ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
KHOA TOÁN TIN
FACULTY OF MATHEMATICS AND INFORMATICS

Phân tích số liệu

Chủ đề: Phân tích thành phần chính

Giáo viên hướng dẫn: ThS. Lê Xuân Lý

Nhóm 8: Trần Ngọc Bảo 20227086

Nguyễn Đức Mạnh 20227021

Nguyễn Việt Hoàng 20227019

Lê Thanh Thảo 20227152

Lê Cảnh Hải 20227107

Nguyễn Trọng Đoàn 20227041

Đặng Quyết Tiến 20227033

Bùi Thị Như Ngọc 20227027

Mai Văn Tưởng 20227073

Trần Anh Vũ 20216905

Ngày 15 tháng 1 năm 2025

Mục lục

Danh sách thành viên

STT	Họ và tên	MSSV	Công việc	Điểm cộng
1	Nguyễn Đức Mạnh	20227021	Giới thiệu chủ đề + Cấu trúc thành phần chính của một tổng thể	1.5
2	Nguyễn Việt Hoàng	20227019	Các thành phần chính thu được từ biến đã chuẩn hóa	1.5
3	Lê Thanh Thảo	20227152	Phân tích thành phần chính cho ma trân hiệp phương sai với cấu trúc đặc biệt	1.5
4	Lê Cảnh Hải	20227107	Cấu trúc thành phần chính của một tổng thể	1.5
5	Trần Ngọc Bảo	20227086	Phân công công việc + Vẽ đồ thị thành phần chính	1.5
6	Nguyễn Trọng Đoàn	20227041	Suy luận dựa trên mẫu cỡ lớn	1.5
7	Đặng Quyết Tiến	20227033	Ứng dụng PCA	1.5
8	Bùi Thị Như Ngọc	20227027	Sự thay đổi của mẫu theo thành phần chính + Số lượng thành phần chính	1.5
9	Mai Văn Tưởng	20227073	Biểu diễn hình học thành phần chính của mẫu + Thành phần chính của mẫu đã chuẩn hóa	1.5
10	Trần Anh Vũ	20216905	Vẽ đồ thị thành phần chính	1.5

I Giới thiệu bài toán

1 Bối cảnh bài toán

Trong thực tế, các biến ngẫu nhiên thường có số chiều và kích thước rất lớn, có thể lên tới hàng nghìn. Thêm vào đó, số lượng các điểm dữ liệu lớn gây khó khăn cho việc lưu trữ và tốc độ tính toán.

Giả sử cho p biến ngẫu nhiên. Liệu có cách nào để xây dựng được p biến mới không tương quan với nhau và được biểu diễn tuyến tính thông qua các biến cũ, đồng thời không làm mất mát dữ liệu ban đầu hoặc làm mất mát ít nhất?

Câu trả lời chính là **Thuật toán phân tích thành phần chính (Principal Component Analysis - PCA)**.

Kỹ thuật Phân tích thành phần chính (PCA) được Karl Pearson giới thiệu vào năm 1901.Kỹ thuật này hoạt động dựa trên điều kiện là khi dữ liệu trong không gian có nhiều chiều hơn được ánh xạ vào trong không gian có ít chiều hơn, thì phương sai của dữ liệu trong không gian có ít chiều hơn phải là lớn nhất. Qua đó giúp giảm số chiều của dữ liệu mà vẫn giữ lại phần lớn thông tin quan trọng.

2 Mục tiêu của PCA

Phân tích thành phần chính (PCA) là một kỹ thuật phân tích cấu trúc ma trận hiệp phương sai $\mathbf{\Sigma}$ của một tập hợp các biến X thông qua các tổ hợp tuyến tính của các biến đó. Mục tiêu chính của PCA bao gồm:

1. Giảm số chiều của dữ liệu.
2. Thay vì giữ lại các trục tọa độ của không gian cũ, PCA xây dựng một không gian mới ít chiều hơn, nhưng lại có khả năng biểu diễn dữ liệu tốt tương đương không gian cũ, tức là đảm bảo độ biến thiên của dữ liệu trên mỗi chiều mới.
3. Các trục tọa độ trong không gian mới là tổ hợp tuyến tính của không gian cũ. Về mặt ngữ nghĩa, PCA xây dựng các biến mới dựa trên các biến đã quan sát được, đồng thời vẫn biểu diễn tốt dữ liệu ban đầu.
4. Trong không gian mới, các liên kết tiềm ẩn của dữ liệu có thể được khám phá. Trong không gian cũ, các liên kết này có thể khó phát hiện hơn hoặc không rõ ràng.

3 Ví dụ về PCA

Giả sử chúng ta có dữ liệu về một số đặc trưng của các ngôi nhà, bao gồm:

1. Diện tích (S).
2. Số phòng ngủ (R).
3. Số phòng tắm (B).
4. Diện tích sân vườn (G).
5. Khoảng cách đến trung tâm thành phố (D).
6. Tuổi của ngôi nhà (A).

Dữ liệu này có thể được biểu diễn bằng một ma trận dữ liệu với m ngôi nhà và $n = 6$ đặc trưng:

$$\begin{bmatrix} S_1 & R_1 & B_1 & G_1 & D_1 & A_1 \\ S_2 & R_2 & B_2 & G_2 & D_2 & A_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ S_m & R_m & B_m & G_m & D_m & A_m \end{bmatrix}$$

Mục tiêu:

- Chúng ta muốn dự đoán giá trị của một ngôi nhà dựa trên các đặc trưng này.
- Tuy nhiên, có những đặc trưng tương quan chặt chẽ với nhau (như diện tích ngôi nhà, số phòng ngủ, số phòng tắm), và một số đặc trưng có thể ít quan trọng hơn (như diện tích sân vườn hoặc khoảng cách đến trung tâm thành phố).
- Vì vậy, PCA có thể giúp chúng ta giảm số chiều dữ liệu từ 6 xuống một số chiều nhỏ hơn mà vẫn giữ được phần lớn thông tin quan trọng.

Chọn k thành phần chính

- Giả sử $k = 2$, chúng ta chọn hai vector riêng tương ứng với hai giá trị riêng lớn nhất, tạo ra một không gian mới với 2 chiều. Các vector riêng này có thể biểu diễn sự kết hợp của các đặc trưng ban đầu.
- **ví dụ**:

  - Vector riêng thứ nhất có thể đại diện cho "tổng kích thước" của ngôi nhà, là sự kết hợp của diện tích, số phòng ngủ và số phòng tắm.
  - Vector riêng thứ hai có thể đại diện cho "vị trí địa lý" của ngôi nhà, là sự kết hợp của khoảng cách đến trung tâm thành phố và diện tích sân vườn.

Dữ liệu sau khi giảm chiều có thể được biểu diễn bởi ma trận Z:

- Giả sử sau khi thực hiện PCA, ta nhận thấy rằng các đặc trưng "diện tích", "số phòng ngủ" và "số phòng tắm" tương quan chặt chẽ và có thể được thay thế bằng một thành phần chính duy nhất.
- Tương tự, "diện tích sân vườn" và "khoảng cách đến trung tâm thành phố" có thể được gộp lại thành một thành phần chính thứ hai.
- Bây giờ, thay vì sử dụng 6 đặc trưng ban đầu, chúng ta chỉ cần sử dụng 2 đặc trưng:

$$Z = \begin{bmatrix} z_1^{(1)} & z_2^{(1)} \\ z_1^{(2)} & z_2^{(2)} \\ \vdots & \vdots \\ z_1^{(m)} & z_2^{(m)} \end{bmatrix}$$

  Trong đó z_1 và z_2 đại diện cho hai thành phần chính (principal components) quan trọng nhất.

4 Các thành phần chính

- Các thành phần chính p được dùng để tạo tổng độ biến thiên của hệ thống, nhưng thường thì phần lớn độ biến thiên này có thể được tính toán bởi một phần nhỏ số k các thành phần chính, mà vẫn giữ gần như toàn bộ thông tin trong p thành phần ban đầu.
- Do đó, k thành phần chính có thể được sử dụng thay thế cho p thành phần ban đầu, giúp giảm độ phức tạp mà vẫn đảm bảo tính chính xác.

Khái quát

Các số liệu được xem như tạo thành một "đám mây". PCA giúp phản ánh tốt nhất hình ảnh của "đám mây", tức là "bóng" của nó chiếu lên các trục tọa độ. Nói cách khác, PCA dịch chuyển gốc tọa độ đến trọng tâm của "đám mây"và xác định các trục tọa độ mới chính là các thành phần chính.

5 Ý tưởng và cách thức triển khai thuật toán

1. **Chuẩn hóa dữ liệu**: Trước khi áp dụng PCA, dữ liệu cần được chuẩn hóa để mỗi đặc trưng có cùng độ lớn. Điều này rất quan trọng vì các đặc trưng khác nhau có thể có **đơn vị đo khác nhau** (ví dụ: chiều dài và trọng lượng), và nếu không chuẩn hóa, các đặc trưng có giá trị lớn hơn sẽ chiếm ưu thế trong kết quả PCA.

2. **Tính ma trận hiệp phương sai**: để hiểu rõ mối quan hệ giữa các đặc trưng. Ma trận hiệp phương sai cho biết mức độ mà hai đặc trưng thay đổi cùng nhau. Nếu hai đặc trưng có tương quan cao, điều này ngụ ý rằng có thể loại bỏ một trong số chúng mà không mất nhiều thông tin.

3. **Tính giá trị riêng và vectơ riêng của ma trận hiệp phương sai**: Từ ma trận hiệp phương sai, chúng ta tiến hành tính giá trị riêng (eigenvalues) và vectơ riêng (eigenvectors). Vectơ riêng xác định hướng của các thành phần chính, còn giá trị riêng cho biết mức độ biến thiên mà mỗi thành phần chính nắm giữ. Các thành phần chính này là các tổ hợp tuyến tính của các đặc trưng ban đầu, được sắp xếp theo thứ tự từ quan trọng nhất đến ít quan trọng nhất

4. **Chọn số lượng thành phần chính**: Giữ lại những thành phần chính có tổng giá trị riêng chiếm 90-95% tổng biến thiên.

5. **Chiếu dữ liệu lên không gian mới**: Dữ liệu được chiếu lên không gian mới với số chiều giảm nhưng vẫn giữ lại phần lớn thông tin.

II Thành phần chính tổng thể

1 Một số khái niệm và ý nghĩa

Ma trận hiệp phương sai (Covariance Matrix): Ma trận hiệp phương sai biểu diễn mức độ mà hai biến cùng thay đổi với nhau. Để tính ma trận hiệp phương sai của một tập hợp các biến, bạn có thể sử dụng công thức:

Công thức:

Giả sử bạn có n quan sát cho p biến $X_1, X_2, \ldots, X_p$. Khi đó, phần tử C_{ij} trong ma trận hiệp phương sai là hiệp phương sai giữa biến X_i và X_j:

$$C_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (X_{ki} - \mu_i)(X_{kj} - \mu_j).$$

Trong đó:
- X_{ki} là giá trị của biến X_i tại quan sát thứ k.
- μ_i là giá trị trung bình của biến X_i.
- n là số lượng quan sát.

Nếu có một ma trận dữ liệu X (kích thước $n \times p$, với n là số mẫu và p là số biến), có thể tính ma trận hiệp phương sai bằng:

$$\text{Cov}(X) = \frac{1}{n-1}(X^\top X).$$

Ma trận hệ số tương quan (Correlation Matrix): Ma trận hệ số tương quan chuẩn hóa ma trận hiệp phương sai sao cho các giá trị nằm trong khoảng từ -1 đến 1, đo lường mức độ tuyến tính giữa các biến.

Công thức:

Giả sử bạn có ma trận hiệp phương sai $\text{Cov}(X)$, hệ số tương quan giữa biến X_i và X_j là:

$$\rho_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sigma_i \sigma_j}.$$

Trong đó

- $\text{Cov}(X_i, X_j)$ là hiệp phương sai giữa X_i và X_j.
- σ_i và σ_j là độ lệch chuẩn của X_i và X_j.

Nếu đã có ma trận hiệp phương sai $\text{Cov}(X)$, có thể chuyển nó thành ma trận tương quan bằng công thức:

$$\text{Corr}(X) = D^{-1/2} \cdot \text{Cov}(X) \cdot D^{-1/2},$$

Trong đó D là ma trận đường chéo chứa các phương sai của các biến trên đường chéo chính.

- **Vecto riêng v** là một vecto không phải là vecto không, thỏa mãn phương trình: $\text{Av} = \lambda \text{v}$ Nó biểu thị hướng mà tại đó ma trận A chỉ làm thay đổi độ lớn của vecto, chứ không thay đổi hướng của nó.
- **Giá trị riêng** ứng với λ trong phương trình đó.
- **Vectơ riêng v** tương ứng với các hướng (hay trục) của không gian mới mà chúng ta muốn chiếu dữ liệu lên.

Cụ thể

Giá trị riêng và vectơ riêng: Khi bạn tính ma trận hiệp phương sai, việc phân tích giá trị riêng và vectơ riêng cho phép xác định những trục nào mang lại nhiều thông tin nhất. Thành phần chính đầu tiên là hướng có nhiều biến thiên nhất trong dữ liệu, và thành phần chính thứ hai sẽ là hướng vuông góc với thành phần chính đầu tiên nhưng vẫn giữ được nhiều thông tin nhất từ dữ liệu.

(2-45)

Tổ Hợp tuyến tính $\mathbf{c}'\mathbf{X} = c_1X_1 + \cdots + c_pX_p$ có

$$\begin{aligned} \text{trung bình } &= \mathbb{E}(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\mu, \\ \text{phương sai} &= \text{Var}(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\Sigma\mathbf{c}. \end{aligned} \tag{2-43}$$

tại $\mu = \mathbb{E}(\mathbf{X})$ and $\Sigma = \text{Cov}(\mathbf{X})$:

Tổng thể q tổ hợp tuyến tính của p biến ngẫu nhiên $X_1, \ldots, X_p$:

$$\begin{aligned} Y_1 &= c_{11}X_1 + c_{12}X_2 + \cdots + c_{1p}X_p, \\ Y_2 &= c_{21}X_1 + c_{22}X_2 + \cdots + c_{2p}X_p, \\ &\vdots \\ Y_q &= c_{q1}X_1 + c_{q2}X_2 + \cdots + c_{qp}X_p. \end{aligned}$$

hay

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{q1} & c_{q2} & \cdots & c_{qp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = C\mathbf{X}. \tag{2-44}$$

Tổ hợp tuyến tính $\mathbf{Y} = C\mathbf{X}$ có

$$\mu_{\mathbf{Y}} = \mathbb{E}(\mathbf{Y}) = \mathbb{E}(C\mathbf{X}) = C\mu_{\mathbf{X}},$$

$$\Sigma_{\mathbf{Y}} = \mathrm{Cov}(\mathbf{Y}) = \mathrm{Cov}(C\mathbf{X}) = C\Sigma_{\mathbf{X}}C'. \tag{2-45}$$

2 Cấu trúc thành phần chính

Xét véc tơ ngẫu nhiên p chiều $\boldsymbol{X}' = [X_1, ..., X_p] \in \mathbb{R}^p$ có ma trận hiệp phương sai $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma} = [\sigma_{ij}]_{p\times p}$ với các giá trị riêng $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_p \geqslant 0$.
Xét p tổ hợp tuyến tính:

$$\begin{aligned} Y_1 &= \boldsymbol{a}_1'\boldsymbol{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p, \\ Y_2 &= \boldsymbol{a}_2'\boldsymbol{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p, \\ &\vdots \\ Y_p &= \boldsymbol{a}_p'\boldsymbol{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p. \end{aligned} \tag{1}$$

Từ (2-45): $Y = CX$, ta có:

$$\mu_Y = E(Y) = E(CX) = C\mu_X,$$
$$\Sigma_Y = Cov(Y) = Cov(CX) = C\Sigma_X C'.$$

Ta được:

$$\mathrm{Var}(Y_i) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_i, \quad i = 1, 2, \ldots, p, \tag{2}$$
$$\mathrm{Cov}(Y_i, Y_k) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_k, \quad i, k = 1, 2, \ldots, p. \tag{3}$$

Các thành phần chính là các tổ hợp tuyến tính không tương quan $Y_1, Y_2, \ldots, Y_p$ mà phương sai của chúng trong công thức (2) càng lớn càng khả thi.

Lý do:

Trong phân tích thành phần chính (PCA), **mục tiêu** là tìm các tổ hợp tuyến tính của các biến đầu vào sao cho **phương sai** của dữ liệu trong các tổ hợp này được **tối đa hóa**. Các thành phần chính (principal components) được xây dựng theo cách tối đa hóa phương sai một cách tuần tự, đồng thời đảm bảo rằng các *thành phần chính này là trực giao (không tương quan)* với nhau.

Thành phần chính đầu tiên:

Thành phần chính đầu tiên là tổ hợp tuyến tính của các biến $\mathbf{X}$ được biểu diễn dưới dạng $a_1'\mathbf{X}$. Tổ hợp này được chọn sao cho nó tối đa hóa phương sai:

$$\text{Var}(a_1'\mathbf{X}) = a_1'\Sigma a_1,$$

Trong đó:

- Σ là ma trận hiệp phương sai của dữ liệu $\mathbf{X}$.
- a_1 là vectơ trọng số của tổ hợp tuyến tính.

Tuy nhiên, $\text{Var}(a_1'\mathbf{X})$ có thể được tăng vô hạn nếu chúng ta nhân a_1 với một hằng số bất kỳ. Để loại bỏ sự không xác định này, ta áp đặt điều kiện $a_1'a_1 = 1$, tức là vectơ trọng số a_1 có độ dài đơn vị (norm bằng 1).

Như vậy, thành phần chính đầu tiên là tổ hợp tuyến tính $a_1'\mathbf{X}$ sao cho nó tối đa hóa $\text{Var}(a_1'\mathbf{X})$ thỏa mãn điều kiện $a_1'a_1 = 1$.

Thành phần chính thứ hai:

Thành phần chính thứ hai là tổ hợp tuyến tính của các biến $\mathbf{X}$ được biểu diễn dưới dạng $a_2'\mathbf{X}$, sao cho nó tối đa hóa phương sai:

$$\text{Var}(a_2'\mathbf{X}) = a_2'\Sigma a_2.$$

Đồng thời, thành phần chính thứ hai phải thỏa mãn các điều kiện

- $a_2'a_2 = 1$, tức là vectơ trọng số a_2 cũng có độ dài đơn vị.
- $\text{Cov}(a_1'\mathbf{X}, a_2'\mathbf{X}) = 0$, nghĩa là thành phần chính thứ hai không tương quan với thành phần chính thứ nhất.

Quy trình này tiếp tục cho các thành phần chính tiếp theo, với mỗi thành phần chính mới tối đa hóa phương sai của tổ hợp tuyến tính, đồng thời đảm bảo tính trực giao với các thành phần trước đó.

Tại bước thứ i

thành phần chính thứ i là tổ hợp tuyến tính $a_i'X$ tối đa hóa $\text{Var}(a_i'X)$ với điều kiện

$$\boldsymbol{a}_i'\boldsymbol{a}_i = 1 \text{ và } \text{Cov}(\boldsymbol{a}_i'\boldsymbol{X}, \boldsymbol{a}_k'\boldsymbol{X}) = 0,\ i = 1, \ldots, k-1. \tag{4}$$

Định nghĩa 2.1:

Các đại lượng $Y_i = (\boldsymbol{a}_i)'\boldsymbol{X}$ thỏa mãn điều kiện (4) được gọi là thành phần chính thứ i của véc tơ X với ma trận hiệp phương sai $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

Kết quả 2.1

Cho các véc tơ ngẫu nhiên $\boldsymbol{X}' = [X_1, X_2, ..., X_p]$ có $\boldsymbol{\Sigma}$ là ma trận hiệp phương sai. Cho $\boldsymbol{\Sigma}$ có các cặp trị riêng, véc tơ riêng tương ứng là $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$ với $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Khi đó, thành phần chính thứ i được tạo thành bởi

$$Y_i = \boldsymbol{e}_i' \boldsymbol{X} = e_{i1}X_1 + e_{i2}X_2 + ... + e_{ip}X_p, \quad i = 1, 2, ..., p \tag{5}$$

Với giả thiết như trên, ta có

$$\text{Var}(Y_i) = \boldsymbol{e}_i' \boldsymbol{\Sigma} \boldsymbol{e}_i = \lambda_i, \quad i = 1, 2, ..., p \tag{6}$$

$$\text{Cov}(Y_i, Y_k) = \boldsymbol{e}_i' \boldsymbol{\Sigma} \boldsymbol{e}_k = 0, \quad i \neq k. \tag{7}$$

Chú ý

Nếu một số giá trị λ_i bằng nhau, việc chọn các vector hệ số tương ứng, e_i, và do đó Y_i, sẽ không duy nhất.

Chứng Minh

Nhắc lại (2-51): Tính cực đại bậc hai cho các điểm trên mặt cầu đơn vị:

$$\max_{x \neq 0} \frac{x'Bx}{x'x} = \lambda_1 \text{ (đạt được khi } x = e_1),$$
$$\min_{x \neq 0} \frac{x'Bx}{x'x} = \lambda_p \text{ (đạt được khi } x = e_p).$$

Từ (2-51), $B = \Sigma$, ta có:

$$\max_{a \neq 0} \frac{a'\Sigma a}{a'a} = \lambda_1 \text{ (đạt được khi } a = e_1).$$

Do các vector riêng đã được chuẩn hóa ${e_1}'e_1 = 1$ nên:

$$\max_{\boldsymbol{a} \neq 0} \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} = \max_{\boldsymbol{a} \neq 0} \frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{a}} = \lambda_1 = \frac{\boldsymbol{e}_1'\boldsymbol{\Sigma}\boldsymbol{e}_1}{\boldsymbol{e}_1'\boldsymbol{e}_1} = \text{Var}(Y_1),$$

Sử dụng (2-52):

$$\max_{x \perp e_1, e_2, .., e_k} \frac{x'Bx}{x'x} = \lambda_{k+1}, x = e_{k+1}, k = 1, 2, ..., p-1.$$

Ta có:

$$\max_{\boldsymbol{a} \perp \boldsymbol{e}_1, ..., \boldsymbol{e}_k} \frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{a}} = \lambda_{k+1}, \quad k = 1, \ldots, p-1$$

Chọn $\boldsymbol{a} = \boldsymbol{e}_{k+1}$ với $\boldsymbol{e}_{k+1}'\boldsymbol{e}_i = 0, \ \forall i \in \{1, \ldots, k\}; k \in \{1, \ldots, p-1\}$

$$\frac{\boldsymbol{e}_{k+1}'\boldsymbol{\Sigma}\boldsymbol{e}_{k+1}}{\boldsymbol{e}_{k+1}'\boldsymbol{e}_{k+1}} = \boldsymbol{e}_{k+1}'\boldsymbol{\Sigma}e_{k+1} = \text{Var}(Y_{k+1}).$$

Như vậy, ta có được điều phải chứng minh.

Nhận xét 2.1

Tất cả các trị riêng của Σ sẽ trực giao nếu các trị riêng $\lambda_1, \lambda_2, ..., \lambda_p$ là riêng biệt. Vì vậy, với hai vctor bất kì e_i và e_k, ${e_i}'e_k = 0, i \neq k$. Mặt khác, $\Sigma e_k = \lambda_k e_k$, nhân phía trước 2 vế với ${e_i}'$ ta được:

$$Cov(Y_i, Y_k) = {e_i}'\Sigma e_k = {e_i}'\lambda_k e_k = \lambda_k {e_i}'e_k = 0 (i \neq k).$$

Kết Luận

Từ kết quả 1.1, các thành phần chính không tương quan và có phương sai bằng trị riêng của ma trận $\boldsymbol{\Sigma}$.

Kết quả 2.2

Cho các véc tơ ngẫu nhiên $\boldsymbol{X}' = [X_1, X_2, ..., X_p]$ có $\boldsymbol{\Sigma}$ là ma trận hiệp phương sai. Cho $\boldsymbol{\Sigma}$ có các cặp trị riêng, véc tơ riêng tương ứng là $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$ với $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Cho các thành phần chính $Y_1 = \boldsymbol{e}_1' X, Y_2 = \boldsymbol{e}_2' X, ..., Y_p = \boldsymbol{e}_p' X$. Ta có

$$\sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \sum_{i=1}^{p} \text{Var}(X_i) = \lambda_1 + \lambda_2 + ... + \lambda_p = \sum_{i=1}^{p} \text{Var}(Y_i).$$

Chứng minh.

Từ định nghĩa 2A.28, $\sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = tr(\Sigma)$.

Từ (2-20) cho:

$$A = \sum_{i=1}^{k} \lambda_i e_i {e_i}' = P \Lambda P'$$

Thay thế $A = \Sigma$ ta có thể viết $\boldsymbol{\Sigma} = \boldsymbol{P\Lambda P'}$, ở đó $\boldsymbol{\Lambda}$ là ma trận không với đường chéo chính là các trị riêng

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

Và ma trận $\boldsymbol{P} = [\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_p]$, vì thế $\boldsymbol{PP'} = \boldsymbol{P'P} = \boldsymbol{I}$.

Sử dụng kết quả 2A.12(c),

$$tr(AB) = tr(BA)$$

ta có

$$tr(\boldsymbol{\Sigma}) = tr(\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}') = tr(\boldsymbol{\Lambda}\boldsymbol{P}'\boldsymbol{P}) = tr(\boldsymbol{\Lambda}) = \lambda_1 + \lambda_2 + ... + \lambda_p.$$

Do đó,

$$\sum_{i=1}^{p} \text{Var}(X_i) = tr(\boldsymbol{\Sigma}) = tr(\boldsymbol{\Lambda}) = \sum_{i=1}^{p} \text{Var}(Y_i)$$

Ta có điều phải chứng minh.

Kết quả 1.2 chỉ ra rằng

$$\begin{aligned} \textbf{Phương sai tổng thể} &= \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_p \end{aligned}$$

Do vậy, tỷ lệ thành phần chính thứ k trên tổng các thành phần chính là

$$\begin{pmatrix} \text{Tỷ lệ thành phần chính} \\ \text{thứ } k \text{ trên tổng} \\ \text{các thành phần chính} \end{pmatrix} = \frac{\lambda_k}{\rho} \tag{8}$$

với $\rho = \lambda_1 + \lambda_2 + ... + \lambda_p,\ k = 1, 2, ..., p$

Nhận xét 2.2

Nếu phần lớn (ví dụ khoảng $80 - 90\%$) **phương sai tổng thể** có thể **quy cho** thành phần chính thứ nhất, thành phần chính thứ hai, thành phần chính thứ ba **đến thành phần chính thứ** i. Và các thành phần chính này có thể thay thế cho p biến ban đầu với sự mất mát thông tin là ít nhất.

Nhận xét 2.3

Mỗi thành phần chính của ma trận hệ số $\boldsymbol{e}_i' = [e_{i1}, ..., e_{ik}, ..., e_{ip}]$ cũng có giá trị kiểm tra. **Độ lớn của** e_{ik} thể hiện **tầm quan trọng** của biến thứ k đối với **thành phần chính thứ** $\boldsymbol{i}$. Cụ thể, e_{ik} tỷ lệ thuận với hệ số tương quan giữa Y_i và X_k.

Kết quả 2.3

Nếu $Y_1 = \boldsymbol{e}_1'\boldsymbol{X}, Y_2 = \boldsymbol{e}_2'\boldsymbol{X}, ..., Y_p = \boldsymbol{e}_p'\boldsymbol{X}$ là những thành phần chính thu được từ ma trận hiệp phương sai $\boldsymbol{\Sigma}$, ta được

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \qquad i, k = 1, 2, ..., p \tag{9}$$

là hệ số tương quan giữa thành phần Y_i và biến X_k.

Chứng minh. Đặt $\boldsymbol{a}_k' = [0, ..., 0, 1, 0, ..., 0]$ ta có

$$X_k = \boldsymbol{a}_k'\boldsymbol{X} \text{ và } \text{Cov}(X_k, Y_i) = \text{Cov}(\boldsymbol{a}_k'\boldsymbol{X}, \boldsymbol{e}_i'\boldsymbol{X}) = \boldsymbol{a}_k'\boldsymbol{\Sigma}\boldsymbol{e}_i.$$

Vì $\boldsymbol{e}_i$ là vecto riêng ứng với giá trị riêng λ_i nên $\boldsymbol{\Sigma e}_i = \lambda_i \boldsymbol{e}_i \Rightarrow \text{Cov}(X_k, Y_i) = \boldsymbol{a}_k' \lambda_i \boldsymbol{e}_i = \lambda_i e_{ik}$, bên cạnh đó, $\text{Var}(Y_i) = \lambda_i$ và $\text{Var}(X_k) = \sigma_{kk}$ kéo theo:

$$\rho_{Y_i, X_k} = \frac{\text{Cov}(X_k, Y_i)}{\sqrt{\text{Var}(Y_i)}\sqrt{\text{Var}(X_k)}} = \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \quad i, k = 1, 2, \ldots, p$$

Từ đây ta có điều phải chứng minh.

Ví dụ 1.1 Tính các thành phần chính của tổng thể.

Giả sử $\boldsymbol{X}' = [X_1, X_2, X_3]$ là véc tơ có ma trận hiệp phương sai là

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

Lời giải:

Có thể thấy rằng 3 cặp giá trị riêng và véc tơ riêng của $\boldsymbol{\Sigma}$ là

$$\begin{aligned} &\lambda_1 = 5.83;\ \boldsymbol{e}_1 = [0.383; -0.924; 0]', \\ &\lambda_2 = 2.00;\ \boldsymbol{e}_2 = [0; 0; 1]', \\ &\lambda_3 = 0.17;\ \boldsymbol{e}_3 = [0.924; 0.383; 0]'. \end{aligned}$$

Ba thành phần chính sẽ là

$$\begin{aligned} Y_1 &= \boldsymbol{e}_1' \boldsymbol{X} = 0.383X_1 - 0.924X_2, \\ Y_2 &= \boldsymbol{e}_2' \boldsymbol{X} = X_3, \\ Y_3 &= \boldsymbol{e}_3' \boldsymbol{X} = 0.924X_1 + 0.383X_2. \end{aligned}$$

Để minh hoạ cho kết quả 2.1 ta xét:

$$\begin{aligned} \text{Var}(Y_1) &= \text{Var}(0.383X_1 - 0.924X_2) \\ &= (0.383)^2\text{Var}(X_1) + (-0.924)^2\text{Var}(X_2) \\ &\quad +2(0.383)(-0.924)\text{Cov}(X_1, X_2) \\ &= 0.147 \times 1 + 0.854 \times 5 - 0.708 \times (-2) \\ &= 5.83 = \lambda_1. \\ \text{Cov}(Y_1, Y_2) &= \text{Cov}(0.383X_1 - 0.924X_2, X_3) \\ &= 0.383 * \text{Cov}(X_1, X_3) - 0.924 * \text{Cov}(X_2, X_3) = 0. \end{aligned}$$

Minh hoạ kết quả 2.2 ta có:

$$\sigma_{11} + \sigma_{22} + \sigma_{33} = \lambda_1 + \lambda_2 + \lambda_3 = 8.$$

Tỷ lệ của thành phần chính thứ nhất trên tổng số các phương sai là

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{5.83}{5.83 + 2.00 + 0.17} = 0.73.$$

Hơn nữa, tỷ lệ của tổng thành phần chính thứ nhất và thứ hai trên tổng số các phương sai

$$\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\lambda_3}=\frac{5.83+2}{8}=0.98.$$

Trong trường hợp này, các thành phần Y_1, Y_2 có thể thay thế cho các biến ban đầu với dữ liệu bị mất là ít nhất.

Khi đó, *với thành phần chính thứ nhất* sử dụng kết quả 2.3, ta được:

$$\begin{aligned}\rho_{Y_1,X_1} &= \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}}=\frac{0.383\sqrt{5.83}}{\sqrt{1}}=0.925,\\ \rho_{Y_1,X_2} &= \frac{e_{12}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}}=\frac{-0.924\sqrt{5.83}}{\sqrt{5}}=-0.998.\end{aligned}$$

Biến X_2, với hệ số -0.924, có trọng số lớn nhất trong thành phần chính Y_1. Đồng thời, X_2 cũng có tương quan lớn nhất (theo giá trị tuyệt đối) với Y_1. Tương quan giữa X_1 và Y_1 là 0.925, gần bằng giá trị tương quan của X_2, cho thấy cả hai biến đều gần như quan trọng như nhau đối với thành phần chính thứ nhất. Tuy nhiên, độ lớn tương đối của các hệ số của X_1 và X_2 chỉ ra rằng X_2 đóng góp nhiều hơn vào việc xác định Y_1 so với X_1. Do cả hai hệ số đều có giá trị đủ lớn và mang dấu trái ngược nhau, có thể lập luận rằng cả hai biến đều hỗ trợ lẫn nhau trong việc diễn giải thành phần chính Y_1.

Cuối cùng,

$$\rho_{Y_2,X_1}=\rho_{Y_2,X_2}=0 \quad \text{và} \quad \rho_{Y_2,X_3}=\frac{\sqrt{\lambda_2}}{\sqrt{\sigma_{33}}}=\frac{\sqrt{2}}{\sqrt{2}}=1. \quad \text{(đúng như mong đợi)}$$

Các tương quan còn lại có thể bỏ qua, vì thành phần thứ ba không quan trọng.

Xét các thành phần chính được suy ra từ các biến ngẫu nhiên phân phối chuẩn nhiều chiều

Giả sử $\mathbf{X}$ được phân phối như $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Chúng ta biết từ (4-7) phương trình elip tâm tại $\boldsymbol{\mu}$:

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

với các trục là $\pm c\sqrt{\lambda_i}\mathbf{e}_i$, $i=1,2,\ldots,p$, trong đó $(\lambda_i, \mathbf{e}_i)$ là các cặp giá trị riêng-vector riêng của $\boldsymbol{\Sigma}$. Một điểm nằm trên trục thứ i của elipsoid sẽ có tọa độ tỷ lệ với $\mathbf{e}_i'=[e_{i1}, e_{i2}, \ldots, e_{ip}]$ trong hệ tọa độ có gốc tại $\boldsymbol{\mu}$ và các trục song song với các trục ban đầu $x_1, x_2, \ldots, x_p$. ta đặt $\boldsymbol{\mu}=0$ trong lập luận tiếp theo.

Lấy kết quả mục 2.3, chúng ta có thể viết:

$$c^2=x'\Sigma^{-1}x=\frac{1}{\lambda_1}({e_1}'x)^2+\frac{1}{\lambda_2}({e_2}'x)^2+\ldots+\frac{1}{\lambda_p}({e_p}'x)^2,$$

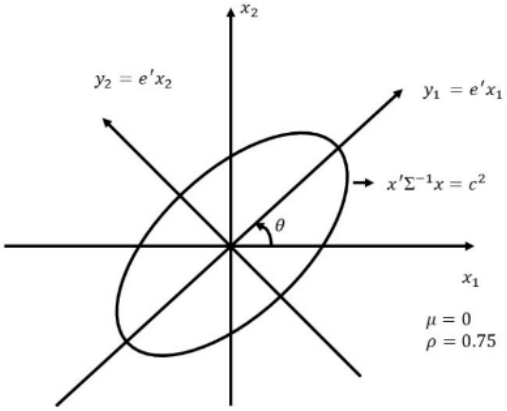
ở đó ${e_1}'x, {e_2}'x, .., {e_p}'x$được xem như các thành phần chính của X.

Đặt $y_1={e_1}'x, y_2={e_2}'x, .., y_p={e_p}'x$, chúng ta có:

$$c^2=\frac{1}{\lambda_1}(y_1)^2+\frac{1}{\lambda_2}(y_2)^2+\ldots+\frac{1}{\lambda_p}(y_p)^2.$$

Và phương trình này được xác định bằng một elipsoid trong một hệ trục tọa độ với các trực $y_1, y_2, ..., y_p$ nằm theo hướng của $e_1, e_2, .., e_p$ tương đương. Nếu λ_1 là trị riêng lớn nhất, thì trục chính nằm theo hướng e_1. Các trục nhỏ còn lại nằm theo hướng xác định bởi $e_2, ..e_p$.

Một elipsoid mật độ không đổi và các thành phần chính cho vectơ chuẩn ngẫu nhiên hai biến với $\boldsymbol{\mu} = 0$ và $\rho = 0.75$ được biểu diễn ở hình 2.1. Ta thấy rằng các thành phần chính được thu được bằng cách xoay hệ trục tọa độ ban đầu qua một góc θ cho đến khi chúng trùng khớp với các trục của elipsoid ban đầu. Kết quả này cũng đúng cho $p > 2$ chiều.



Hình 2.1: Ellipse mật độ không đổi $x'\mu^{-1}x = c^2$ và các thành phần y_1, y_2 với vector X chuẩn ngẫu nhiên hai biến có trung bình bằng 0.

3 Các thành phần chính thu được từ biến đã chuẩn hóa

3.1 Biến chuẩn hóa

Chuẩn hóa các biến giúp cho ta:

- Chia đều độ quan trọng của các biến được quan sát.
- Giúp tăng tốc độ tính toán và giảm bộ nhớ.
- Tránh vấn đề các biến không cùng thứ nguyên.

Công thức chuẩn hóa biến X thành biến Z:

$$Z = \frac{(X - \mu)}{\sqrt{\sigma}}, \quad \text{khi đó } Z \sim N(0, 1).$$

Với mỗi biến X_i ta chuẩn hóa nó bằng một biến không có thứ nguyên Z_i:

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}}$$
$$Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}}$$
$$\dots$$
$$Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}}$$

Đặt $\boldsymbol{Z} = (Z_1, ..., Z_p)'$ ta có:

$$\boldsymbol{Z} = \boldsymbol{V}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \tag{10}$$

trong đó $\boldsymbol{V} = diag(\sigma_{11}, ..., \sigma_{pp})$.

3.2 Bài toán

Với các biến được chuẩn hóa Z_i, ta có:

- $E(\boldsymbol{Z}) = \boldsymbol{0}$ và $\text{Cov}(\boldsymbol{Z}) = \boldsymbol{\rho}$
- Phương sai tổng thể của các **thành phần chính** vẫn bằng phương sai tổng thể của các **biến đã chuẩn hóa** và bằng **p**:

$$\sum_{i=1}^{p} \text{Var}(Y_i) = \sum_{i=1}^{p} \text{Var}(Z_i) = p$$

- Hệ số tương quan của **thành phần chính thứ** i so với **biến chuẩn hóa thứ** k được tính theo công thức:

$$\rho_{Y_i;Z_k} = e_{ik}\sqrt{\lambda_i}, \quad i, k = 1, 2, \dots, p$$

Chứng minh: $E(\boldsymbol{Z}) = \boldsymbol{0}$

Ta có:

$$\bar{Z} = \frac{1}{n}\mathbf{Z}'1 = \frac{1}{n}\begin{bmatrix} \sum_{j=1}^{n} \frac{x_{j1}-\mu_1}{\sqrt{\sigma_{11}}} \\ \sum_{j=1}^{n} \frac{x_{j2}-\mu_2}{\sqrt{\sigma_{22}}} \\ \vdots \\ \sum_{j=1}^{n} \frac{x_{jp}-\mu_p}{\sqrt{\sigma_{pp}}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \boldsymbol{0}$$

$$\left(\frac{1}{n}\sum_{j=1}^{n} \frac{x_{ji}-\mu_i}{\sqrt{\sigma_{ii}}} = \frac{\frac{1}{n}\sum_{j=1}^{n}(x_{ji}-\mu_i)}{\sqrt{\sigma_{ii}}} = \frac{\frac{1}{n}(\sum_{j=1}^{n} x_{ji} - n.\mu_i)}{\sqrt{\sigma_{ii}}} = \frac{\mu_i - \frac{n}{n}\mu_i}{\sqrt{\sigma_{ii}}} = 0\right)$$

Chứng minh: $Cov(\boldsymbol{Z}) = \boldsymbol{\rho}$

Ta có:

$$Cov(Z_i, Z_j) = \mathbb{E}[(Z_i - \mathbb{E}[Z_i])(Z_j - \mathbb{E}[Z_j])]$$

Do $\mathbb{E}[Z_i] = 0 \implies Cov(Z_i, Z_j) = \mathbb{E}[Z_i Z_j]$
Thay $Z_i = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}$ và $Z_j = \frac{X_j - \mu_j}{\sqrt{\sigma_{jj}}}$:

$$\Leftrightarrow Cov(Z_i, Z_j) = \mathbb{E}[\frac{(X_i - \mu_i)(X_j - \mu_j)}{\sqrt{\sigma_{ii}\sigma_{jj}}}]$$
$$\Leftrightarrow Cov(Z_i, Z_j) = \frac{\mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$
$$\Leftrightarrow Cov(Z_i, Z_j) = \frac{Cov(X_j, X_j)}{\sqrt{\sigma_{ii}\sigma_{jj}}} = \rho$$

Chứng minh: $\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \sum_{i=1}^{p} \mathrm{Var}(Z_i) = p$

Thành phần chính thứ i:

$$Y_i = \boldsymbol{e}_i' \boldsymbol{Z} = e_{i1} Z_1 + e_{i2} Z_2 + ... + e_{ip} Z_p, \quad i = 1, 2, ..., p \tag{11}$$

với vector $\boldsymbol{Z}$ là các vector quan sát đã được chuẩn hoá, có ma trận hiệp phương sai $\boldsymbol{\rho}$, cặp trị riêng, vector riêng tương ứng là $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$ với $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$.

- Theo định nghĩa: $tr(\boldsymbol{\rho}) = \sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \sum_{i=1}^{p} \mathrm{Var}(Z_i) = p$
- Theo tính chất: $tr(\boldsymbol{\rho}) = \lambda_1 + \lambda_2 + ... + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(Y_i)$

$$\Rightarrow \sum_{i=1}^{p} \mathrm{Var}(Z_i) = \sum_{i=1}^{p} \mathrm{Var}(Y_i) = p \quad \text{(đpcm)}$$

Chứng minh: $\rho_{Y_i, Z_k} = e_{ik}\sqrt{\lambda_i}$

Hệ số tương quan giữa thành phần chính Y_i và thành phần X_j của véc tơ X là

$$\rho(Y_i, X_j) = \frac{\sqrt{\lambda_i} e_{ij}}{\sqrt{\sigma_{jj}}}$$

trong đó $\boldsymbol{e}_i = [e_{i1}, ..., e_{ip}]'$, $\forall i = 1, ...,$p.
Tương tự với $\boldsymbol{Z}$ đã được chuẩn hóa, ta có $\sigma_{ii} = 1 \qquad \forall i = 1, ..., p$. Ta suy ra được

$$\rho(Y_i, Z_j) = e_{ij}\sqrt{\lambda_i} \qquad i, j = 1, ..., p$$

Ta có được điều phải chứng minh.

Bằng cách thay $\boldsymbol{Z}$ cho $\boldsymbol{X}$, chúng ta thấy rằng tỷ lệ của tổng phương sai được giải thích bởi thành phần chính thứ k của $\boldsymbol{Z}$ xác định bởi

$$\begin{pmatrix} \text{Tỷ lệ của} \\ \text{phương sai tổng thể} \\ \text{(chuẩn hóa) do} \\ \text{thành phần chính thứ k} \end{pmatrix} = \frac{\lambda_k}{p} \qquad k = 1, 2, ..., p \tag{12}$$

trong đó λ_k là trị riêng của $\boldsymbol{\rho}$.

Ví dụ 1.2. Giả sử $\boldsymbol{X} = [X_1, X_2]'$ có ma trận hiệp phương sai

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

và ma trận tương quan tương ứng là

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$$

$\boldsymbol{\Sigma}$ có các cặp giá trị riêng và véc tơ riêng là

$$\lambda_1 = 100.16; \quad \boldsymbol{e}_1' = [0.04; 0.999]$$

$$\lambda_2 = 0.84; \quad \boldsymbol{e}_2' = [0.999; -0.04]$$

$\boldsymbol{\rho}$ có các cặp giá trị riêng và véc tơ riêng là

$$\lambda_1 = 1.4; \quad \boldsymbol{e}_1' = [0.707; 0.707]$$

$$\lambda_2 = 0.6; \quad \boldsymbol{e}_2' = [0.707; -0.707]$$

Các thành phần chính của $[X_1, X_2]'$ là

$$Y_1 = 0.040X_1 + 0.999X_2$$
$$Y_2 = 0.999X_1 - 0.040X_2$$

Các thành phần chính của $[Z_1, Z_2]'$ là

$$\begin{aligned} Y_1' &= 0,707Z_1 + 0.707Z_2 = 0.707\left(\frac{X_1 - \mu_1}{1}\right) + 0.707\left(\frac{X_2 - \mu_2}{10}\right) \\ &= 0.707(X_1 - \mu_1) + 0.0707(X_2 - \mu_2) \\ Y_2' &= 0,707Z_1 - 0.707Z_2 = 0.707\left(\frac{X_1 - \mu_1}{1}\right) - 0.707\left(\frac{X_2 - \mu_2}{10}\right) \\ &= 0.707(X_1 - \mu_1) - 0.0707(X_2 - \mu_2) \end{aligned}$$

Nhận xét: Các thành phần chính xuất phát từ $\boldsymbol{\Sigma}$ khác với $\boldsymbol{\rho}$. Hơn nữa, một tập hợp các thành phần chính không phải là một hàm đơn giản của tập hợp kia.

Thành phần chính thứ nhất:

$$\boxed{Y_1 = 0.040X_1 + 0.999X_2}$$

do có phương sai lớn, nên X_2 chiếm ưu thế hoàn toàn của thành phần chính đầu tiên được xác định từ $\boldsymbol{\Sigma}$. Hơn thế nữa, thành phần chính đầu tiên này giải thích một tỷ lệ

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{100.16}{101} = 0.992$$

của phương sai tổng thể.

Vì vậy, nếu giữ nguyên các biến ban đầu, thông tin từ X_1 sẽ bị xem nhẹ.

Thành phần chính thứ nhất của biến đã chuẩn hóa:

$$\boxed{Y_1' = 0.707(X_1 - \mu_1) + 0.0707(X_2 - \mu_2)}$$

Từ kết quả 1.2, ta tính được:

$$\rho_{Y_1,Z_1} = e_{11}\sqrt{\lambda_1} = 0.707\sqrt{1.4} = 0.837$$
$$\rho_{Y_1,Z_2} = e_{21}\sqrt{\lambda_1} = 0.707\sqrt{1.4} = 0.837$$

Nhận xét: Hai biến đầu vào có mức độ đóng góp cân bằng vào thành phần chính đầu tiên. Trong trường hợp này, thành phần chính đầu tiên giải thích một tỷ lệ

$$\frac{\lambda_1}{p} = \frac{1.4}{2} = 0.7$$

của phương sai tổng thể.

Nhận xét: Kết quả giải thích của thành phần chính giảm xuống phản ánh rằng việc chuẩn hóa giúp phân bổ lại phương sai giữa các thành phần chính một cách hợp lý hơn.

Ví dụ dữ liệu thực tế: Bộ dữ liệu từ các chỉ số của 20 quốc gia từ năm 2010-2019. Bảng chứa 200 bản ghi về 6 chỉ số bao gồm: GDP(USD), Dân số, Tuổi thọ kỳ vọng, Tỷ lệ thất nghiệp, Lượng khí thải CO2, Khả năng tiếp cận điện:

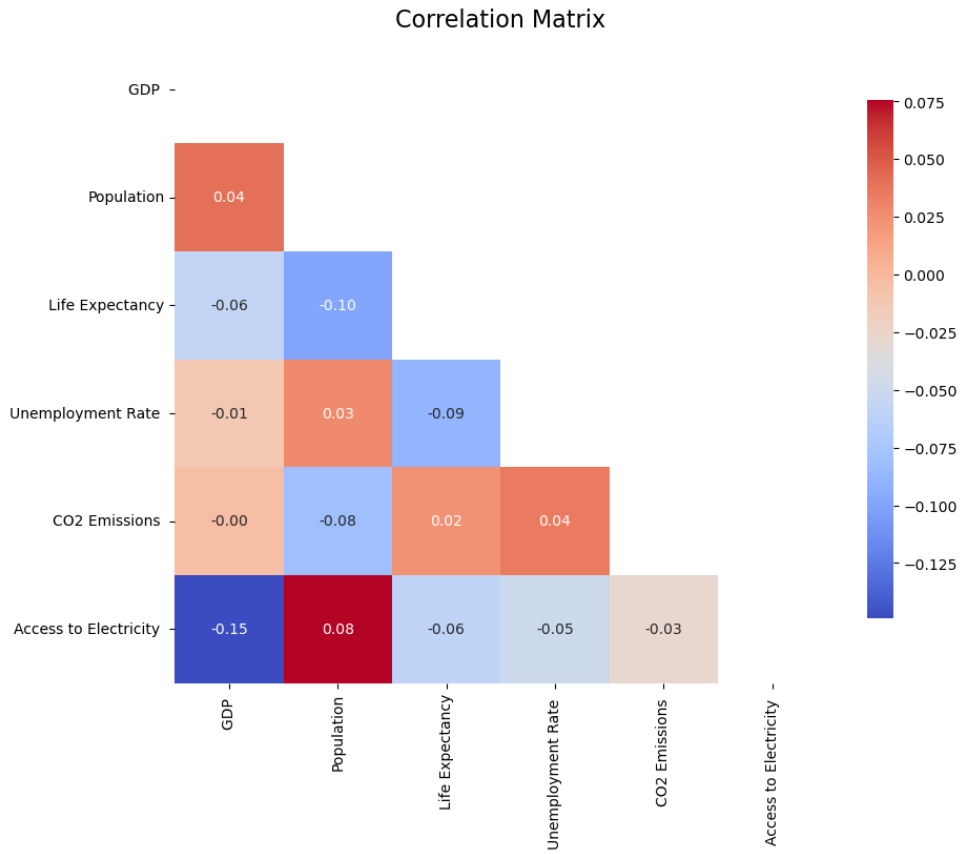
	Country	Year	GDP (USD)	Population	Life Expectancy	Unemployment Rate (%)	CO2 Emissions (metric tons per capita)	Access to Electricity (%)
0	Brazil	2010	1.493220e+12	8.290200e+08	66.7	3.81	10.79	76.76
1	Japan	2011	1.756270e+13	8.970100e+08	61.4	17.98	15.67	67.86
2	India	2012	1.642688e+13	6.698500e+08	69.1	16.02	2.08	81.08
3	Mexico	2013	1.189001e+13	1.138000e+08	80.1	6.26	19.13	53.46
4	India	2014	2.673020e+12	2.971000e+07	62.7	3.10	15.66	82.17
...	...	...	...	...	...	...	...	...
195	India	2015	1.110880e+12	6.727800e+08	78.5	22.30	11.36	66.04
196	Australia	2016	9.210290e+12	6.896500e+08	64.9	9.82	15.59	67.29
197	United States	2017	4.937150e+12	1.234820e+09	61.0	4.14	14.66	91.16
198	Canada	2018	1.088696e+13	2.625900e+08	76.5	2.28	7.25	59.21
199	United Kingdom	2019	4.525980e+12	5.354700e+08	73.7	17.29	16.68	65.48

200 rows × 8 columns

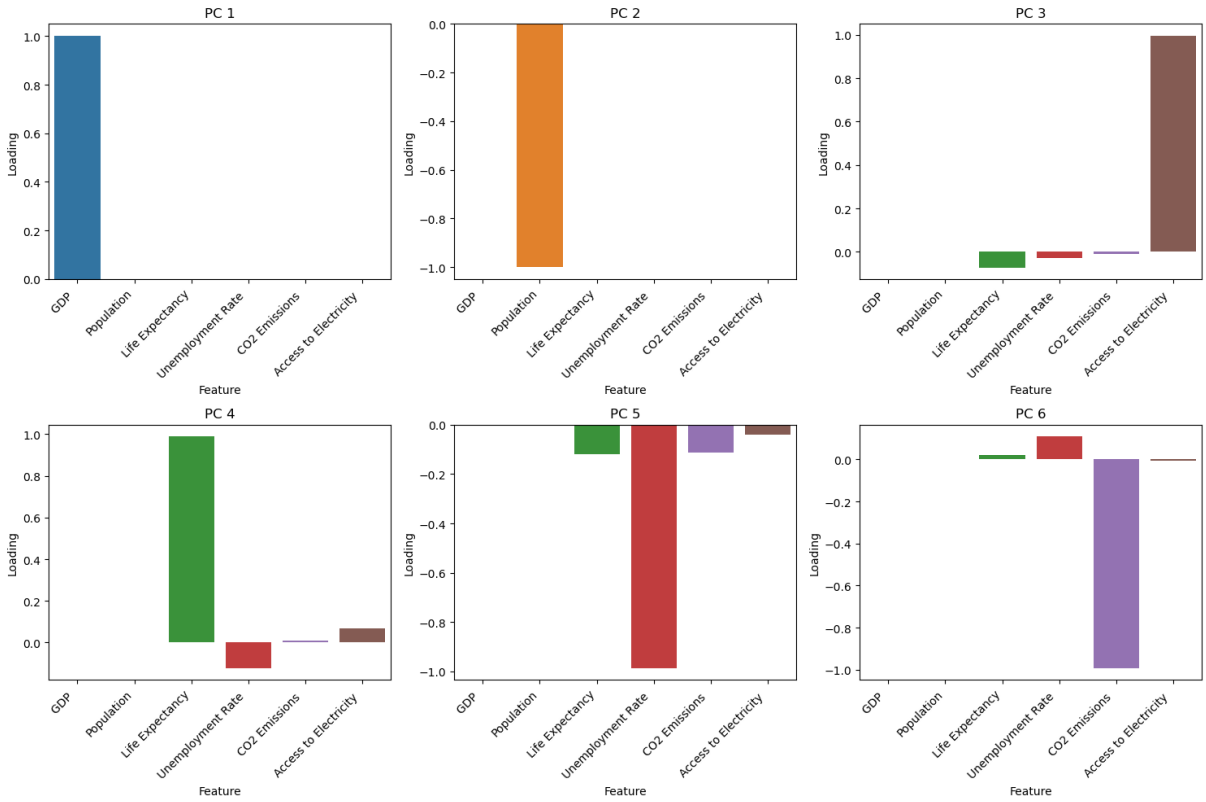
Tính toán phương sai của các đặc trưng (chỉ số):

```
GDP                          307770110479626309635932 16
Population                                 192717102653042784
Access to Electricity                                     219
Life Expectancy                                            96
Unemployment Rate                                          46
CO2 Emissions                                              33
```

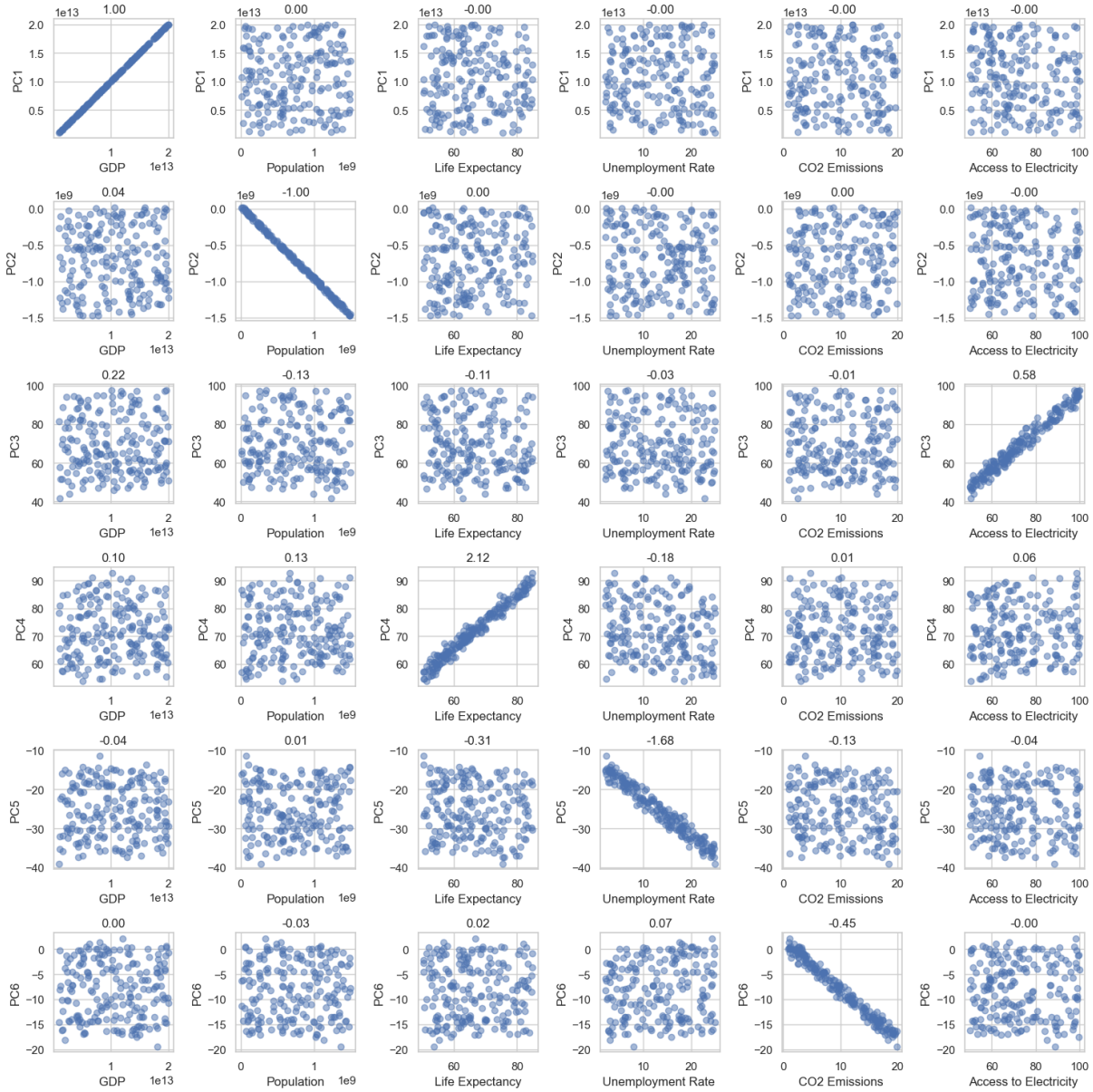
Tính ma trận tương quan:



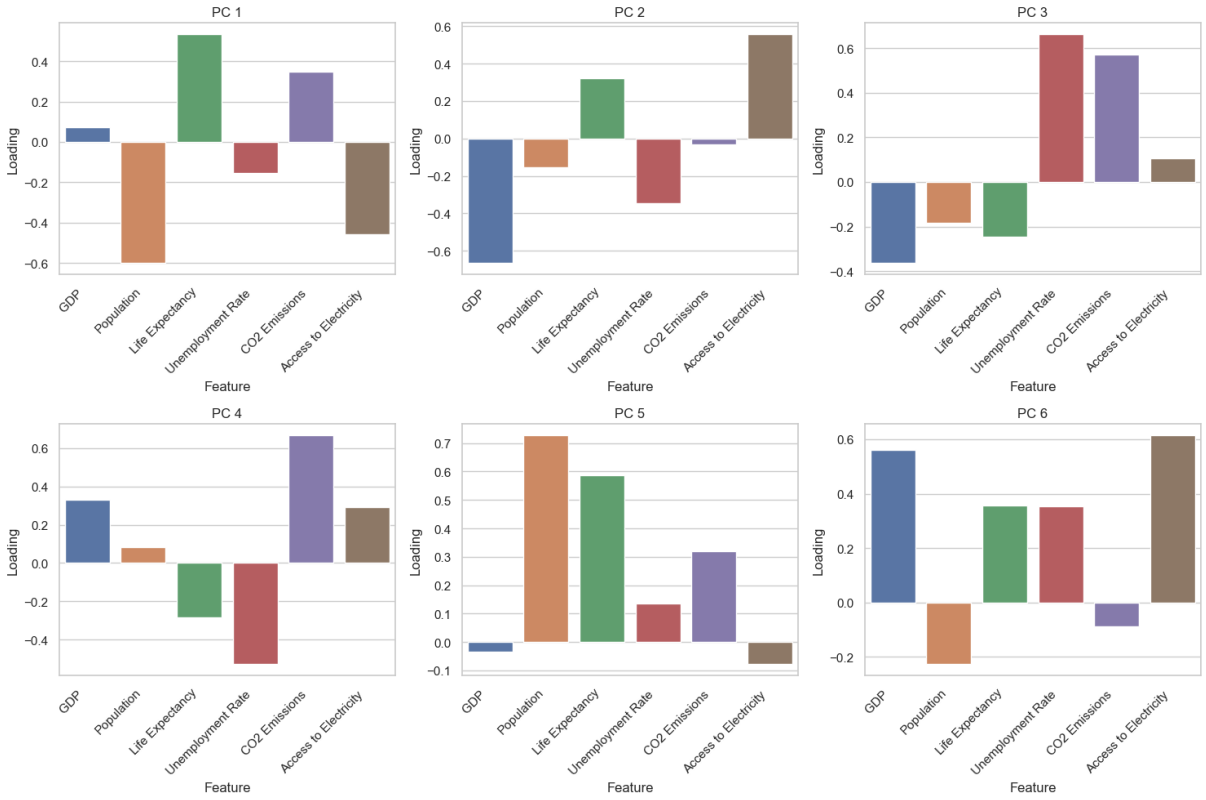
Áp dụng PCA cho các biến ban đầu không chuẩn hóa, các trọng số của từng đặc trưng được biểu diễn bởi các cột trong từng thành phần chính:



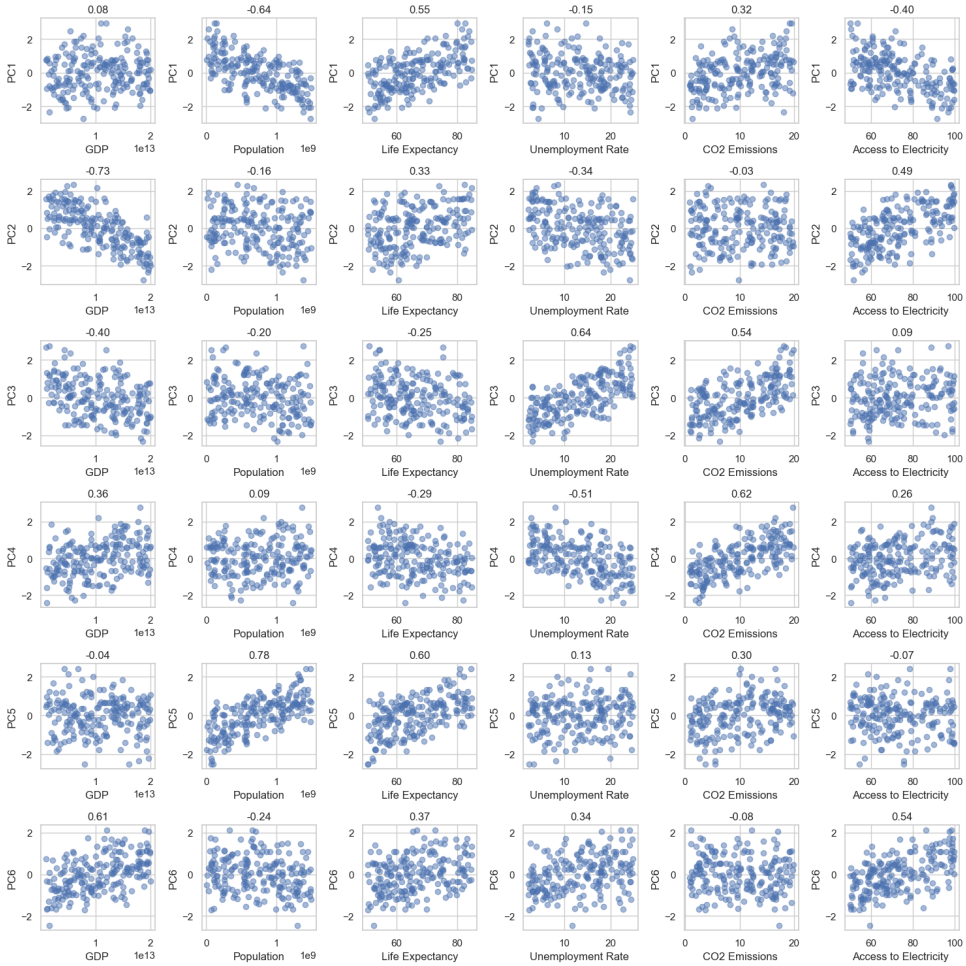
Mỗi thành phần chính thể hiện mạnh mẽ một đặc trưng ban đầu, các đặc trưng còn lại biểu diễn lượng thông tin rất thấp. Điều này làm cho việc phân tích thành phần chính trở nên không có nhiều ý nghĩa. Scatter plot thể hiện sự tương quan mạnh mẽ của cặp 1 thành phần chính với 1 đặc trưng ban đầu:



Tuy nhiên, sau khi chuẩn hóa, trong số của các đặc trưng trong từng thành phần chính không bị quá áp đảo bởi một đặc trưng bất kỳ nào. Điều này cho thấy một thành phần chính có khả năng biểu diễn thông tin của nhiều đặc trưng, chia đều độ quan trọng của các đặc trưng ban đầu:



Độ tương quan giữa thành phần chính với đặc trưng cũng vì thế mà không xuất hiện sự biểu diễn mạnh mẽ của một đặc trưng bất kỳ:



4 Phân tích thành phần chính cho ma trận hiệp phương sai với cấu trúc đặc biệt

Có một số ma trận hiệp phương sai và ma trận tương quan có một khuôn mẫu nhất định mà các thành phần chính của nó có thể được biểu diễn một cách đơn giản.

Trường hợp 1: Ma trận $\mathbf{\Sigma}$ là ma trận đường chéo

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{bmatrix} \tag{13}$$

Thiết lập ma trận hệ số $\boldsymbol{e}_i' = [0, ..., 0, 1, 0, ..., 0]$ với 1 ở vị trí thứ i, ta thấy rằng:

$$\begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sigma_{ii} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad hay \quad \mathbf{\Sigma}\boldsymbol{e}_i = \sigma_{ii}\boldsymbol{e}_i$$

Ví dụ: Xét ma trận đường chéo:

$$\mathbf{\Sigma} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix} \tag{14}$$

$$\mathbf{\Sigma}\boldsymbol{e}_1 = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \sigma_{11}\boldsymbol{e}_1 \quad \text{với } \sigma_{11} = 2$$

Tương tự:

$$\mathbf{\Sigma}\boldsymbol{e}_2 = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \\ 0 \\ 0 \end{bmatrix} = \sigma_{22}\boldsymbol{e}_2 \quad \text{với } \sigma_{22} = 3$$

Hay $(\sigma_{ii}, \boldsymbol{e}_i)$ là cặp trị riêng-vector riêng thứ i. Từ tổ hợp tuyến tính $Y_i = \boldsymbol{e}_i'\boldsymbol{X} = X_i$, tập hợp các thành phần chính chỉ là tổ hợp ban đầu của các biến ngẫu nhiên không tương quan. Đối với ma trận hiệp phương sai có mẫu là ma trận (13), ta sẽ không thu được gì bằng cách trích xuất các thành phần chính.

Trường hợp 2: Khi nghiên cứu các đặc tính về kích thước của các loài sinh vật sống, thường các biến như chiều dài, chiều rộng, chiều cao của một nhóm loài sinh vật sẽ có mối tương quan chặt chẽ với nhau.

Khi đó, ma trận $\mathbf{\Sigma}$ có dạng:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{bmatrix} \tag{15}$$

Ma trận tương quan:

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \tag{16}$$

Cũng là ma trận hiệp phương sai của các biến chuẩn hóa.

Trong ma trận (16), các biến $X_1, X_2, ..., X_p$ có tương quan một cách đồng đều bằng ρ.

Không khó để chỉ ra rằng ρ là trị riêng của ma trận tương quan (16) có thể được chia thành hai nhóm.

Khi ρ dương, giá trị riêng lớn nhất là:

$$\lambda_1 = 1 + (p-1)\rho \tag{17}$$

với véc tơ riêng tương ứng:

$$\boldsymbol{e}_1' = \left[\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}, \ldots, \frac{1}{\sqrt{p}}\right] \tag{18}$$

Khi đó, $p-1$ giá trị riêng còn lại là:

$$\lambda_2 = \lambda_3 = ... = \lambda_p = 1 - \rho$$

Các véc tơ riêng tương ứng như sau

$$\begin{aligned}
\boldsymbol{e}_2' &= \left[\frac{1}{\sqrt{1\times 2}}, \frac{-1}{\sqrt{1\times 2}}, 0, ..., 0\right] \\
\boldsymbol{e}_3' &= \left[\frac{1}{\sqrt{2\times 3}}, \frac{1}{\sqrt{2\times 3}}, \frac{-2}{\sqrt{2\times 3}}, 0, ..., 0\right] \\
&\vdots \qquad\qquad \vdots \\
\boldsymbol{e}_i' &= \left[\frac{1}{\sqrt{(i-1)\times i}}, ..., \frac{1}{\sqrt{(i-1)\times i}}, \frac{-(i-1)}{\sqrt{(i-1)\times i}}, 0, ..., 0\right] \\
&\vdots \qquad\qquad \vdots \\
\boldsymbol{e}_p' &= \left[\frac{1}{\sqrt{(p-1)\times p}}, ..., \frac{1}{\sqrt{(p-1)\times p}}, \frac{-(p-1)}{\sqrt{(p-1)\times p}}\right]
\end{aligned}$$

Ta có thành phần chính đầu tiên

$$Y_1 = \boldsymbol{e}_1' \boldsymbol{Z} = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} Z_i$$

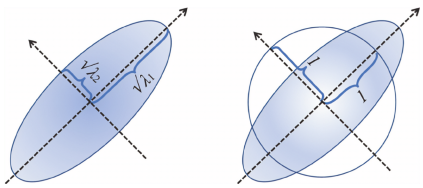
tỷ lệ với tổng của p biến được chuẩn hóa. Nó có thể được coi là một chỉ số có trọng số bằng nhau. Thành phần chính này giải thích một tỷ lệ:

$$\frac{\lambda_1}{p} = \frac{1 + (p-1)\rho}{p} = \rho + \frac{1-\rho}{p} \tag{19}$$

là tỷ lệ biểu diễn mức độ ảnh hưởng của p biến được chuẩn hóa lên sự biến thiên của tổng thể.

Nhận xét 1.9. Chúng ta thấy rằng $\dfrac{\lambda_1}{p} = \rho$ đối với ρ gần bằng 1 hoặc p lớn.
Ví dụ, nếu $\rho = 0.8$ và p = 5, thành phần chính đầu tiên giải thích 84% tổng biến thiên. Khi ρ gần 1, các thành phần cuối cùng $p-1$ cộng gộp lại đóng góp rất ít vào tổng biến thiên và có thể bị bỏ qua. Trong trường hợp đặc biệt này, chỉ giữ lại thành phần chính đầu tiên $Y_1 = \dfrac{1}{\sqrt{p}}[1, 1, ..., 1]\boldsymbol{X}$ là một phép đo về tổng kích thước, vẫn giải thích cùng một tỷ lệ (19) của tổng phương sai.

Nhận xét 1.10. Nếu các biến chuẩn hóa $Z_1, Z_2, ..., Z_p$ có phân phối chuẩn đa biến với ma trận hiệp phương sai cho trong (16), thì các ellipsoid của mật độ không đổi sẽ có "dạng điếu xì gà", với trục chính tỉ lệ với thành phần chính đầu tiên $Y_1 = \dfrac{1}{\sqrt{p}}[1, 1, ..., 1]\boldsymbol{X}$. Thành phần chính này là hình chiếu của Z lên đường đẳng giác $\mathbf{1}' = [1, 1, ..., 1]$ là đường nối các điểm có cùng giá trị của Y_1. Các trục phụ (và các thành phần chính còn lại) nằm trong các hướng đối xứng cầu vuông góc với trục chính (trục của Y_1 - thành phần chính đầu tiên).



Hình 4.1: Ellipsoid mật độ không đổi có dạng "điếu xì gà".

III Phân tích thành phần chính dựa trên mẫu

Bây giờ chúng ta đã có những lý thuyết cần thiết để nghiên cứu vấn đề tóm tắt sự biến thiên của n phép đo trên p biến bằng một vài tổ hợp tuyến tính được lựa chọn.

Giả sử dữ liệu ta thu được n biến ngẫu nhiên $x_1, x_2, ..., x_n$ từ một tổng thể p chiều nào với vector trung bình μ và ma trận hiệp phương sai Σ. Từ những dữ liệu này, ta tính được vector trung bình mẫu $\overline{x}$, ma trận hiệp phương sai mẫu S và ma trận hệ số tương quan mẫu R.

1 Sự thay đổi của mẫu theo thành phần chính

Ở phần này, ta sẽ xây dựng các tổ hợp tuyến tính giữa các thành phần của quan sát x_i sao cho các tổ hợp này không tương quan với nhau và có phương sai mẫu lớn nhất. Khi đó, những tổ hợp tuyến tính này được gọi là các thành phần chính của mẫu.
Với mỗi $a_i \in \mathbb{R}^{p\times 1}$, ta có tổ hợp tuyến tính:

$$a_i' x_j = a_{i1}' x_{j1} + a_{i2}' x_{j2} + ... + a_{ip}' x_{jp}, \;\; j = \overline{1,n}$$

có trung bình mẫu bằng $a_i'\overline{x}$ và phương sai mẫu bằng $a_i' S a_i$. Đồng thời cặp mẫu $(a_1' x_j, a_2' x_j)$ có hiệp phương sai mẫu bằng $a_1' S a_2$.

Để phân tích các thành phần chính của mẫu, ta cần xác định các tổ hợp tuyến tính có phương sai mẫu lớn nhất và không tương quan với nhau. Giống như đối với tổng thể, ta cũng sẽ hạn chế các vector hệ số a_i sao cho $a_i' a_i = 1$. Các tổ hợp tuyến tính đó được xác định như sau:
Gọi $\hat{y}_1, \hat{y}_2, ..., \hat{y}_k$ là thành phần chính mẫu thứ $1, 2, ..., k$ của mẫu x_i với $i = \overline{1,n}$ nếu:

- Thành phần chính mẫu thứ nhất $\hat{y}_1 = a_1' x_j$, mà ở đó a_1 làm cực đại phương sai mẫu $a_1' S a_1$ với điều kiện $a_1' a_1 = 1$.
- Thành phần chính mẫu thứ hai $\hat{y}_2 = a_2' x_j$, mà ở đó a_2 làm cực đại phương sai mẫu $a_2' S a_2$ với điều kiện $a_2' a_2 = 1$ và hiệp phương sai mẫu $a_1' S a_2 = 0$.
- ...
- Thành phần chính mẫu thứ k $\hat{y}_k = a_k' x_j$, mà ở đó a_k làm cực đại phương sai mẫu $a_k' S a_k$ với điều kiện $a_k' a_k = 1$ và hiệp phương sai mẫu $a_u' S a_k = 0 (u < k)$.

Thành phần chính đầu tiên làm cực đại phương sai mẫu $a_1' S a_1$ hoặc tương đương:

$$\frac{a_1' S a_1}{a_1' a_1} \tag{20}$$

- Theo kết quả chứng minh maximum của dạng toàn phương trên hình cầu đơn vị, giá trị cực đại là giá trị riêng $\hat{\lambda}_1$ lớn nhất đạt được cho sự lựa chọn $a_1 = \hat{e}_1$ của S.

- Chọn các a_i tiếp theo làm cực đại (20) với giả thuyết: $0 = a_i' S\hat{e}_k = a_i'\hat{\lambda}_k\hat{e}_k$, hoặc a_i trực giao với $\hat{e}_k$.

Các kết quả thu được

Nếu $S = [s_{ik}]$ là ma trận hiệp phương sai mẫu $p \times p$ với các cặp giá trị riêng và vector riêng $(\hat{\lambda}_1, \hat{e}_1), (\hat{\lambda}_2, \hat{e}_2), ..., (\hat{\lambda}_p, \hat{e}_p)$, thành phần chính mẫu thứ i được đưa ra bởi:

$$\hat{y}_i = \hat{e}_i' x = \hat{e}_{i1}x_1 + \hat{e}_{i2}x_2 + \ldots + \hat{e}_{ip}x_p \text{ với } i = \overline{1,p}$$

trong đó $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p \geq 0$ và x là một quan sát của vector ngẫu nhiên $X = [X_1, X_2, ...X_p]'$
Đồng thời,

- Phương sai mẫu $(\hat{y}_k) = \hat{\lambda}_k$, $k = \overline{1,p}$
- Hiệp phương sai mẫu $(\hat{y}_i, \hat{y}_k) = 0$, $i \neq k$
- Tổng phương sai mẫu $= \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + ... + \hat{\lambda}_p$
- Hệ số tương quan mẫu $r(\hat{y}_i, x_k) = \dfrac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}$, $i, k = \overline{1,p}$

Ta ký hiệu các thành phần chính mẫu bởi $\hat{y}_1, \hat{y}_2, ..., \hat{y}_p$ không phân biệt trường hợp chúng được tính từ S hay R. Các thành phần chính được tính từ S và R là khác nhau nhưng trong một ngữ cảnh cụ thể, ta sẽ nêu rõ ràng ma trận nào được sử dụng. Bên cạnh đó trong cả hai trường hợp này, ta đều ký hiệu $\hat{e}_i$ cho vector riêng của thành phần chính và $\hat{\lambda}_i$ cho phương sai của thành phần chính.
Bên cạnh đó các thành phần chính mẫu cũng có thể tính được từ $\hat{\Sigma} = S_n$. $\hat{\Sigma}$ có các giá trị riêng là $[(n-1)/n]\hat{\lambda}_i$ và các vector riêng $\hat{e}_i$ tương ứng. Vì vậy cả S và $\hat{\Sigma}$ đều cho cùng các thành phần chính mẫu và cùng tỷ lệ phương sai của thành phần chính mẫu thứ i là $\hat{\lambda}_i/(\hat{\lambda}_1 + \hat{\lambda}_2 + ... + \hat{\lambda}_p)$

Lưu ý

Các quan sát x_j được gọi là "quy tâm" bằng cách trừ $\overline{x}$. Điều này không ảnh hưởng đến ma trận hiệp phương sai mẫu S và cho ta thành phần chính mẫu thứ i là:

$$\hat{y}_i = \hat{e}_i'(x - \overline{x}),\ i = \overline{1,p}$$

với x là một quan sát bất kỳ. Cụ thể hơn, nếu ta muốn xét đến giá trị của thành phần chính mẫu thứ i thì:

$$\hat{y}_{ji} = \hat{e}_i'(x_j - \overline{x}),\ j = \overline{1,n}$$

là thành phần chính mẫu thứ i tương ứng với quan sát x_j. Khi đó ta có:

$$\overline{\hat{y}_i} = \frac{1}{n}\sum_{j=1}^{n}\hat{e}_i'(x_j - \overline{x}) = \frac{1}{n}\hat{e}_i'\left(\sum_{j=1}^{n}(x_j - \overline{x})\right) = \frac{1}{n}\hat{e}_i' 0 = 0$$

Lúc này, từng thành phần chính thứ i có vector trung bình mẫu bằng 0 và phương sai mẫu vẫn bằng $\hat{\lambda}_i$

Ví dụ 2.1 Khi thực hiện điều tra dân số ở vùng Madison và Wisconsin, người ta thu được các số liệu về năm biến số kinh tế xã hội lần lượt là: tổng dân số (đơn vị: nghìn), trình độ đại học trở lên (đơn vị: phần trăm), lao động trên 16 tuổi (đơn vị: phần trăm), lao động làm việc cho chính phủ (đơn vị: phần trăm), giá trị nhà ở trung bình (đơn vị: \$100.000). Thu được tổng cộng 61 quan sát và chúng được thể hiện qua thống kê sau:

$$\overline{x}' = \begin{bmatrix} 4.47, & 3.96, & 71.42, & 26.91, & 1.64 \end{bmatrix}$$

và

$$S = \begin{bmatrix} 3.397 & -1.102 & 4.306 & -2.078 & 0.027 \\ -1.102 & 9.673 & -1.513 & 10.953 & 1.203 \\ 4.306 & -1.513 & 55.626 & -28.937 & -0.044 \\ -2.078 & 10.953 & -28.937 & 89.067 & 0.957 \\ 0.027 & 1.203 & -0.044 & 0.957 & 0.319 \end{bmatrix}$$

$\Longrightarrow$ Liệu có thể tóm tắt sự biến đổi mẫu bằng một hay hai thành phần chính hay không?

Đầu tiên ta thực hiện tính các vector riêng $\hat{e}_i$ ứng với các thành phần chính $\hat{y}_i$ và hệ số tương quan giữa thành phần chính $\hat{y}_i$ và biến x_k qua bảng sau đây:

Biến	$\hat{e}_1(r_{\hat{y}_1,x_k})$	$\hat{e}_2(r_{\hat{y}_2,x_k})$	$\hat{e}_3$	$\hat{e}_4$	$\hat{e}_5$
Tổng dân số	-0.039(-.22)	0.071(.24)	0.188	-0.977	-0.058
Trình độ đại học trở lên	0.105(.35)	0.130(.26)	0.091	0.171	-0.139
Lao động trên 16 tuổi	-0.492(-.68)	0.864(.73)	0.046	-0.091	0.005
Lao động làm việc cho chính phủ	0.863(.95)	0.480(.32)	0.153	-0.030	0.007
Giá trị nhà trung bình	0.009(.16)	0.015(.17)	0.125	0.082	0.989
Phương sai	107.02	39.67	8.37	2.87	0.15
Phần trăm tích lũy của tổng phương sai	67.7	92.8	98.1	99.9	100.0

Từ bảng trên, ta có một số nhận xét sau:

- Thành phần chính đầu tiên giải thích được 67.7% tổng phương sai mẫu.
- Hai thành phần chính đầu tiên giải thích được 92.8% tổng phương sai mẫu.

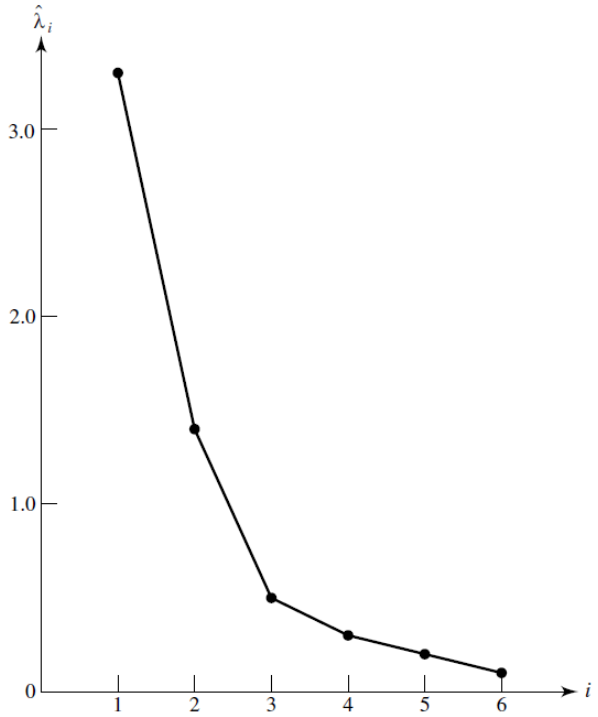
Từ đó, ta thấy rằng sự biến đổi của mẫu này được tổng quát khá tốt thông qua hai thành phần chính đầu tiên (92.8%). Như vậy, ta có thể giảm từ 61 quan sát trên 5 biến xuống 61 quan sát trên 2 thành phần chính này.

2 Số lượng thành phần chính

Sau khi hiểu được cách tìm ra các thành phần chính của một mẫu, một câu hỏi khác đặt ra là: Ta nên giữ lại bao nhiêu thành phần chính? Tuy nhiên không có một câu trả lời chính xác cho câu hỏi này trong mọi trường hợp nhưng trong một trường hợp cụ thể, ta có thể xét đến một số yếu tố sau:

- Tổng phương sai mẫu được giải thích.
- Giá trị riêng (phương sai của các thành phần mẫu).
- Sự giải thích (ý nghĩa) của mỗi thành phần.
- Thành phần với giá trị tương ứng tương đối nhỏ (so với các giá trị riêng còn lại) thì có thể coi là không quan trọng.

Một công cụ trực quan giúp ta xác định số lượng các thành phần chính cần giữ lại là **đồ thị scree**.



(a)

Hình 2.1: Đồ thị screen

Trên đồ thị scree, các giá trị riêng được biểu diễn ở trục tung và số lượng thành phần chính được thể hiện ở trục hoành. Các thành phần chính được sắp xếp tương ứng theo thứ tự giá trị riêng giảm dần.

Để xác định số lượng thành phần chính cần giữ lại, ta tìm "điểm gãy" trên đồ thị:

- Điểm gãy là điểm mà tại đó đường biểu diễn bắt đầu "chững lại", độ dốc không thay đổi đột ngột.
- Những thành phần ở bên trái của điểm gãy sẽ được giữ lại.
- Những thành phần ứng với giá trị riêng tương đối nhỏ (so với các giá trị riêng còn lại) có thể bỏ qua.

Quay lại ví dụ ở Hình **??**, áp dụng các quy tắc ở trên ta có thể chọn điểm gãy tương ứng với thành phần chính thứ hai (hoặc ba). Có nghĩa là ta sẽ chỉ giữ lại thành phần chính thứ nhất (hoặc giữ lại thành phần chính thứ nhất hoặc thứ hai).

Kết luận

Số lượng thành phần chính cần được giữ lại:

- Dựa vào thông tin muốn giữ lại.
- Với phương sai càng lớn tức là dữ liệu có độ phân tán cao, thể hiện thông tin lớn.
- Vì trong hệ trục tọa độ, tổng phương sai của dữ liệu là như nhau và bằng tổng giá trị riêng của ma trận hiệp phương sai. Vậy có thể coi biểu thức:

$$r_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}, \quad (p\text{: số chiều gốc})$$

là lượng thông tin được giữ lại, khi số chiều dữ liệu mới sau PCA là k.

3 Biểu diễn hình học thành phần chính của mẫu

Cho vectơ ngẫu nhiên gốc $X' = [X_1, X_2, \ldots, X_p]$ tuân theo phân phối chuẩn $N_p(\mu, \Sigma)$. Giả sử có ma trận dữ liệu gồm n quan sát x. Với :

- $\hat{e_i'}$ là vectơ đơn vị của thành phần chính thứ i,

- x_j là quan sát thứ j,

- $\bar{x}$ là trung bình mẫu của các quan sát. Ta có

$$\hat{y_{ij}} = \hat{e_i'}(x_j - \bar{x}), \quad j = 1, 2, \ldots, n$$

là thành phần chính của mẫu cụ thể ứng với các thành phần chính tổng thể Y_i.

Các thành phần chính tổng thể Y_i có thể được biểu diễn dưới dạng:

$$Y_i = \hat{e_i'}(X - \mu)$$

có phân phối chuẩn $N_p(0, \lambda)$, trong đó $\lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ là ma trận chéo với các giá trị riêng $\lambda_1, \lambda_2, \ldots, \lambda_p$ của ma trận hiệp phương sai Σ. Các giá trị riêng này tương ứng với các phương sai của các thành phần chính và (λ_i, e_i) là cặp giá trị riêng và vectơ riêng của ma trận hiệp phương sai Σ, trong đó λ_i là giá trị riêng của ma trận hiệp phương sai và e_i là vectơ riêng tương ứng với giá trị riêng đó.

Giả thuyết

Ta có phương trình xác định mặt mức, trên đó hàm mật độ $f(X)$ của biến ngẫu nhiên gốc X là hằng số, là một hyperellipsoid tâm tại μ và các trục $\pm c\sqrt{\lambda_i} e_i$, với $i = 1, 2, \ldots, p$. Phương trình xác định mặt mức là:

$$(X - \mu)' \Sigma^{-1} (X - \mu) = c^2$$

Giả sử ta có một mẫu ngẫu nhiên kích thước n và thực hiện các quan sát cụ thể thu được dữ liệu $\mathbf{x}_{n\times p}$ xấp xỉ μ bởi $\bar{x}$ và ma trận hiệp phương sai Σ được xấp xỉ bởi S. Với điều kiện S xác định dương, mặt mức của n dữ liệu trong không gian p-chiều được xác định bởi phương trình:

$$(\mathbf{x}-\bar{x})'S^{-1}(\mathbf{x}-\bar{x}) = c^2$$

Nhận xét

* Khi biểu diễn dữ liệu trên biểu đồ phân tán, về mặt hình học, có thể hiểu vẽ thành n điểm trong không gian p chiều. Sau đó dữ liệu được biểu diễn trong các tọa độ mới là hình chiếu xuống các trục của mặt mức xác định bởi phương trình

$$(\mathbf{x}-\bar{x})'S^{-1}(\mathbf{x}-\bar{x}) = c^2$$

*Phương trình

$$(\mathbf{x}-\bar{x})'S^{-1}(\mathbf{x}-\bar{x}) = c^2$$

xác định một hyperellipsoid có tâm ở $\bar{x}$, các trục được xác định bởi các vector riêng của ma trận S^{-1} .

* Nếu (λ_i, e_i) là trị riêng, vectơ riêng của ma trận S, thì $\left(\lambda_i^{-1}, e_i\right)$ là trị riêng, vectơ riêng của ma trận S^{-1}. Vì vậy, việc xây dựng và biểu diễn hình học dựa trên ma trận S không làm mất tính tổng quát.

* Như vậy, dữ liệu được biểu diễn trong hệ tọa độ mới với các trục là các thành phần chính của mẫu $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$, nằm dọc theo hướng các trục của hyperellipsoid.

* Với việc xấp xỉ ma trận Σ bằng ma trận hiệp phương sai mẫu S, độ dài trục của hyperellipsoid tỷ lệ thuận với $\sqrt{\hat{\lambda}_i}$, $i = 1, 2, \ldots, p$. Trong đó $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ là các trị riêng của ma trận S.

* Nếu λ_1 là trị riêng lớn nhất thì trục chính của hệ tọa độ mới nằm theo hướng của e_1. Các trục nhỏ còn lại nằm theo các hướng xác định bởi các vectơ $e_2, e_3, \ldots, e_p$ tương ứng.

* Do e_i là các vectơ đơn vị, độ dài vectơ thành phần chính thứ i là:

$$|\hat{y}_i| = |\hat{e_i'}(x-\bar{x})|$$

điều này cho biết độ dài hình chiếu của vectơ $(x-\bar{x})$ trên trục $\hat{e_i'}$.

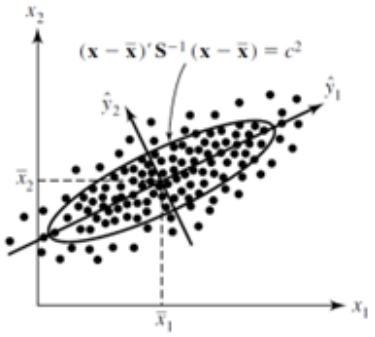
* Do đó, việc biểu diễn các thành phần chính của mẫu có thể được xem như là kết quả của việc dịch gốc của hệ trục tọa độ ban đầu sang $\bar{x}$ và quay các trục tọa độ một góc Φ sao cho chúng trùng với các trục của hyperellipsoid hoặc có thể hiểu là các hướng có phương sai lớn nhất.

Chú ý:

*Trong trường hợp các giả định về phân phối chuẩn của biến ngẫu nhiên không được thỏa mãn, biểu đồ phân tán của mẫu có thể sai lệch với hình elip. Tuy nhiên, ta vẫn có thể sử dụng giá trị riêng của ma trận hiệp phương sai mẫu S để tìm các thành phần chính của mẫu này.

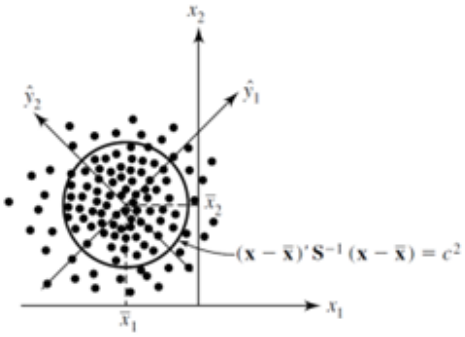
*Trong thực tế khi làm việc với bộ dữ liệu kích thước lớn, rất khó để xác định biến ngẫu nhiên quan sát có tuân theo phân phối chuẩn hay không. Do đó, ta cần các đánh giá, xử lý dữ liệu hoặc sử dụng định lý giới hạn trung tâm để xấp xỉ phân phối chuẩn và có thể sử dụng các kết quả đã trình bày ở trên.

Phân tích qua ví dụ



(a)

Hình 1: đồ thị phân tán với trường hợp trường $\lambda_1 > \lambda_2$ Hai thành phần chính của mẫu đường mức nằm dọc theo các trục của hình elip mà có thể (vuông góc với nhau và có phương sai (bao lớn nhất)



Hình 2 : đồ thị phân tán với hợp $\lambda_1 = \lambda_2$ Các trục của không xác định duy nhất hai đường vuông góc bất kì gồm cả trục gốc)

Nhận xét

*Khi các giá trị riêng của ma trận hiệp phương sai S gần bằng nhau, đường mức có dạng gần với đường tròn và các thành phần chính của mẫu có thể nằm theo những hướng vuông góc bất kỳ, kể cả hướng của các trục tọa độ ban đầu. Khi đó, sự biến thiên của mẫu là đồng nhất theo mọi hướng. Điều này dẫn đến việc không thể biểu diễn dữ liệu tốt với mẫu nhỏ hơn p chiều.

* Nếu tồn tại một số trị riêng $\hat{\lambda}_i$ đủ nhỏ sao cho sự biến thiên theo hướng của những $\hat{e}_i'$ tương ứng là không đáng kể thì các thành phần chính đó có thể bỏ qua mà dữ liệu vẫn được biểu diễn tương đối đầy đủ thông qua các thành phần chính được giữ lại.

4 Thành phần chính của mẫu đã chuẩn hóa

Với n quan sát của biến ngẫu nhiên p chiều, ta xây dựng ma trận kích thước $n \times p$ của các giá trị quan sát sau khi được chuẩn hóa:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1' \\ \mathbf{z}_2' \\ \vdots \\ \mathbf{z}_n' \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}$$

trong đó:

$$\mathbf{z}_j = \mathbf{D}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}}) = \begin{bmatrix} \frac{x_{j1}-\bar{x}_1}{\sqrt{s_{11}}} \\ \frac{x_{j2}-\bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \frac{x_{jp}-\bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix}, \quad j = 1, 2, \ldots, n$$

với:

$$\mathbf{D}^{-1/2} = \operatorname{diag}\left(\frac{1}{\sqrt{s_{ii}}}\right), \quad i = 1, 2, \ldots, p$$

Như vậy, ta có:

$$Z = \begin{bmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{12}-\bar{x}_2}{\sqrt{s_{22}}} & \ldots & \frac{x_{1p}-\bar{x}_p}{\sqrt{s_{pp}}} \\ \frac{x_{21}-\bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{22}-\bar{x}_2}{\sqrt{s_{22}}} & \ldots & \frac{x_{2p}-\bar{x}_p}{\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1}-\bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{n2}-\bar{x}_2}{\sqrt{s_{22}}} & \ldots & \frac{x_{np}-\bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix}$$

Từ đây, ta tính được vectơ trung bình mẫu của dữ liệu sau khi chuẩn hóa:

$$\bar{z} = \frac{1}{n} 1' Z = \frac{1}{n} Z' 1 = \frac{1}{n} \begin{bmatrix} \sum_{j=1}^{n} \frac{x_{j1}-\bar{x}_1}{\sqrt{s_{11}}} \\ \sum_{j=1}^{n} \frac{x_{j2}-\bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \sum_{j=1}^{n} \frac{x_{jp}-\bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} = 0$$

Tính được ma trận hiệp phương sai

$$S_z = \frac{1}{n-1}\left(Z - \frac{1}{n}11'Z\right)'\left(Z - \frac{1}{n}11'Z\right)$$

$$= \frac{1}{n-1}\left(Z - 1\bar{z}'\right)'\left(Z - 1\bar{z}'\right)$$

$$= \frac{1}{n-1}Z'Z$$

$$= \frac{1}{n-1}\begin{bmatrix} (n-1)s_{11} & \frac{(n-1)s_{12}}{\sqrt{s_{11}s_{22}}} & \cdots & \frac{(n-1)s_{1p}}{\sqrt{s_{11}s_{pp}}} \\ \frac{(n-1)s_{12}}{\sqrt{s_{11}s_{22}}} & (n-1)s_{22} & \cdots & \frac{(n-1)s_{2p}}{\sqrt{s_{22}s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{(n-1)s_{1p}}{\sqrt{s_{11}s_{pp}}} & \frac{(n-1)s_{2p}}{\sqrt{s_{22}s_{pp}}} & \cdots & (n-1)s_{pp} \end{bmatrix}$$

$$= R$$

Với $z_1, z_2, \ldots, z_n$ là các quan sát đã được chuẩn hóa với ma trận hiệp phương sai R như công thức ở trên, ta có thành phần chính mẫu thứ i là:

$$\hat{y}_i = e_i'z = e_{i1}z_1 + e_{i2}z_2 + \cdots + e_{ip}z_p, \quad i = 1, 2, \ldots, p$$

trong đó (λ_i, e_i) là cặp giá trị riêng - vectơ riêng thứ i của R với $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ và z là một quan sát bất kỳ trong n các quan sát đã được chuẩn hóa ở trên.

Hơn nữa, ta cũng thu được:

- Phương sai mẫu của $\hat{y}_i$ bằng $\lambda_i, \quad i = 1, 2, \ldots, p$
- Hiệp phương sai mẫu $\hat{y}_i, \hat{y}_k$ bằng 0, $\quad \forall i \neq j$

Từ đây, ta tính được tổng phương sai mẫu đã chuẩn hóa:

$$\text{tr}(\mathbf{R}) = \sum_{i=1}^{p} r_{ii} = p = \lambda_1 + \lambda_2 + \cdots + \lambda_p$$

và hệ số tương quan giữa thành phần chính mẫu $\hat{y}_i$ và biến đã chuẩn hóa z_k:

$$r(\hat{y}_i, z_k) = e_{ik}\sqrt{\lambda_i}, \quad i, k = 1, 2, \ldots, p$$

Từ công thức tính tổng phương sai mẫu, ta cũng có công thức tính tỷ lệ của tổng phương sai mẫu được giải thích bởi thành chính thứ i của mẫu là:

$$\frac{\lambda_i}{p}$$

Ví dụ: Tỷ suất lợi nhuận hàng tuần cho năm cổ phiếu(JPMorgan, Citibank, Wells Fargo, Royal Dutch Shell and ExxonMobil) được niêm yết trên sàn chứng khoán New York được xác định trong khoảng thời gian từ tháng 1 năm 2004 đến tháng 12 năm 2005.Tỉ suất lợi nhuận hàng tuần được định nghĩa là(giá đóng cửa tuần hiện tại -giá đóng cửa tuần trước)/(giá đóng cửa tuần trước).Dữ liệu được quan sát trong 103 tuần liên tiếp dường như được phân phối độc lập , nhưng tỷ suất lợi nhuận trên các cổ phiếu có mối tương quan , bởi vì người ta mong đợi các cổ phiếu có xu hướng di chuyển cùng nhau để đáp ứng các điều kiện kinh tế chung. Dữ liệu được thể hiện trong bảng dưới đây:

Week	**JP Morgan**	**Citibank**	**Wells Fargo**	**Royal Dutch Shell**	**ExxonMobil**
1	0.01303	0.00784	-0.00319	-0.04477	0.00522
2	0.00849	0.01669	-0.00621	0.01196	0.01349
3	-0.01792	-0.00864	0.01004	0	-0.00614
4	0.02156	-0.00349	0.01744	-0.02859	-0.00695
5	0.01082	0.00372	-0.01013	0.02919	0.04098
⋮	⋮	⋮	⋮	⋮	⋮
102	0.01039	-0.00266	0.00443	-0.00248	-0.01645
103	-0.01279	-0.01437	-0.01874	-0.00498	-0.01637

Bảng 1: Weekly data for various companies.

Đặt x_1, x_2, x_3, x_4, x_5 lần lượt biểu thị tỷ lệ hoàn vốn hằng tuần của 5 cổ phiếu, khi đó ta tính được:

$$\bar{x}' = [0.0011, 0.0007, 0.0016, 0.0040, 0.0040]$$

Chuẩn hóa dữ liệu:

$$z_1 = \frac{x_1 - \bar{x}_1}{\sqrt{s_{11}}}, \quad z_2 = \frac{x_2 - \bar{x}_2}{\sqrt{s_{22}}}, \ldots, \quad z_5 = \frac{x_5 - \bar{x}_5}{\sqrt{s_{55}}}$$

Khi đó ta tính được ma trận hiệp phương sai của dữ liệu sau khi đã chuẩn hóa:

$$R = \begin{bmatrix} 1.000 & 0.632 & 0.511 & 0.115 & 0.155 \\ 0.632 & 1.000 & 0.574 & 0.322 & 0.213 \\ 0.511 & 0.574 & 1.000 & 0.183 & 0.146 \\ 0.115 & 0.322 & 0.183 & 1.000 & 0.683 \\ 0.155 & 0.213 & 0.146 & 0.683 & 1.000 \end{bmatrix}$$

Các giá trị riêng và vectơ riêng chuẩn hóa tương ứng của $\mathbf{R}$ được tính như sau:

$$.\lambda_1 = 2.437, \mathbf{e}_1' = \begin{bmatrix} 0.469 & 0.532 & 0.465 & 0.387 & 0.361 \end{bmatrix}$$

$$\lambda_2 = 1.407, \mathbf{e}_2' = \begin{bmatrix} -0.368 & -0.236 & -0.315 & 0.585 & 0.606 \end{bmatrix}$$
$$\lambda_3 = 0.501, \mathbf{e}_3' = \begin{bmatrix} -0.604 & -0.136 & 0.772 & 0.093 & -0.109 \end{bmatrix},$$

$$\lambda_4 = 0.400, \mathbf{e}_4' = \begin{bmatrix} -0.363 & 0.629 & -0.289 & -0.381 & 0.493 \end{bmatrix},$$
$$\lambda_5 = 0.255, \mathbf{e}_5' = \begin{bmatrix} 0.384 & -0.496 & 0.071 & 0.595 & -0.498 \end{bmatrix}$$

Sử dụng các biến được chuẩn hóa, ta thu được hai thành phần chính như sau:

$$\hat{y}_1 = \mathbf{e}_1'\mathbf{z} = 0.496z_1 + 0.532z_2 + 0.465z_3 + 0.387z_4 + 0.361z_5$$

$$\hat{y}_2 = \mathbf{e}_2'\mathbf{z} = -0.368z_1 - 0.236z_2 - 0.315z_3 + 0.585z_4 + 0.606z_5$$

Hai thành phần chính này đóng góp tới:

$$\frac{\lambda_1 + \lambda_2}{p} \cdot 100\% = \frac{2.437 + 1.407}{5} \cdot 100\% = 77\%$$

của tổng phương sai mẫu (biến đã chuẩn hóa).

Nhận xét:

* Ta nhận thấy thành phần chính đầu tiên là tổng của các x_i với trọng số khác nhau. Thành phần này có thể được gọi là thành phần thị trường chứng khoán chung hoặc đơn giản hơn là thành phần thị trường chung.

* Thành phần chính thứ hai thể hiện sự tương phản giữa nhóm cổ phiếu ngân hàng (JP Morgan, Citibank, Wells Fargo) và nhóm cổ phiếu dầu (Royal Dutch Shell và Exxon-Mobil). Thành phần này có thể được gọi là một thành phần của ngành.

Như vậy, ta thấy rằng hầu hết sự thay đổi trong lợi nhuận của những cổ phiếu này đều bị ảnh hưởng bởi các hoạt động thị trường và hoạt động ngành không tương quan. Từ các thành phần còn lại ngoài hai thành phần đầu tiên, ta không dễ để suy luận ra sự biến đổi cụ thể với mỗi cổ phiếu. Do đó, trong hầu hết các trường hợp, chúng không giải thích nhiều cho tổng phương sai mẫu.

IV Vẽ đồ thị các thành phần chính

1 Bối cảnh

Trong lĩnh vực phân tích dữ liệu, Edward Tufte được biết đến như một chuyên gia hàng đầu về trực quan hóa dữ liệu, người đã đưa ra nhiều nguyên tắc quan trọng về cách thể hiện thông tin một cách hiệu quả. Câu nói nổi tiếng của ông "...there are right ways and wrong ways to show data; there are displays that reveal the truth and displays that do not" đã thể hiện một chân lý quan trọng về việc trình bày dữ liệu: không phải mọi cách thể hiện dữ liệu đều có giá trị như nhau.

Trong phân tích thành phần chính (Principal Component Analysis - PCA), nguyên lý này càng trở nên quan trọng hơn bao giờ hết. Khi chúng ta chuyển đổi dữ liệu từ không gian ban đầu sang không gian mới của các thành phần chính, việc lựa chọn cách biểu diễn phù hợp trở thành yếu tố then chốt để hiểu được bản chất của dữ liệu. Một đồ thị được thiết kế tốt có thể giúp chúng ta phát hiện những mẫu dữ liệu bất thường, kiểm tra các giả định thống kê, và quan trọng hơn cả, là hiểu được cấu trúc tiềm ẩn của dữ liệu.

Trong phần này, chúng ta sẽ tập trung vào việc khám phá các phương pháp trực quan hóa khác nhau trong PCA, từ việc sử dụng đồ thị Q-Q để kiểm tra tính chuẩn của dữ liệu, đến việc dùng đồ thị phân tán để phát hiện các quan sát bất thường. Mỗi loại đồ thị đều có vai trò riêng của nó trong việc "reveal the truth" về dữ liệu, và việc hiểu được khi nào nên sử dụng loại đồ thị nào là chìa khóa để thực hiện phân tích thành công.

2 Cơ sở lý thuyết

Trong phân tích thành phần chính, việc vẽ đồ thị đóng vai trò quan trọng trong việc phát hiện các quan sát bất thường và kiểm tra giả thuyết về phân phối chuẩn của dữ liệu. Để hiểu được cơ sở của việc này, ta cần xem xét cách biểu diễn các quan sát trong không gian của các thành phần chính.

Mỗi quan sát có thể được biểu diễn dưới dạng một tổ hợp tuyến tính của tập đầy đủ các vector riêng $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, ..., \hat{\mathbf{e}}_p$ của ma trận hiệp phương sai mẫu $\mathbf{S}$:

$$\mathbf{x}_j = (\mathbf{x}_j'\hat{\mathbf{e}}_1)\hat{\mathbf{e}}_1 + (\mathbf{x}_j'\hat{\mathbf{e}}_2)\hat{\mathbf{e}}_2 + ... + (\mathbf{x}_j'\hat{\mathbf{e}}_p)\hat{\mathbf{e}}_p \tag{21}$$

hay có thể viết gọn lại thành:

$$\mathbf{x}_j = \hat{y}_{j1}\hat{\mathbf{e}}_1 + \hat{y}_{j2}\hat{\mathbf{e}}_2 + ... + \hat{y}_{jp}\hat{\mathbf{e}}_p \tag{22}$$

Trong quá trình phân tích thành phần chính, ta có thể phân tách biểu diễn này thành hai phần:

$$\mathbf{x}_j = (\hat{y}_{j1}\hat{\mathbf{e}}_1 + ... + \hat{y}_{j,q-1}\hat{\mathbf{e}}_{q-1}) + (\hat{y}_{jq}\hat{\mathbf{e}}_q + ... + \hat{y}_{jp}\hat{\mathbf{e}}_p) \tag{23}$$

Mục đích chính của phân tích thành phần chính là giảm số chiều của dữ liệu. Do đó, khi sử dụng công thức trên để khôi phục dữ liệu cho mỗi quan sát $\mathbf{x}_j$, các giá trị thu được thường không khác quá nhiều so với dữ liệu ban đầu. Tuy nhiên, nếu trong số các thành phần chính cuối mà ta bỏ đi có giá trị rất lớn hoặc rất nhỏ, điều này sẽ dẫn đến sai khác lớn so với giá trị ban đầu. Những quan sát như vậy được xem là quan sát bất thường và cần được kiểm tra kỹ lưỡng.
Với cơ sở đó, dựa vào đồ thị, các quan sát bất thường có thể được xác định dựa trên yếu tố như:

- Các quan sát có giá trị rời rạc, nằm ngoài phạm vi dự kiến hoặc quá xa so với các điểm dữ liệu chính
- Các quan sát có mối quan hệ tương phản hoặc không tuân theo mô hình phân bố chung của dữ liệu
- Các quan sát có giá trị cực đại hoặc cực tiểu trên một số thành phần chính

3 Các loại đồ thị được sử dụng

Trong phần này, chúng ta sẽ sử dụng bộ dữ liệu về kích thước mai rùa đực cho bởi bảng sau

Table 6.9 Carapace Measurements (in Millimeters) for Painted Turtles

Female			Male		
Length (x_1)	Width (x_2)	Height (x_3)	Length (x_1)	Width (x_2)	Height (x_3)
98	81	38	93	74	37
103	84	38	94	78	35
103	86	42	96	80	35
105	86	42	101	84	39
109	88	44	102	85	38
123	92	50	103	81	37
123	95	46	104	83	39
133	99	51	106	83	39
133	102	51	107	82	38
133	102	51	112	89	40
134	100	48	113	88	40
136	102	49	114	86	40
138	98	51	116	90	43
138	99	51	117	90	41
141	105	53	117	91	41
147	108	57	119	93	41
149	107	55	120	89	40
153	107	56	120	93	44
155	115	63	121	95	42
155	117	60	125	93	45
158	115	62	127	96	45
159	118	63	128	95	45
162	124	61	131	95	46
177	132	67	135	106	47

Quy tâm bảng dữ liệu

Length	Width	Height
-20.375	-14.29167	-3.70833
-19.375	-10.29167	-5.70833
-17.375	-8.29167	-5.70833
-12.375	-4.29167	-1.70833
-11.375	-3.29167	-2.70833
-10.375	-7.29167	-3.70833
-9.375	-5.29167	-1.70833
-7.375	-5.29167	-1.70833
-6.375	-6.29167	-2.70833
⋮	⋮	⋮
14.625	6.70833	4.29167
17.625	6.70833	5.29167
21.625	17.70833	6.29167

Tính được các giá trị riêng:

$$\hat{\lambda_1} = 195,275$$
$$\hat{\lambda_2} = 3,689$$
$$\hat{\lambda_3} = 1,104$$

Các vector riêng tương ứng:

$$\hat{e}_1' = [-0,840; -0,488; -0,237]$$
$$\hat{e}_2' = [-0,492; 0,869; -0,047]$$
$$\hat{e}_3' = [-0,229; -0,077; 0,971]$$

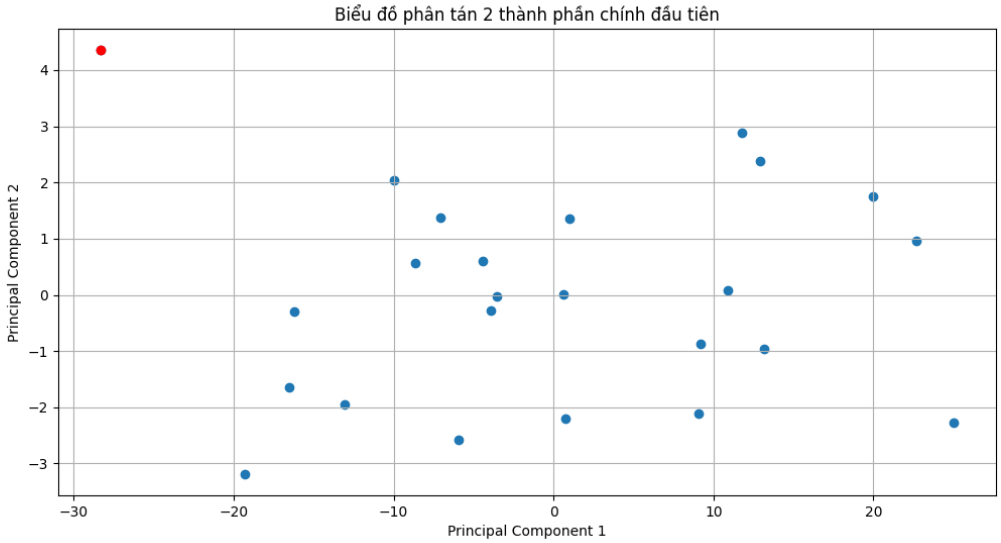
Ta thu được các thành phần chính:

$$\hat{y}_1 = \hat{\mathbf{e}}_1'\mathbf{z} = -0,840z_1 - 0,488z_2 - 0,237z_3$$
$$\hat{y}_2 = \hat{\mathbf{e}}_2'\mathbf{z} = -0,492z_1 + 0,869z_2 - 0,047z_3$$
$$\hat{y}_3 = \hat{\mathbf{e}}_3'\mathbf{z} = -0,229z_1 - 0,077z_2 + 0,971z_3$$

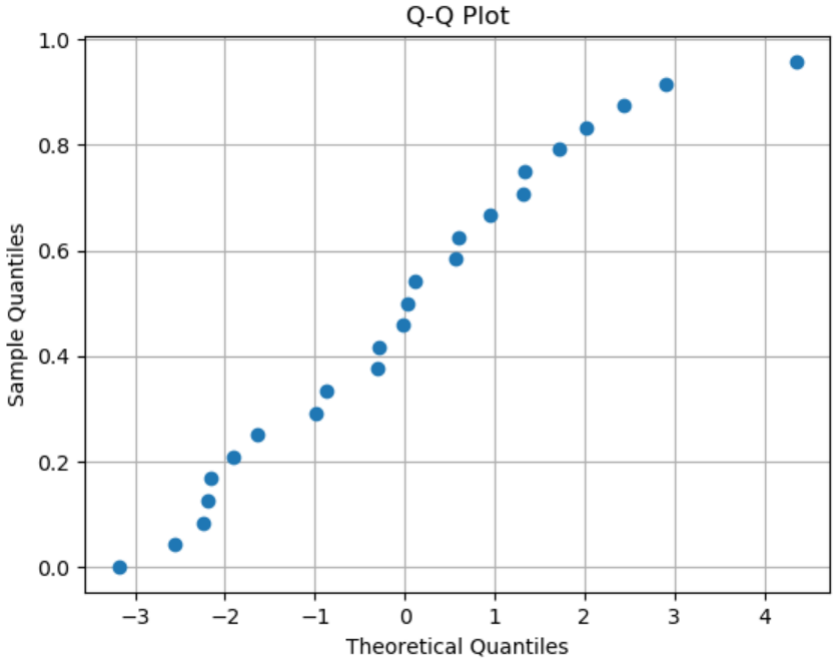
Bảng dữ liệu các thành phần chính:

$\boldsymbol{y}_1$	$\boldsymbol{y}_2$	$\boldsymbol{y}_3$
24.96821	-2.22067	2.16555
22.65021	0.85733	-0.31345
19.99421	1.61133	-0.92545
12.89421	2.43933	1.50555
11.80321	2.86333	0.22855
13.15221	-1.05767	-0.66345
⋮	⋮	⋮
-16.57579	-1.56767	0.30155
-19.33279	-3.09067	0.58555
-28.29779	4.45333	-0.20645

Đồ thị phân tán của hai thành phần chính đầu tiên sẽ như sau:



Dựa vào đồ thị ta có thể thấy có một điểm nằm cách xa so với phân bố dữ liệu.
Tiếp theo, ta sẽ vẽ đồ thị Q-Q để kiểm tra xem các dữ liệu của thành phần chính có tuân theo phân phối chuẩn hay không.



Nhận xét

Từ hai đồ thị trên, ta nhận thấy có một điểm dữ liệu bất thường. Cần kiểm tra lại xem đây là lỗi khi ghi dữ liệu hay con rùa có bị dị thường về cơ thể hay không. Loại trừ điểm này, các điểm dữ liệu khác đều tuân theo phân phối chuẩn. Đồ thị cho thấy các thành phần chính khác không có sự bất thường đáng kể nào.

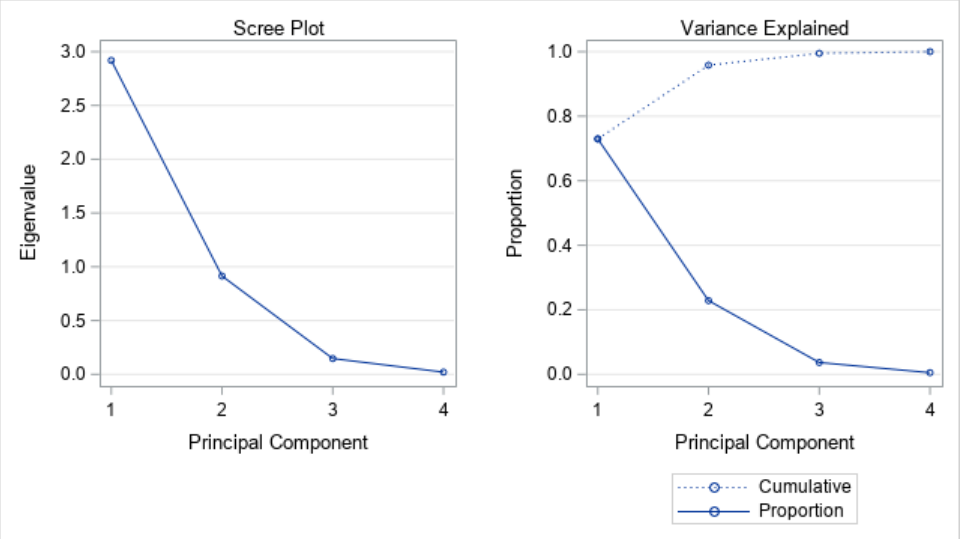
4 Một số đồ thị khác

Trong phần này chúng ta sẽ sử dụng bộ dữ liệu hoa Iris để lấy ví dụ. Thông tin bộ dữ liệu như sau:

- **Tổng số mẫu:** 150 mẫu hoa.
- **Số lượng lớp (class):** 3, tương ứng với ba loại hoa Iris:
  - *Iris setosa*
  - *Iris versicolor*
  - *Iris virginica*
- **Số lượng đặc trưng (features):** 4, gồm:
  - Chiều dài đài hoa (*sepal length*, cm)
  - Chiều rộng đài hoa (*sepal width*, cm)
  - Chiều dài cánh hoa (*petal length*, cm)
  - Chiều rộng cánh hoa (*petal width*, cm)

4.1 Đồ thị scree

Ý tưởng chính của phân tích thành phần chính (PCA) là phần lớn phương sai trong dữ liệu có chiều cao có thể được nắm bắt trong một không gian con có chiều thấp hơn, được tạo ra bởi vài thành phần chính đầu tiên. Do đó, ta có thể "giảm chiều" bằng cách chọn một số ít thành phần chính để giữ lại. Tuy nhiên, ta nên giữ lại bao nhiêu thành phần chính (PCs)? Đồ thị scree là một đồ thị đường của các giá trị riêng của ma trận tương quan, sắp xếp từ lớn đến nhỏ.



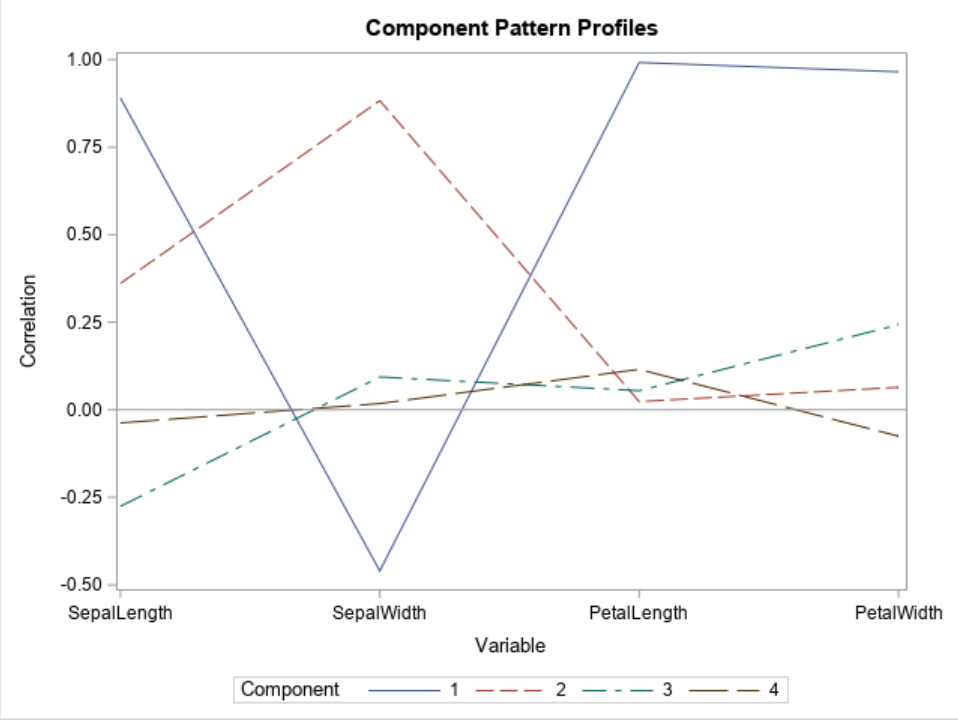
Ta có thể sử dụng đồ thị scree như một công cụ đồ họa để giúp ta chọn số lượng thành phần chính cần giữ lại. Trong đồ thị scree cho dữ liệu iris, ta có thể thấy (trong đồ thị "Variance Explained") rằng hai giá trị riêng đầu tiên giải thích khoảng 96% phương sai trong dữ liệu bốn chiều. Điều này gợi ý rằng ta nên giữ lại hai thành phần chính đầu tiên, và rằng một phép chiếu dữ liệu lên hai thành phần chính đầu tiên sẽ cung cấp cho ta một cách tốt để trực quan hóa dữ liệu trong không gian con tuyến tính có chiều thấp.

4.2 Đồ thị profile

Đồ thị profile cho thấy sự tương quan giữa mỗi thành phần chính (PC) và các biến gốc. Đến một mức độ nào đó, ta có thể ước lượng dấu hiệu và độ lớn xấp xỉ của các tương quan bằng cách nhìn vào các hệ số định nghĩa mỗi thành phần chính như một kết hợp tuyến tính của các biến gốc. Các tương quan được thể hiện trong đồ thị "Component Pattern Profiles" dưới đây.

Đồ thị profile tiết lộ các thông tin sau về các thành phần chính:

- Thành phần chính đầu tiên (đường màu xanh dương đậm) có sự tương quan mạnh mẽ dương với SepalLength, PetalLength, và PetalWidth. Nó có sự tương quan tiêu cực vừa phải với SepalWidth.
- Thành phần chính thứ hai (đường gạch đỏ) có sự tương quan dương với SepalLength và SepalWidth.
- Thành phần chính thứ ba và thứ tư chỉ có sự tương quan nhỏ với các biến gốc.



V Suy luận dựa trên mẫu lớn

Giá trị riêng, vector riêng của ma trận hiệp phương sai (hay ma trận tương quan) là yếu tố thiết yếu trong phân tích các thành phần chính. Vector riêng quyết định hướng có phương sai lớn nhất, giá trị riêng sẽ xác định phương sai.

Việc giữ lại các thành phần chính dựa trên các giá trị riêng của ma trận hiệp phương sai là phụ thuộc vào độ lớn của các giá trị riêng đó. Tuy nhiên các λ_i là chưa biết. Hơn nữa chất lượng của thành phần chính thứ i phụ thuộc vào λ_i và vectơ riêng e_i. Vì vậy việc tìm khoảng tin cậy của các giá trị riêng λ_i, các vectơ riêng e_i và qua đó là thành phần chính Y_i dựa trên một mẫu ngẫu nhiên $X_1, X_2, \ldots, X_n$ cần được nghiên cứu.

1 Thuộc tính mẫu lớn $\hat{\lambda}_i$ và $\hat{e}_i$

Giả sử các quan sát $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên từ một tổng thể có phân phối chuẩn. Đồng thời, cũng giả sử rằng các giá trị riêng (chưa biết) của Σ là phân biệt và đều dương, tức là:

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$$

Trong đó, ma trận xác định dương là ma trận mà mọi định thức con chính (Định thức con chính cấp k của ma trận vuông cấp n là định thức của ma trận khối tạo thành từ dòng và cột đầu tiên của A) đều dương.

Giả sử $(\hat{\lambda}_i, \hat{e}_i)$ là cặp giá trị riêng và vector riêng của ma trận hiệp phương sai mẫu S, sao cho:

$$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$$

Anderson và Girshick đã thiết lập các lý thuyết phân phối mẫu lớn cho các giá trị riêng $\hat{\lambda}' = [\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p]$ và vector riêng $\hat{e_1}, \hat{e_2}, \ldots, \hat{e_p}$ của tập S.

Kết luận 1

Cho Λ là ma trận đường chéo các giá trị riêng $\lambda_1, \lambda_2, \ldots, \lambda_p$ của Σ, khi đó:

$$\sqrt{n}(\hat{\lambda} - \lambda) \sim N_p(0, 2\Lambda^2)$$

Kết luận 1 cho rằng, với n lớn, $\hat{\lambda}_i$ được phân phối độc lập. Và $\hat{\lambda}_i$ sẽ có phân phối xấp xỉ phân phối chuẩn $N(\lambda_i, \frac{2\lambda_i^2}{n})$.

Với phân phối chuẩn này, tồn tại:

$$P\left|\hat{\lambda}_i - \lambda_i\right| \leq z_\alpha \frac{2\lambda_i^2}{n} = 1 - \alpha$$

trong đó, $z\left(\frac{\alpha}{2}\right)$ là phân vị trên thứ $100\frac{\alpha}{2}$ của phân phối chuẩn tắc.

Khoảng tin cậy với mức tin cậy $100(1-\alpha)\%$ của λ_i sẽ được xác định:

$$\frac{\hat{\lambda}_i}{1+z_{\frac{\alpha}{2}}\frac{2}{n}} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{1-z_{\frac{\alpha}{2}}\frac{2}{n}} \tag{28}$$

Khoảng tin cậy đồng thời với mức tin cậy $100(1-\alpha)\%$ cho m giá trị riêng λ_i thu được theo phương pháp Bonferroni bằng cách thay $z\left(\frac{\alpha}{2}\right)$ bằng $z\left(\frac{\alpha}{2m}\right)$.

Kết luận 2

Cho:

$$E_i = \lambda_i \sum_{k=1,k\neq i}^{p} \frac{\lambda_k}{(\lambda_k-\lambda_i)^2} e_k e_k'$$

với e_k là vector riêng ứng với giá trị riêng λ_k của Σ. Khi đó:

$$\sqrt{n}(\hat{e}_i - e_i) \sim N_p(0, E_i)$$

Kết luận 2 cho rằng, $\hat{e}_i$ được phân phối chuẩn tương ứng với các e_i của mẫu lớn. Các phần tử của mỗi $\hat{e}_i$ có tương quan và mối tương quan phụ thuộc rất nhiều vào sự phân tách của các giá trị riêng $\lambda_1, \lambda_2, \ldots, \lambda_p$. Sai số chuẩn gần đúng cho các hệ số $\hat{e}_{ik}$ được xác định bởi căn bậc hai các phần tử đường chéo của $\frac{1}{n}\hat{E}_i$.

Trong đó $\hat{E}_i$ được tính từ E_i bằng cách thay $\lambda_i = \hat{\lambda}_i$ và $e_i = \hat{e}_i$.

Kết luận 3

Mỗi $\hat{\lambda}_i$ được phân phối độc lập với các phần tử $\hat{e}_i$.

Ví dụ 5.1: Xây dựng khoảng tin cậy cho λ_1

Bảng 2: Tỷ lệ hoàn vốn của 5 cổ phiếu: JPMorgan, Citibank, Wells Fargo, Royal Dutch Shell và ExxonMobil từ tháng 1 năm 2004 đến tháng 12 năm 2005 theo tuần

Week	JPMorgan	Citibank	Wells Fargo	Royal Dutch Shell	ExxonMobil
1	0.01303	-0.00784	-0.00319	-0.04477	0.00522
2	0.00849	0.01669	-0.00621	0.01196	0.01349
3	-0.01792	-0.00864	0.01004	0	-0.00614
4	0.02156	-0.00349	0.01744	-0.02859	-0.00695
5	0.01082	0.00372	-0.01013	0.02919	0.04098
6	0.01017	-0.01220	-0.00838	0.01371	0.00323
8	0.04848	-0.00515	0.01825	0.00633	0.00768
9	-0.03449	-0.01380	-0.00805	-0.02990	-0.01081
10	-0.00466	0.02099	-0.00608	-0.02039	-0.01267
...	...	...	...	...	...
100	0.00336	0.00290	-0.00305	-0.00122	-0.00970
101	0.01701	0.00951	0.01820	-0.01618	-0.00756
102	0.01039	-0.00266	0.00443	-0.00248	-0.01645
103	-0.01279	-0.01437	-0.01874	-0.00498	-0.01637

Đặt X_1, X_2, X_3, X_4, X_5 lần lượt biểu thị tỷ lệ hoàn vốn hàng tuần của JPMorgan, Citibank, Wells Fargo, Royal Dutch Shell và ExxonMobil.

Giả sử rằng tỷ lệ hoàn vốn là các biến độc lập từ một tổng thể với phân phối $N_5(\mu, \Sigma)$. Trong đó Σ xác định dương với các giá trị riêng phân biệt $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \lambda_5 > 0$.

Vì $n = 103$ là một mẫu lớn, ta có thể dùng công thức (28) với $i = 1$ để xây dựng khoảng tin cậy 95%. Ta có ma trận hiệp phương sai mẫu S:

$$S = \begin{bmatrix} 0.00043327 & 0.00027567 & 0.00015903 & 0.00006412 & 0.00008897 \\ 0.00027567 & 0.00043872 & 0.00017997 & 0.00018145 & 0.00012326 \\ 0.00015903 & 0.00017997 & 0.00022397 & 0.00007341 & 0.00006055 \\ 0.00006412 & 0.00018145 & 0.00007341 & 0.00072250 & 0.00050828 \\ 0.00008897 & 0.00012326 & 0.00006055 & 0.00050828 & 0.00076567 \end{bmatrix}$$

Các giá trị riêng của S được xác định là:

$$\hat{\lambda}_1 = 0.00137, \quad \hat{\lambda}_2 = 0.00070, \quad \hat{\lambda}_3 = 0.00025, \quad \hat{\lambda}_4 = 0.00014, \quad \hat{\lambda}_5 = 0.00012$$

Ta có $\hat{\lambda}_1 = 0.00137$ và $z\left(\frac{\alpha}{2}\right) = z(0.025) = 1.96$. Vậy với độ tin cậy 95%, thay vào công thức (28), ta được:

$$\frac{0.00137}{1 + \frac{1.96 \cdot 2}{103}} \leq \lambda_1 \leq \frac{0.00137}{1 - \frac{1.96 \cdot 2}{103}}$$

Tính toán cho ra:

$$0.00108 \leq \lambda_1 \leq 0.00188$$

Chú ý:

- Khi một giá trị riêng lớn, chẳng hạn như 100 hoặc thậm chí 1000, các khoảng được tạo bởi công thức (28) có thể khá rộng, đối với các mức độ tin cậy hợp lý, mặc dù n là khá lớn.
- Nói chung, khoảng tin cậy sẽ trở nên rộng hơn cùng với tốc độ $\hat{\lambda}_i'$ tăng. Do đó, cần thận trọng trong việc loại bỏ hoặc giữ lại các thành phần chính dựa trên việc kiểm tra các $\hat{\lambda}_i'$.

2 Kiểm tra cấu trúc tương quan bình đẳng

Cấu trúc tương quan đặc biệt $Cov(X_i, X_k) = \sqrt{\sigma_{ii}\sigma_{kk}} \cdot \rho$ hay $Corr(X_i, X_k) = \rho$ với mọi $i \neq k$ là một cấu trúc quan trọng, trong đó các giá trị riêng của Σ không khác biệt và các kết quả trước đó không được áp dụng.

Để kiểm tra cấu trúc này, xét cặp giả thuyết:

$$H_0 : \rho = \rho_0$$
$$H_1 : \rho \neq \rho_0$$

Nếu H_0 đúng, thì ma trận tương quan sẽ bằng với ma trận tương quan đặc biệt, tức là các giá trị nằm ngoài đường chéo chính là như nhau.

Phép thử theo Lawley

Phép thử H_0 so với H_1 có thể được thực hiện bằng thống kê tỷ lệ khả năng xảy ra. Tuy nhiên, Lawley đã chứng minh rằng một quy trình thử nghiệm tương đương có thể được xây dựng từ các phần tử nằm ngoài đường chéo của R.

Phương pháp Lawley yêu cầu:

$$r_k = \frac{1}{p-1} \sum_{i=1, i\neq k}^{p} r_{ik}, \quad r = \frac{2}{p(p-1)} \sum_{i<k} \hat{r}_{ik}$$

$$\gamma = \frac{(p-1)^2[1-(1-r)^2]}{p-(p-2)(1-r)^2}$$

trong đó:

- r_k là giá trị trung bình của các phần tử nằm ngoài đường chéo ở cột (hoặc hàng) thứ k của R,
- r là giá trị trung bình tổng thể của các phần tử nằm ngoài đường chéo,
- γ là tham số xác định độ chênh lệch.

Kiểm định mẫu lớn với mức ý nghĩa α, bác bỏ H_0 và chấp nhận H_1 nếu:

$$T = (n-1) \cdot \frac{(1-r)^2}{\sum_{i<k}(r_{ik}-r)^2 - \hat{\gamma}\sum_{k=1}^{p}(r_k-r)^2} > \chi^2_{(p+1)(p-2)/2}(\alpha)$$

Trong đó:

- $\chi^2_{(p+1)(p-2)/2}(\alpha)$ là phân vị trên thứ 100α của phân phối χ^2 với bậc tự do $\frac{(p+1)(p-2)}{2}$.

Ví dụ 5.2

Từ ví dụ 2.3, ta có ma trận tương quan mẫu được xây dựng từ cân nặng của 150 cá thể chuột cái sau khi sinh:

$$R = \begin{bmatrix} 1.0 & 0.7501 & 0.6329 & 0.6363 \\ 0.7501 & 1.0 & 0.6925 & 0.7386 \\ 0.6329 & 0.6925 & 1.0 & 0.6625 \\ 0.6363 & 0.7386 & 0.6625 & 1.0 \end{bmatrix}$$

Sử dụng ma trận tương quan trên để minh họa cho quá trình kiểm tra mẫu lớn trong công thức (30). Với $p = 4$, ta đặt:

$$H_0 : \rho = \rho_0 = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}, \quad H_1 : \rho \neq \rho_0$$

Sử dụng công thức (29) và (30), ta tính toán:

$$r_1 = \frac{1}{3}(0.7501 + 0.6329 + 0.6363) = 0.6731$$

$$r_2 = \frac{1}{3}(0.7501 + 0.6925 + 0.7386) = 0.7271$$

$$r_3 = \frac{1}{3}(0.6329 + 0.6925 + 0.6625) = 0.6626$$

$$r_4 = \frac{1}{3}(0.6363 + 0.7386 + 0.6625) = 0.6791$$

$$r = \frac{2}{4 \cdot 3}(0.7501 + 0.6329 + 0.6925 + 0.6363 + 0.7386 + 0.6625) = 0.6855$$

$$\hat{\gamma} = \frac{(4-1)^2[1-(1-0.6855)^2]}{4-(4-2)(1-0.6855)^2} = 2.1329$$

Tổng bình phương sai lệch của các phần tử ngoài đường chéo:

$$\sum_{i<k}(r_{ik}-r)^2 = (0.7501-0.6855)^2 + (0.6329-0.6855)^2 + \cdots + (0.6625-0.6855)^2 = 0.01277$$

Tổng bình phương sai lệch của các giá trị trung bình r_k:

$$\sum_{k=1}^{4}(r_k-r)^2 = (0.6731-0.6855)^2 + (0.7271-0.6855)^2 + (0.6626-0.6855)^2 + (0.6791-0.6855)^2 = 0.00245$$

Thống kê kiểm định:

$$T = (150 - 1) \cdot \frac{(1 - 0.6855)^2}{0.01277 - 2.1329 \cdot 0.00245} = 11.4$$

Với $(p+1)(p-2)/2 = \frac{5 \cdot 2}{2} = 5$ và mức ý nghĩa $\alpha = 0.05$, giá trị tới hạn:

$$\chi^2_5(0.05) = 11.07$$

Do $T = 11.4 > 11.07$, ta bác bỏ giả thuyết H_0 và chấp nhận H_1. Tức là các giá trị riêng của ma trận Σ khác nhau.

Nhận xét

Trong ví dụ trên, các giá trị riêng nhỏ nhất $\hat{\lambda}_2, \hat{\lambda}_3$ và $\hat{\lambda}_4$, trong đó $\hat{\lambda}_4$ nhỏ hơn một chút so với hai giá trị còn lại. Với cỡ mẫu lớn, những khác biệt nhỏ trong cấu trúc tương quan bình đẳng có ý nghĩa thống kê.

VI Ứng dụng của PCA

1 Thuật toán PCA

Bước 1: Tính vector kỳ vọng của toàn bộ tập dữ liệu

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

trong đó N là số lượng điểm dữ liệu, x_n là vector dữ liệu thứ n.

Bước 2: Quy tâm và chuẩn hóa dữ liệu

Để chuẩn hóa dữ liệu, mỗi giá trị x_{ij} trong ma trận dữ liệu X được chuẩn hóa bằng công thức:

$$\hat{X}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

Trong đó:

- x_{ij} là giá trị của đặc trưng j tại mẫu i.
- $\bar{x}_j$ là trung bình của đặc trưng j.
- σ_j là độ lệch chuẩn của đặc trưng j.
- $\hat{X}_{ij}$ là giá trị đã chuẩn hóa của đặc trưng j tại mẫu i

Bước 3: Tính ma trận hiệp phương sai của dữ liệu đã chuẩn hóa

Ma trận hiệp phương sai được tính theo công thức:

$$\Sigma = \frac{1}{n-1}\hat{X}'\hat{X}$$

với $\hat{X}$ là ma trận dữ liệu sau khi chuẩn hóa.

Bước 4: Tính các giá trị riêng và vector riêng của ma trận hiệp phương sai, sau đó sắp xếp các giá trị riêng và các vector riêng tương ứng theo thứ tự trị riêng không tăng.

Bước 5: Chọn K vector riêng ứng với K giá trị riêng lớn nhất

Tập hợp K vector riêng này tạo thành ma trận U_K, trong đó các cột của U_K là các vector riêng. Các vector riêng này được gọi là *các thành phần chính* và tạo thành một không gian con U_K.

Bước 6: Chiếu dữ liệu chuẩn hóa xuống không gian con

Chiếu dữ liệu $\hat{X}$ lên không gian do U_K sinh ra:

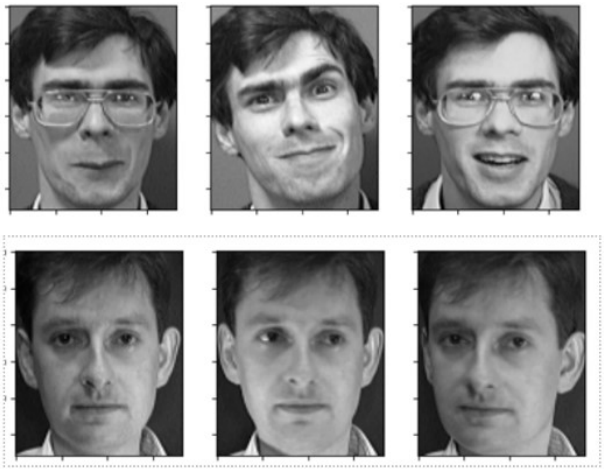
$$Z = \hat{X}U_K$$

Trong đó, Z là ma trận ảnh trong không gian mới U_K.

2 PCA với bộ dữ liệu khuôn mặt

3 Giới thiệu bài toán đặt ra

- **Bộ dữ liệu**:
  - Bao gồm hình ảnh của 40 người, mỗi người có 10 bức ảnh.
  - Định dạng: `.pgm`, kích thước mỗi ảnh: 92×112 pixel.
  - Mỗi pixel có giá trị từ $0 - 255$, biểu thị mức độ sáng.



Hình 3.1: Bộ dữ liệu khuôn mặt

- **Mục tiêu bài toán**: Sử dụng **khoảng cách Euclidean** để tìm kiếm hình ảnh tương đồng trong 3 trường hợp:
  1. Mẫu test nằm trong bộ dữ liệu train dùng cho phân tích PCA.
  2. Mẫu test hoàn toàn nằm ngoài tập train.
  3. Mẫu test thuộc người có ảnh trong tập train nhưng chính ảnh đó không tham gia phân tích.
- **Đặt vấn đề**:
  - Dữ liệu ảnh gốc có $92 \times 112 = 10304$ biến.
  - Khối lượng tính toán lớn $\Rightarrow$ cần giảm chiều dữ liệu bằng PCA.

4 Thực hiện PCA giảm chiều dữ liệu

1. **Mảng hóa dữ liệu hình ảnh**:

	0	1	2	...	10301	10302	10303
0	48	49	45	...	47	46	46
1	34	34	33	...	37	40	33
2	60	60	62	...	32	34	34
3	39	44	53	...	29	26	29
4	63	53	35	...	41	10	24
..	...	...	...	...	...	...	...
384	87	93	84	...	125	117	125
385	97	94	92	...	144	143	129
386	93	93	88	...	112	111	111
387	88	85	89	...	117	117	116
388	86	90	87	...	127	124	126

[389 rows x 10304 columns]

Hình 4.1: Ma trận dữ liệu ảnh

- Chuyển đổi từng ảnh thành mảng 1 chiều với 10,304 phần tử.
- Kết quả: Ma trận X, mỗi hàng là một ảnh được mảng hóa.

2. **Chuẩn hóa dữ liệu hình ảnh**:

- **Tính vector kỳ vọng**:

$$x_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij}$$

- **Tính vector độ lệch chuẩn**:

$$\sigma_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(x_{ij} - x_i)^2}$$

- **Chuẩn hóa dữ liệu**:

$$\hat{X}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

Trong đó:

  - x_{ij} là giá trị của đặc trưng j tại mẫu i.
  - $\bar{x}_j$ là trung bình của đặc trưng j.
  - σ_j là độ lệch chuẩn của đặc trưng j.
  - $\hat{X}_{ij}$ là giá trị đã chuẩn hóa của đặc trưng j tại mẫu i

3. **Tính ma trận hiệp phương sai**:

$$\Sigma = \frac{1}{n-1}\hat{X}'\hat{X}$$

	0	1	2	...	10301	10302	10303
0	-1.024861	-0.998482	-1.124400	...	-0.638089	-0.671914	-0.664412
1	-1.417088	-1.419233	-1.462655	...	-0.849499	-0.805952	-0.957753
2	-0.688666	-0.689931	-0.645206	...	-0.955204	-0.939990	-0.935188
3	-1.277007	-1.138732	-0.898897	...	-1.018627	-1.118708	-1.048012
4	-0.604618	-0.886281	-1.406280	...	-0.764935	-1.476143	-1.160836
..	...	...	...	...	...	...	...
384	0.067772	0.235722	-0.025072	...	1.010910	0.914205	1.118200
385	0.347934	0.263772	0.200431	...	1.412589	1.495037	1.208459
386	0.235869	0.235722	0.087680	...	0.736077	0.780167	0.802294
387	0.095788	0.011321	0.115868	...	0.841782	0.914205	0.915118
388	0.039756	0.151571	0.059492	...	1.053192	1.070583	1.140765

[389 rows x 10304 columns]

Hình 4.2: Ma trận dữ liệu ảnh sau khi chuẩn hóa

	0	1	2	...	10301	10302	10303
0	1.000000	0.993528	0.992293	...	-0.085137	-0.065080	-0.031155
1	0.993528	1.000000	0.993357	...	-0.084246	-0.064680	-0.031532
2	0.992293	0.993357	1.000000	...	-0.082862	-0.061237	-0.030283
3	0.992021	0.994036	0.993257	...	-0.080963	-0.062751	-0.027884
4	0.989298	0.990680	0.991714	...	-0.080613	-0.059661	-0.026966
...	...	...	...	...	...	...	...
10299	-0.024184	-0.025980	-0.029307	...	0.831624	0.793601	0.805532
10300	-0.076421	-0.076910	-0.079058	...	0.905192	0.855483	0.849512
10301	-0.085137	-0.084246	-0.082862	...	1.000000	0.925694	0.863093
10302	-0.065080	-0.064680	-0.061237	...	0.925694	1.000000	0.940179
10303	-0.031155	-0.031532	-0.030283	...	0.863093	0.940179	1.000000

Hình 4.3: Ma trận hiệp phương sai Σ

4. **Tìm giá trị riêng và vector riêng của** Σ:

   - Sử dụng hàm `numpy.linalg.eig()` để tính giá trị riêng và vector riêng của Σ.

```
0  1646.455582
1  1313.760849
2   827.436616
3   596.967018
4   522.462928
```

Hình 4.4: 5 trị riêng lớn nhất của Σ

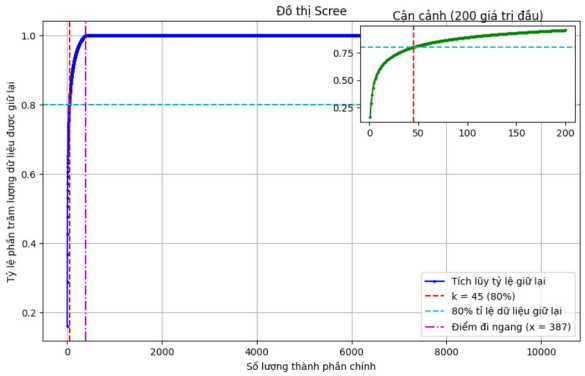
	0	1	2	3	4
0	-0.008317	0.014316	0.019648	0.014096	0.000416
1	-0.008302	0.014293	0.019872	0.014209	0.000397
2	-0.008325	0.014370	0.019728	0.014318	0.000551
3	-0.008277	0.014479	0.019831	0.014101	0.000278
4	-0.008392	0.014340	0.019780	0.014442	0.000481
...	...	...	...	...	...
10299	0.008074	0.002054	0.009849	-0.013245	-0.002178
10300	0.009115	-0.000246	0.010695	-0.012271	-0.002452
10301	0.009905	-0.000932	0.011033	-0.011388	-0.002165
10302	0.009565	-0.000964	0.011928	-0.011274	-0.001429
10303	0.009765	0.000311	0.012442	-0.011325	-0.001664

Hình 4.5: 5 vector riêng ứng với trị riêng lớn nhất của Σ

5. **Lựa chọn số lượng thành phần chính**:

- Tính tỷ lệ phương sai tích lũy:

$$P_k = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{n} \lambda_j}$$



Hình 4.6: Đồ thị Scree

Nhận xét

(a) Tại vị trí $x = 387$, ta có số lượng % dữ liệu được giữ lại là 99.99...996%

(b) Đồ thị gần như đi ngang ở giá trị 100% từ thành phần chính thứ 387

(c) Với 45/10304 thành phần chính, lượng dữ liệu được giữ lại là 80%. Khi đó, chọn $k = 45$ là số thành phần chính được giữ lại để tạo thành U_K.

6. **Chiếu dữ liệu lên không gian thành phần chính**:

$$Z = \hat{X}U_K$$

	0	1	2	...	42	43	44
0	-47.045052	0.238878	-55.935134	...	10.132375	-3.452007	5.566764
1	-61.402609	-26.002331	-24.782342	...	5.504146	3.341045	-0.947852
2	-74.845695	-29.761621	5.518660	...	10.789179	-2.303527	-2.577113
3	-57.610167	-31.629274	-34.828884	...	-1.153679	-3.526984	-0.419212
4	-78.681595	-27.926053	8.526993	...	-7.282582	4.212271	-4.390958
..	...	...	...	...	...	...	...
384	82.346875	11.252185	19.432508	...	-6.845263	3.799949	-5.225650
385	97.470538	2.012309	19.910143	...	-4.111413	-5.960370	-4.888065
386	68.971811	23.172991	27.788684	...	-4.707913	-1.479011	-6.027686
387	64.294417	21.036009	28.226400	...	-5.022119	-5.609199	-3.238753
388	78.940497	16.417965	32.633225	...	-0.979481	4.980681	-5.941218

[389 rows x 45 columns]

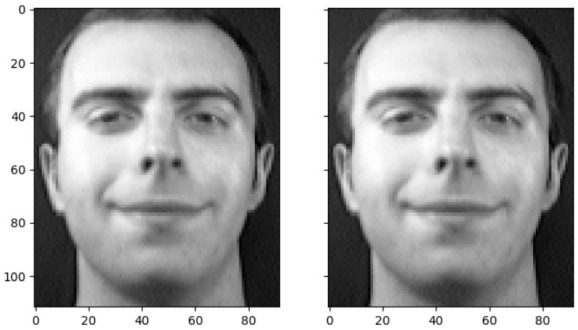
Hình 4.7: Các quan sát thu được sau khi chiếu

7. **Tìm kiếm hình ảnh tương đồng**:

- Thực hiện huẩn hóa và chiếu mẫu test lên hệ 45 thành phần chính.
- Áp dụng công thức khoảng cách Euclidean:

$$d = \sqrt{\sum_{i=1}^{k} (z_i - z_{i,\text{test}})^2}$$

- So sánh khoảng cách và tìm hình ảnh gần nhất.
- Đối với mẫu test trong tập train:



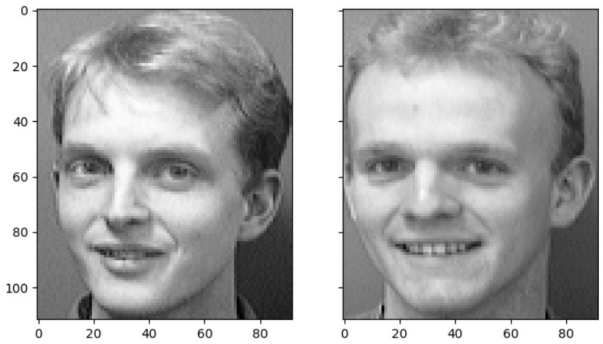
Hình 4.8: Kết quả thu được

=> Kết quả thu được trùng khớp hoàn toàn với mẫu test

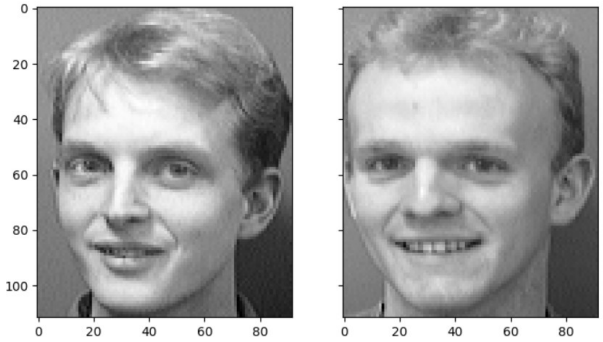
- Đối với mẫu test hoàn toàn ngoài tập train:

=> Kết quả thu được là hình ảnh của quan sát có nét tương đồng với người trong mẫu test

- Đối với mẫu test nằm ngoài tập train, là hình ảnh của bức ảnh được đem vào phân tích thành phần chính:



Hình 4.9: Kết quả thu được



Hình 4.10: Kết quả thu được

=> Kết quả thu được là cùng 1 người, sắc thái của 2 bức ảnh có nét giống nhau

Tài liệu tham khảo

1. Johnson, R. A., Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*. Englewood Cliffs, N.J: Prentice Hall.

2. Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). New York: Springer-Verlag New York, Inc.

3. UCI Machine Learning. (2017). *Red Wine Quality Dataset*. Kaggle.

4. Bộ dữ liệu gương mặt ứng dụng: *AT&T Database of Faces*. Truy cập tại: `https://www.kaggle.com`.

5. Mã nguồn và bộ dữ liệu đi kèm: `https://drive.google.com/drive/folders/1c_r04-o4Fe1EUQiRAo0WpEvObWb11dLZ?usp=share_link`.