i

1. Miền bác bỏ $S = \{\frac{|\hat{\beta}_0|}{S_{\hat{\beta}_0}} \geq t_{n-k-1}(\frac{\alpha}{2})\}$

2. Trong đó $\begin{cases} S_{\hat{\beta}_0} = \sqrt{D(\beta_0)} \\ T = \frac{\hat{\beta}_0}{S_{\hat{\beta}_0}} \end{cases}$

6.4 Ví dụ minh họa:

Cho bộ dữ liệu sau:

x_1	x_2	y
10	2	15
5	3	9
7	3	3
19	6	25
11	7	7
8	9	13

$$=> X = \begin{pmatrix} 1 & 10 & 2 \\ 1 & 5 & 3 \\ 1 & 7 & 3 \\ 1 & 19 & 6 \\ 1 & 11 & 7 \\ 1 & 8 & 9 \end{pmatrix} ; Y = \begin{pmatrix} 15 \\ 9 \\ 3 \\ 25 \\ 7 \\ 13 \end{pmatrix}$$

a) Sử dụng mô hình HQTT cổ điển để ước lượng các hệ số sau đây của mô hình: $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}, R^2, cov(\hat{\beta})$
b) Kiểm tra mô hình có ý nghĩa thống kê không? ($\alpha = 5\%$)

Bài làm

a) Sử dụng mô hình HQTT để ước lượng các hệ số:

- $\hat{\beta} = (X^T X)^{-1} X^T Y => \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} -0.46487259 \\ 1.2760181 \\ -0.05906168 \end{pmatrix}$

- $\epsilon = Y - \hat{Y} = Y - X\hat{\beta} = (2.823; 3.262; -5.29; 1.575; -6.15; 3.78)^T$
- $\hat{\sigma} = \sqrt{\frac{\Sigma\epsilon_i^2}{n-k-1}} = 5.81218851$ với n = 6; k = 2
- $R^2 = \dfrac{\displaystyle\sum_1^N \hat{y_j}^2 - N(\overline{y}^2)}{\displaystyle\sum_1^N y_j^2 - N(\overline{y}^2)} = 0.6552904$
- $cov(\hat{\beta}) = \sigma^2(X^TX)^{-1} = \begin{bmatrix} 45.0514 & -2.2928 & -3.298 \\ -2.2928 & 0.305 & -0.152 \\ -3.2985 & -0.1528 & 0.965 \end{bmatrix}$

b) Kiểm tra mô hình có ý nghĩa thống kê không?

- Ta có $\begin{cases} H_0 : \beta_1 = \beta_2 = 0 \\ H_1 : \ \textit{ngược lại} \end{cases}$
- Miền bác bỏ $S = \{F \geq F_{2,n-3}\}$ => $\begin{cases} F = \frac{R^2(N-k-1)}{k(1-R^2)} = 2.851 \\ F_{2,3} = 9.507 \end{cases}$

  => Do $F_{2,3} > F$ => Chấp nhận giải thiết H_0

=> Vậy mô hình có ý nghĩa thống kê

7 Mô hình hồi quy tuyến tính nhiều biến phụ thuộc

7.1 Mô hình hồi quy:

Cho $\begin{cases} Y_1 = \beta_{01} + \beta_{11}X_1 + ... + \beta_{k1}X_k + \epsilon_1 \\ Y_2 = \beta_{02} + \beta_{12}X_1 + ... + \beta_{k2}X_k + \epsilon_2 \\ ... \\ Y_m = \beta_{0m} + \beta_{1m}X_1 + ... + \beta_{km}X_k + \epsilon_m \end{cases}$ với $\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_m \end{bmatrix}$

trong đó: $X = \begin{pmatrix} 1 & x_{11} & ... & x_{1k} \\ 1 & x_{21} & ... & x_{2k} \\ ... \\ 1 & x_{N1} & ... & x_{Nk} \end{pmatrix}$; $Y = \begin{pmatrix} y_{11} & y_{12} & ... & y_{1m} \\ y_{21} & y_{22} & ... & y_{2m} \\ ... \\ y_{N1} & y_{N2} & ... & y_{Nm} \end{pmatrix}$

Lưu ý: $E(\epsilon) = 0; cov(\epsilon) = \Sigma$

7.2 Ước lượng hệ số bình phương tối thiểu:

$$Y = X\beta + \epsilon$$

- $\beta = (X^TX)^{-1}X^TY$
- $\beta = (\beta_1, \beta_2, ..., \beta_m) = \begin{pmatrix} \beta_{01} & \beta_{02} & ... & \beta_{0m} \\ \beta_{11} & \beta_{12} & ... & \beta_{1m} \\ ... \\ \beta_{k1} & \beta_{k2} & ... & \beta_{km} \end{pmatrix}$

7.3 Ví dụ minh họa:

Xét hệ 2 mô hình $\begin{cases} y_{j1} = \beta_{01} + \beta_{11}x_{j1} + \epsilon_{j1} \\ y_{j2} = \beta_{02} + \beta_{12}x_{j1} + \epsilon_{j2} \end{cases}$ với $j = \overline{1,5}$

có bảng số liệu sau:

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?

Bài làm

- Ước lượng hệ số $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 & \hat{\beta}_2 \end{pmatrix}$:

$$\hat{\beta}_1 = \begin{pmatrix} \hat{\beta_{01}} \\ \hat{\beta_{11}} \\ ... \\ \hat{\beta_{n1}} \end{pmatrix} = (X^T X)^{-1} X^T \begin{pmatrix} y_{11} \\ y_{12} \\ ... \\ y_{1n} \end{pmatrix}$$

$$\hat{\beta}_2 = \begin{pmatrix} \hat{\beta_{02}} \\ \hat{\beta_{12}} \\ ... \\ \hat{\beta_{n2}} \end{pmatrix} = (X^T X)^{-1} X^T \begin{pmatrix} y_{21} \\ y_{22} \\ ... \\ y_{2n} \end{pmatrix}$$

- Có: $X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} => \begin{cases} X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} 5 & 10 \\ 10 & 30 \end{pmatrix} \\ (X^T X)^{-1} = \begin{pmatrix} 0.6 & -0.2 \\ -0.2 & 0.1 \end{pmatrix}_{2*2} \end{cases}$

$=> \hat{\beta} = (X^T X)^{-1} X^T \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}^T = \begin{pmatrix} \hat{\beta_{01}} & \hat{\beta_{02}} \\ \hat{\beta_{11}} & \hat{\beta_{12}} \end{pmatrix}$

$=> \begin{pmatrix} 0.6 & 0.4 & 0.2 & 0 & -0.2 \\ -0.2 & -0.1 & 0 & 0.1 & 0.2 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 4 & -1 \\ 3 & 2 \\ 8 & 3 \\ 9 & 2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix}$

- Ước lượng phần dư $\hat{\epsilon}$:

$$\hat{\epsilon} = \begin{pmatrix} \hat{\epsilon_{11}} & \hat{\epsilon_{21}} \\ \hat{\epsilon_{12}} & \hat{\epsilon_{22}} \\ \hat{\epsilon_{13}} & \hat{\epsilon_{23}} \\ \hat{\epsilon_{14}} & \hat{\epsilon_{24}} \\ \hat{\epsilon_{15}} & \hat{\epsilon_{25}} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 4 & -1 \\ 3 & 2 \\ 8 & -3 \\ 9 & 2 \end{pmatrix} - \begin{pmatrix} 1 & -1 \\ 3 & 0 \\ 5 & 1 \\ 7 & 2 \\ 9 & 3 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & -1 \\ -2 & 1 \\ 1 & -5 \\ 0 & -1 \end{pmatrix}$$

8 Hồi quy theo các biến ngẫu nhiên

8.1 Dự báo tuyến tính Y theo X_i:

Mô hình tuyến tính:

$$\tilde{Y} = b_0 + b_1X_1 + ... + b_kX_k = b_0 + b'X$$

với $i \in \{1, 2, ..., k\}$; $\mu = \begin{bmatrix} \mu_y \\ ... \\ \mu_X \end{bmatrix} = \begin{bmatrix} \mu_Y \\ \mu_{X_1} \\ \mu_{X_2} \\ ... \end{bmatrix}$; $\Sigma = \begin{bmatrix} \Sigma_{YY} & ... & \Sigma_{YX} \\ ... & ... & ... \\ \Sigma_{XY} & ... & \Sigma_{XX} \end{bmatrix}$

=> Nghiệm bài toán $\begin{cases} b = \beta = \sum_{XX}^{-1} \sum_{XY} \\ b_0 = \beta_0 = \mu_Y - \beta' \mu_X \end{cases}$

8.2 Sai số bình phương cực tiểu:

$$E(Y - \beta_0 - \beta' X)^2 = \sum_{YY} - \sum_{XY}' \sum_{XX}^{-1} \sum_{YY} = \sum_{YY} - \sum_{XY}' \beta$$

8.3 Hệ số tương quan bội giữa Y và X:

$$\rho_{Y/X} = \sqrt{\frac{\sum_{XY}' \sum_{XX}^{-1} \sum_{XY}}{\sum_{YY}}}$$

8.4 Ví dụ minh họa:

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

Xác định phương trình hồi quy tuyến tính của Y theo X_1, X_2 và sai số bình phương trung bình $E(\hat{\epsilon})^2$ và hệ số tương quan tuyến tính bội $\rho_{Y/X}$?

Bài làm

Lưu ý: Ma trận hiệp phương sai Σ luôn là ma trận đối xứng

Ta có: $\Sigma = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix} = \begin{bmatrix} 10 & 1 & -1 \\ 1 & 7 & 3 \\ -1 & 3 & 2 \end{bmatrix} = \left[\begin{array}{c|cc} \Sigma_{YY} & 1 & 1 \\ \hline 1 & \Sigma_{X_1X_1} & \Sigma_{X_1X_2} \\ \text{-1} & \Sigma_{X_2X_1} & \Sigma_{X_2X_2} \end{array}\right]$

$$\left[\begin{array}{c|cc} 10 & 1 & -1 \\ \hline 1 & 7 & 3 \\ -1 & 3 & 2 \end{array}\right]$$

Xác định phần tử trong ma trận hiệp phương sai

1. Tính $\hat{\beta} = \Sigma_{XX}^{-1}\Sigma_{XY} => \begin{cases} \Sigma_{XX} = \begin{bmatrix} 7 & 3 \\ 3 & 2 \end{bmatrix} => \Sigma_{XX}^{-1} \begin{bmatrix} 2 & -3 \\ -3 & 7 \end{bmatrix} \\ \Sigma_{XY} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \end{cases}$

   $=> \beta = \frac{1}{5}\begin{bmatrix} 2 & -3 \\ -3 & 7 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{6}{5} \\ \frac{-10}{5} \end{bmatrix} = \begin{bmatrix} 1.2 \\ -2 \end{bmatrix}$

2. $\hat{\beta}_0 = EY - \hat{\beta}^T EX = 5 - \begin{bmatrix} 1.2 & -2 \end{bmatrix}\begin{bmatrix} 2 \\ 0 \end{bmatrix} = 2.6$

3. $\hat{Y} = 2.6 + 1.2X_1 - 2X_2$

4. $\hat{\epsilon} = Y - \hat{Y} => E(\hat{\epsilon})^2 = E(Y - 2.6 - 1.2X_1 + 2X_2)^2$

5. $\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} = \frac{1}{5}\begin{bmatrix} 1 & -1 \end{bmatrix}\begin{bmatrix} 2 & -3 \\ -3 & 7 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = 3.2$

6. $\rho = (\frac{\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}}{\Sigma_{YY}})^{\frac{1}{2}} = \sqrt{\frac{3.2}{10}}$

9 Phân biệt lớp bằng quy tắc Bayes chấp nhận được

Dựa trên quan sát p dấu hiệu U = $(U_1, U_2, ..., U_p)$ của 1 đối tượng cá thể, bài toán phân biệt là xác định xem đối tượng cá thể đó thuộc 1 trong K nhóm đã xác định.

Cho $U = (U_1, U_2, ..., U_p)$. Hãy phân loại U thuộc nhóm nào?

Lưu ý: $\begin{cases} \textit{U là số đo p dấu hiệu} \\ \textit{S là tập giá trị của U} \end{cases}$

9.1 Quy tắc phân biệt Gauss thuộc nhóm i:

$$\overline{S_i}(U) = \hat{\mu_i}'A^{-1}U - \frac{1}{2}\hat{\mu_i}'A^{-1}\hat{\mu_i} + ln\hat{\pi_i}$$

với $i = \overline{1,n}; \pi_i = \frac{n_i}{n}$

Lưu ý: Các nhóm i luôn dùng tới A là ma trận hiệp phương sai

=> Nếu $S_i = max\{\overline{S_1(U)}, \overline{S_2(U)}, ...\}$ thì U thuộc lớp S_i (i)

9.2 Ví dụ minh họa:

Cho ma trận $A = \begin{pmatrix} 2.3008 & 0.2516 & 0.4742 \\ 0.2516 & 0.6075 & 0.0358 \\ 0.4742 & 0.0358 & 0.5951 \end{pmatrix}$ và có ma trận nghịch đảo $A^{-1} = \begin{pmatrix} 0.5432 & -0.2002 & -0.4208 \\ -0.2002 & 1.7258 & 0,0558 \\ -0.4208 & 0.0558 & 2.0123 \end{pmatrix}$ với n = 256

Hãy phân loại $U = (0.8201; 1.6; 0.68)$ thuộc lớp nào sau đây:

- $\hat{\mu_1} = (2.9298; 1.667; 0.7281)$ có $n_1 = 114$
- $\hat{\mu_2} = (3.0303; 1.2424; 0.5455)$ có $n_2 = 33$
- $\hat{\mu_3} = (3.8125; 1.8438; 0.8125)$ có $n_3 = 32$
- $\hat{\mu_4} = (4.7059; 1.5882; 1.1176)$ có $n_4 = 17$
- $\hat{\mu_5} = (1.4; 0.2; 0)$ có $n_5 = 5$
- $\hat{\mu_6} = (0.6; 0.1455; 0.2182)$ có $n_6 = 44$

Bài làm

1. $\hat{\mu_1} = (2.9298; 1.667; 0.7281)$ có $n_1 = 114$
   $n_1 = 114; n = 256 => \hat{\pi}_1 = \frac{n_1}{n} = \frac{114}{256}$
   $=> \overline{S_1}(U) = \hat{\mu_1}^T A^{-1} U - \frac{1}{2}\hat{\mu_1}^T A^{-1} \hat{\mu_1} + ln\hat{\pi}_1 = 0.4671$

2. $\hat{\mu_2} = (3.0303; 1.2424; 0.5455)$ có $n_2 = 33$
   $n_2 = 33; n = 256 => \hat{\pi}_2 = \frac{n_2}{n} = \frac{33}{256}$
   $=> \overline{S_2}(U) = \hat{\mu_2}^T A^{-1} U - \frac{1}{2}\hat{\mu_2}^T A^{-1} \hat{\mu_2} + ln\hat{\pi}_2 = -1.3699$

3. $\hat{\mu_3} = (3.8125; 1.8438; 0.8125)$ có $n_3 = 32$
   $n_3 = 32; n = 256 => \hat{\pi}_3 = \frac{n_3}{n} = \frac{32}{256}$
   $=> \overline{S_3}(U) = \hat{\mu_3}^T A^{-1} U - \frac{1}{2}\hat{\mu_3}^T A^{-1} \hat{\mu_3} + ln\hat{\pi}_3 = -1.849043$

4. $\hat{\mu_4} = (4.7059; 1.5882; 1.1176)$ có $n_4 = 17$
   $n_4 = 17; n = 256 => \hat{\pi}_4 = \frac{n_4}{n} = \frac{17}{256}$
   $=> \overline{S_4}(U) = \hat{\mu_4}^T A^{-1} U - \frac{1}{2}\hat{\mu_4}^T A^{-1} \hat{\mu_4} + ln\hat{\pi}_4 = -3.879$

5. $\hat{\mu_5} = (1.4; 0.2; 0)$ có $n_5 = 5$
   $n_5 = 5; n = 256 => \hat{\pi}_5 = \frac{n_5}{n} = \frac{5}{256}$
   $=> \overline{S_5}(U) = \hat{\mu_5}^T A^{-1} U - \frac{1}{2}\hat{\mu_5}^T A^{-1} \hat{\mu_5} + ln\hat{\pi}_5 = -4.1440$

6. $\hat{\mu_6} = (0.6; 0.1455; 0.2182)$ có $n_6 = 55$
   $n_6 = 55; n = 256 => \hat{\pi}_6 = \frac{n_6}{n} = \frac{55}{256}$
   $=> \overline{S_6}(U) = \hat{\mu_6}^T A^{-1} U - \frac{1}{2}\hat{\mu_6}^T A^{-1} \hat{\mu_6} + ln\hat{\pi}_6 = -1.1016$

Ta có: $S = max\{S_1, S_2, S_3, S_4, S_5, S_6\} = S_1$
=> Vậy U thuộc lớp S_1

10 Các phương pháp phân cụm trong bài toán phân lớp

10.1 Khoảng các giữa 2 phần tử:

Có N đối tượng, phân chia thành các nhóm khác nhau (phân cụm) biết rằng tọa độ vector $\begin{cases} x = (x_1, x_2, ..., x_n) \\ y = (y_1, y_2, ..., y_n) \end{cases}$

10.1.1 Khoảng cách Euclide:

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

10.1.2 Khoảng cách thống kê:

Cho A là ma trận đối xứng xác định dương, ta có:

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

10.1.3 Khoảng cách Minkowski:

$$d_3(x.y) = (\sum_{i=1}^{k} |x_i - y_i|^m)^{\frac{1}{m}} \; với \; m = \overline{1, 2, ...}$$

10.1.4 Khoảng cách Canberra:

$$d_4(x, y) = \sum_{i=1}^{k} \frac{|x_i - y_i|}{(x_i + y_i)} \; với \; (x_i, y_i > 0)$$

10.1.5 Hệ số Czekanowski:

$$d_5(x, y) = 1 - \frac{2\sum_{i=1}^{k} min(x_i, y_i)}{\sum_{i=1}^{k} (x_i + y_i)} \; \textbf{với} \; (x_i, y_i > 0)$$

10.2 Phân cụm bằng phương pháp kết nối đơn:

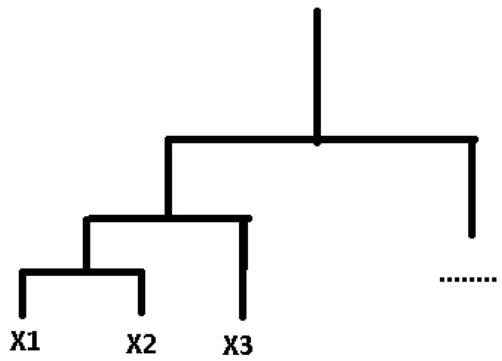
10.2.1 Ma trận khoảng cách đối xứng:

$D = [d_{ik}] \longrightarrow$ phân cụm cho N dữ liệu sau đây:

	X_1	X_2	X_3	...	X_n
X_1	0	...	...	...	...
X_2	...	0	...	...	...
X_3	...	...	0	...	...
...	...	...	...	...	...
X_n	...	...	...	...	0

10.2.2 Tìm cặp có khoảng cách bé nhất:

$$d((X_i, X_j), X_k) = min(d(X_i, X_k), d(X_j, X_k))$$



Biết cách phân K cụm bằng

10.2.3 Ví dụ phương pháp kết nối đơn:

Xét ma trận khoảng cách sau:

$D = [d_{ik}] =>$

	1	2	3	4	5
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

Hãy phân cụm cho 5 cá thể trên?

Bài làm

- Cặp có khoảng cách bé nhất là (3, 5) có d(3, 5) = 2

	(3, 5)	1	2	4
(3, 5)	0			
1	3	0		
2	7	9	0	
4	8	6	5	0

  - d((3, 5), 1) = min{d(3, 1), d(5, 1)} = 3
  - d((3, 5), 2) = min{d(3, 2), d(5, 2)} = 7
  - d((3, 5), 4) = min{d(3, 4), d(5, 4)} = 8
  - d(1, 2) = 9
  - d(1, 4) = 6
  - d(2, 4) = 5

- Cặp có khoảng cách bé nhất là (3, 5) và 1 có d((3, 5), 1) = 3

	(3, 5, 1)	2	4
(3, 5, 1)	0		
2	7	0	
4	6	5	0

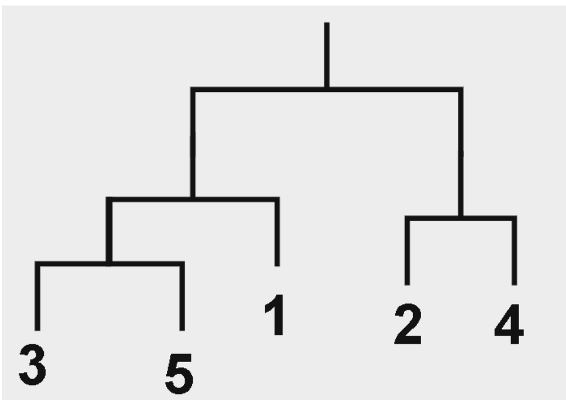
  - d((3, 5, 1), 2) = min{d(3, 2), d(5, 2), d(1, 2)} = 7
  - d((3, 5, 1), 4) = min{d(3, 4), d(5, 4), d(1, 4)} = 6
  - d(4, 2) = 5

- Cặp có khoảng cách bé nhất là (4, 2) có d(4, 2) = 5

	(3, 5, 1)	(2, 4)
(3, 5, 1)	0	
(2, 4)	6	0

  - d((3, 5, 1), (2, 4)) = 6

=> Thuật toán phân cụm dừng



Phân cụm bằng phương pháp kết nối đơn

10.3 Phân cụm bằng phương pháp K-means:

10.3.1 Triển khai thuật toán:

Phân chia làm K cụm khác nhau:

	x	y
A	x_A	y_A
B	x_B	y_B
C	x_C	y_C
D	x_D	y_D

Phương pháp giải:

1. Chọn K cụm bất kỳ nhóm lại với nhau
2. Tính toán tâm từng cụm
3. Tìm tất cả khoảng cách từ các điểm với tâm mỗi cụm (điểm nào gần tâm cụm hơn thì cho vào cụm đó)
4. Thuật toán dừng khi tâm cụm không thay đổi

10.3.2 Ví dụ minh họa:

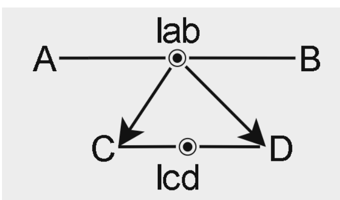
Ta có bảng số liệu sau:

	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Phân chia các đối tượng vào k = 2 cụm mà thỏa mãn mỗi đối tượng gồm tâm của cụm chứa nó nhất?

Bài làm

Ta chọn 2 cụm bất kỳ AB và CD:



Chọn k = 2 cụm bất kỳ

1. Lặp 1 gồm 2 cụm (A, B) và (C, D):

- $$\begin{cases} d^2(C, AB) = (x_C - x_{AB})^2 + (y_C - y_{AB})^2 = 17 \\ d^2(D, AB) = (x_D - x_{AB})^2 + (y_D - y_{AB})^2 = 41 \\ d^2(A, AB) = d^2(B, AB) = 10 \end{cases}$$
- $$\begin{cases} d^2(C, CD) = d^2(D, CD) = 4 \\ d^2(A, CD) = (x_A - x_{CD})^2 + (y_A - y_{CD})^2 = 61 \\ d^2(B, CD) = (x_B - x_{CD})^2 + (y_B - y_{CD})^2 = 9 \end{cases}$$

$$=> \begin{cases} A \textit{ thuộc cụm } (A, B) \\ B \textit{ thuộc cụm } (C, D) \\ C \textit{ thuộc cụm } (C, D) \\ D \textit{ thuộc cụm } (C, D) \end{cases}$$

2. Lặp 2 gồm 2 cụm mới (A) và (BCD):

	x_1	x_2
A	5	3
BCD	-1	-1

$$=> (I_{BCD} = \frac{B + C + D}{3})$$

- $$\begin{cases} d^2(A, (BCD)) = (5 + 1)^2 + (3 + 1)^2 = 52 \\ d^2(B, (BCD)) = (-1 + 1)^2 + (1 + 1)^2 = 4 \\ d^2(C, (BCD)) = (1 + 1)^2 + (-2 + 1)^2 = 5 \\ d^2(D, (BCD)) = (-3 + 1)^2 + (-2 + 1)^2 = 5 \end{cases}$$
- $$\begin{cases} d^2(A, A) = 0 \\ d^2(A, B) = 40 \\ d^2(A, C) = 41 \\ d^2(A, D) = 89 \end{cases} \Leftrightarrow$$

	A	B	C	D
A	0	40	41	89
BCD	52	4	5	5

$$=> \begin{cases} A \textit{ thuộc cụm } (A) \\ B \textit{ thuộc cụm } (BCD) \\ C \textit{ thuộc cụm } (BCD) \\ D \textit{ thuộc cụm } (BCD) \end{cases}$$

=> Vậy có 2 cụm là (A) và (BCD)

11 Hướng dẫn sử dụng RStudio

11.1 Hồi quy tuyến tính đơn:

11.1.1 Phương trình đường thẳng HQTT:

```
model = lm(y ~ x, data = dataset)
summary(model)
```

11.1.2 Lấy hệ số a, b từ Y = a + bX:

```
coefficients = coef(model)
a = coefficients[1]
b = coefficients[2]
```

11.1.3 Kiểm định phần dư có phân bố chuẩn với GTTB = 0

```
test = shapiro.test(residuals(model))
pValue = test$p.values
```

Điều kiện: $\begin{cases} pValue < 0.05 \longrightarrow \ \textit{phần dư không chuẩn} \\ pValue > 0.05 \longrightarrow \ \textit{phần dư chuẩn} \end{cases}$

11.1.4 Khoảng tin cậy $\alpha\%$ cho hàm số hồi quy:

```
confint(model, level = alpha)
```

trong đó $\alpha = alpha$

11.1.5 Với Y = newValue, đưa ra dự đoán về giá trị của X với khoảng tin cậy $\alpha\%$ cho GTTB của X

```
new_data = data.frame(Y = newValue)
pre = predict(model, newdata = new_data,
              interval = "confidence", level = alpha)
```

trong đó $\alpha = alpha$

11.1.6 Bài toán kiểm định phần dư:

```
# GTTB nếu chuẩn
t.test(model$residuals, mu = 0)

# GTTB nếu không chuẩn
wilcox.test(model$residuals)
```

11.1.7 Kẻ đường tuyến tính:

```
abline(model, lwd = "mức độ nét", col)
```

11.1.8 Kiểm định phương sai không đổi:

```
library(car)
ncvTest(model)
```

11.2 Phân bố chuẩn:

Lưu ý: Điều kiện có thể biểu diễn tuyến tính pValue nhỏ và R_squared cao (model tốt $\longrightarrow$ max)

11.2.1 Các hàm trong phân phối chuẩn $X \sim N(mean, sd^2)$:

- Hàm mật độ f(x) = dnorm(x, mean, sd)
- Hàm phân phối P(x < q) = pnorm(x, mean, sd)
- Hàm phân vị $\longrightarrow$ tìm a để P(X < a) = p là qnorm(p, mean, sd)
- Hàm sinh ngẫu nhiên $\longrightarrow$ rnorm(n, mean, sd)
- Hàm sinh ngẫu nhiên phân phối đa biến:
  - Tải thư viện $\longrightarrow$ install.packages('mvtnorm')
  - Gọi thư viện $\longrightarrow$ library(mvtnorm)
  - Thực thi $\longrightarrow$ rmvnorm(n, mean, sigma)

11.2.2 Các đặc trưng cơ bản:

- mean(sample) $\longrightarrow$ kỳ vọng
- var(sample) $\longrightarrow$ phương sai

- sd(sample) $\longrightarrow$ độ lệch chuẩn
- colMeans(data) $\longrightarrow$ GTTB theo cột
- cov(data) $\longrightarrow$ ma trận hiệp phương sai
- cor(data) $\longrightarrow$ ma trận tương quan mẫu

11.3 Phân phối chuẩn nhiều chiều:

Lưu ý: "alternative" luôn theo đối thiết H_1

11.3.1 Kiểm định GTTB của 1 biến có bằng value không?

- Kiểm định $\begin{cases} H_0 : EX = value \\ H_1 : EX \neq value \end{cases}$
  => t.test(data$colname, mu = value, conf.level = 1 - myn)

- Kiểm định $\begin{cases} H_0 : EX = value \\ H_1 : EX > value \text{ hoặc } EX < value \end{cases}$
  => t.test(data$colname, mu = value, alternative = "greater/less", conf.level = 1 - myn)

11.3.2 Kiểm định GTTB 2 mẫu có sự khác biệt:

- Kiểm định $\begin{cases} H_0 : EX = EY \\ H_1 : EX \neq EY \end{cases}$
  => t.test(col1, col2, alternative = ..., conf.level = 1 - myn)

11.3.3 Kiểm định nhiều chiều với phân phối chuẩn:

- sample = sample1 - sample2 tương tự với phần trên
- Độ tin cậy $\alpha = 1 - myn$

  => t.test(sample1, sample2, conf = alpha)

11.3.4 Kiểm định nhiều chiều:

```
matrix = data.frame(c1, c2, ..., cn)
HotellingsT2(matrix, mu = c(0, ..., 0))
HotellingsT2(matrix1, matrix2)
```

11.3.5 Kiểm tra phân phối chuẩn nhiều chiều:

1. Kiểm tra tính chuẩn 1 chiều từng biến (nếu 1 biến không tuân theo thì tất cả không tuân theo)
2. Nếu B1 đúng, kiểm tra tính chuẩn nhiều chiều n biến:
   - mah = mahalanobis(data, colMeans(data), var(data))
   - shapiro.test(qnorm(pchisq(mah, n)))

11.4 Phân tích thành phần chính:

```
pc = princomp(data)
summary(pc)
```

- standar deviation $\longrightarrow$ độ lệch tiêu chuẩn các thành phần chính
- proportion of varience $\longrightarrow$ tỷ lệ biến sai tổng cộng $\frac{DY_i}{\sum DY_j}$
- cumulative proportion $\longrightarrow$ tìm ra số TPC cần thiết khi đề bài yêu cầu biểu ?% thông tin về bộ dữ liệu ban đầu

11.4.1 Phân tích TPC trên ma trận hiệp phương sai:

```
pc = princomp(covmat = cov(data))
summary(pc)
```

trong đó $\begin{cases} DY_i = lambda_i \\ Y_i = e_i'X \longrightarrow pc\$sdev \end{cases}$

- Phương sai các TPC $\longrightarrow (pc\$sdev)^2$
- Giá trị riêng $\longrightarrow$ eigen(cov(data))$values
- Vector riêng $\longrightarrow$ eigen(cov(data))$vectors
- Ma trận tải trọng (các hệ số tải l_{ij}) $\longrightarrow$ pc$loadings

11.4.2 Phân tích TPC trên ma trận tương quan:

Lưu ý: sử dụng khi mẫu (bộ dữ liệu) khác nhau về thang đo

```
pcacor = princomp(data, cor = TRUE)
summary(pcacor)
```

Để thu được α% thông tin về bộ dữ liệu, ta cần m thành phần chính, trong đó λ_α gần nhất với sai số
=> Sai số = 1 - λ_α

11.5 Phân tích nhân tố:

```
factanal(hemangioma, factor = <number>)
```

trong đó <number> là số nhân tố muốn phân tích

- Uniquenesses ⟶ sai số ϵ
- Loadings ⟶ tỷ lệ chi phối từng Factor
- Dựa vào ma trận tải trọng:
  1. Có thể biểu diễn $X_i = F_1 + F_2 + F_3$
  2. Biến X_i bị chi phối bởi những biến nào

11.6 Mô hình hồi quy tuyến tính bội:

```
mohinhVaiBien = lm(formula = <var> ~ val1 + val2 + ...,
                    data = dat)
mohinhToanBien = lm(formula = <var> ~ . , data = dat)
```

11.6.1 Kiểm định phần dư có tương quan không?

Kiểm định $\begin{cases} H_0 : \textit{phần dư không tương quan} \\ H_1 : \textit{phần dư có tương quan} \end{cases}$ => resid = mohinh$residuals
=> So sánh p_value = y - y_h

1. Cách 1: library(car); durbinWatsonTest(...)
2. Cách 2: library(lmtest); dwtest(...)

11.6.2 Phần dư có tuân theo phân phối chuẩn với GTTB = 0 không?

```
shapiro.test(model$residuals)
+ Chuẩn --> t.test(s$residuals)
+ Không chuẩn --> wilcox.test(s$residuals)
```

11.6.3 Các hệ số trong mô hình có thực sự khác 0 không?

Cho $\begin{cases} H_0 : a_i = 0 \\ H_1 : a_i \neq 0 \end{cases}$ => summary(modelName)

Dựa vào Pr(>|t|) $\begin{cases} pValue < 0.05 => & \textit{Bác bỏ } H_0 \\ pValue > 0.05 => & \textit{Chấp nhận } H_0 \end{cases}$

11.6.4 Phân tích theo phương pháp forward/backward/both:

```
# Mô hình đơn giản
only = lm(npg ~ 1, data = dat)

# Mô hình phức tạp nhất
all = lm(mpg ~ . , data = dat)
```

1. forward (only $\longrightarrow$ all)
   => fw = step(only, formula(all), direction = "forward", trace = 0)
   => fw$anova

2. backward (all $\longrightarrow$ only)
   => bw = step(all, formula(all), direction = "backward", trace = 0)
   => bw$anova

3. both (cả 2 mô hình kết hợp)
   => both = step(all, formula(all), direction = "both", trace = 0)

11.6.5 Kiểm tra sự phục thuộc của từng biến:

Lưu ý: quan sát Pr(>F) = pValue

```
univ = aov(y ~ x1 + x2 + ..., data = dat)
summary(univ)
```

Cho $\begin{cases} y, x_i \textit{ độc lập} \\ y, x_i \textit{ phụ thuộc} \end{cases}$

- Độ tương quan cor(y, x) với mức độ mạnh yếu dựa vào F
- Hệ số xác định mô hình là R^2 để đo mức độ phù hợp mô hình

11.6.6 Ước lượng khoảng tin cậy $\alpha\%$ cho các hệ số (β_n và KTC tương ứng)

$$y = b_0 + b_1x_1 + ... \ (\textit{ước lượng} b_0, b_1, ...)$$

```r
# Mô hình phù hợp KTC alpha%
confint(univ, level = alpha%)
```

11.6.7 Bài toán dự đoán mô hình:

```r
# Khoảng dự đoán
predict(univ, newdata = data.frame(x1=a1, x2=a2, ...),
        interval = "prediction", level = alpha)

# Khoảng tin cậy
predict(univ, newdata = data.frame(x1=a1, x2=a2, ...),
        interval = "prediction", conf.alpha = alpha)
```

11.7 Các phép toán với ma trận:

Ký hiệu	**Ý nghĩa**
$A\% * \%B$	Nhân ma trận
$t(A)$	Chuyển vị ma trận
$\det(A)$	Định thức ma trận
solve(A)	Nghịch đảo ma trận
diag(A)	Ma trận đơn vị trên đường chéo chính

Các ký hiệu ma trận và ý nghĩa của chúng

11.8 Một số loại biểu đồ:

- Tán xạ (biến < 10) $\longrightarrow$ pairs(data)
- Nhiệt (số lượng biến lớn) $\longrightarrow$
  - Thư viện: library(ggplot2)
  - ggplot(...)
- Biểu đồ TPC $\longrightarrow$ biplot(...)

Tài liệu

[1] Pisces Kibo. *Bộ công thức Tony*, 2024.