ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
KHOA TOÁN - TIN
Faculty of Mathematics and Informatics

Factor Analysis

Báo cáo

# PHÂN TÍCH SỐ LIỆU

## Chủ đề PHÂN TÍCH NHÂN TỐ

Giảng viên hướng dẫn: ThS. Lê Xuân Lý
Nhóm sinh viên thực hiện: Nhóm 6

| | |
|---|---|
| Mai Thành Duy | 20227225 |
| Nguyễn Minh Anh | 20227212 |
| Nguyễn Thị Yến Dung | 20227222 |
| Lê Trung Kiên | 20195893 |
| Vũ Thị Minh Nguyệt | 20227250 |
| Dương Ánh Thơ | 20227265 |
| Nguyễn Hoàng Anh Kiệt | 20227237 |
| Nguyễn Phương Nhi | 20227251 |
| Lê Đức Đại | 20227217 |
| Trần Quốc Anh | 20216910 |
| Mã lớp học: | 155349 |

Hà Nội
Tháng 1 năm 2025

# Mục lục

# Bảng đánh giá

| Bảng đánh giá thành viên | | | | |
|---|---|---|---|---|
| STT | Tên | MSSV | Nhiệm vụ | Điểm cộng |
| 1 | Mai Thành Duy | 20227225 | • Nhóm trưởng.<br>• Chương 4: 4.3, 4.4.<br>• Chương 6: Đặt vấn đề, Bước 1: a), b), Bước 2: b), Bước 3,Mã nguồn. | 2 |
| 2 | Nguyễn Minh Anh | 20227212 | • Chương 1.<br>• Chương 2: 2.1, 2.2. | 2 |
| 3 | Nguyễn Thị Yến Dung | 20227222 | Chương 2: Ví dụ 2.1, 2.2, Nhận xét. | 1 |
| 4 | Lê Trung Kiên | 20195893 | Chương 6: Bước 1 c), d). | 1 |
| 5 | Vũ Thị Minh Nguyệt | 20227250 | • Chương 3: 3.1.<br>• Chương 6: Bước 2 a). | 2 |
| 6 | Dương Ánh Thơ | 20227265 | Chương 3: 3.2 | 1 |
| 7 | Nguyễn Hoàng Anh Kiệt | 20227237 | Chương 3: 3.3 | 2 |
| 8 | Nguyễn Phương Nhi | 20227251 | Chương 4: 4.1 , 4.2, 4.5. | 2 |
| 9 | Lê Đức Đại | 20227217 | Chương 5: 5.1,5.2 | 1 |
| 10 | Trần Quốc Anh | 20216910 | Chương 5: 5.3 | 1 |

# Lời mở đầu

Phân tích nhân tố là một công cụ quan trọng dành cho các nhà nghiên cứu và nhà phân tích, giúp khám phá và hiểu rõ các cấu trúc cũng như mối quan hệ tiềm ẩn trong dữ liệu. Công cụ này cho phép giảm bớt sự phức tạp của một tập hợp lớn các biến quan sát bằng cách xác định một số lượng nhỏ hơn các nhân tố cơ bản, từ đó nắm bắt được những thông tin cốt lõi có trong dữ liệu.

Phân tích nhân tố có thể được thực hiện bằng các phần mềm thống kê như SPSS, R, SAS hoặc Python. Các phần mềm này cung cấp nhiều phương pháp trích xuất và xoay nhân tố, giúp đáp ứng các loại dữ liệu cũng như mục tiêu nghiên cứu khác nhau.

Báo cáo của nhóm chúng em bao gồm các nội dung chính sau:

- Tổng quan về phân tích nhân tố
- Mô hình nhân tố trực giao
- Các phương pháp ước lượng
- Xoay nhân tố
- Điểm nhân tố
- Ví dụ thực tiễn

Chúng em xin gửi lời cảm ơn sâu sắc đến thầy Lê Xuân Lý vì sự tận tình giảng dạy và hướng dẫn trong học phần này. Trong quá trình thực hiện báo cáo, dù đã nỗ lực tìm hiểu và hoàn thiện, nhóm không tránh khỏi thiếu sót. Chúng em rất mong nhận được ý kiến đóng góp từ thầy để hoàn thiện báo cáo tốt hơn.

Chúng em xin chân thành cảm ơn!

**NHÓM 6**

# Chương 1

# Giới thiệu về phân tích nhân tố

## 1.1 Giới thiệu

Ý tưởng đầu tiên về phân tích nhân tố đã được nêu ra bởi nhà toán học người Anh Karl Pearson (1857-1936) và nhà tâm lý học Charles Spearman (1863-1945) vào đầu thế kỷ XX, để định nghĩa và đo lường trí thông minh. Với sự phát triển không ngừng của máy tính hiện đại, việc tính toán ngày càng dễ dàng hơn và những tiến bộ trong các khía cạnh nghiên cứu lý thuyết, tính toán của phân tích nhân tố cũng được quan tâm hơn.

Mục đích cơ bản của phân tích nhân tố là mô tả (nếu có thể) những mối quan hệ tương quan giữa nhiều biến thông qua ít biến. Các biến này không quan sát được và được gọi là các nhân tố (factors). Phân tích nhân tố mô tả mối quan hệ hiệp phương sai giữa nhiều biến số bằng số lượng ít hơn các nhân tố không quan sát được.

Cách tiếp cận này được thúc đẩy bởi ý tưởng rằng các biến có thể được nhóm lại dựa trên mối tương quan của chúng. Các biến trong một nhóm thể hiện mối tương quan cao giữa chúng nhưng có tương quan tương đối nhỏ với các biến trong các nhóm khác. Phân tích nhân tố nhằm mục đích xác định các cấu trúc cơ bản hoặc các yếu tố chịu trách nhiệm về mối tương quan quan sát được giữa các biến.

Phân tích nhân tố có thể được coi là một phần mở rộng của phân tích thành phần chính vì cả hai phương pháp đều nhằm mục đích ước tính ma trận hiệp phương sai của dữ liệu. Tuy nhiên, phân tích nhân tố cung cấp một phép tính gần đúng phức tạp hơn, vì nó cố gắng xác định các yếu tố cơ bản giải thích các mối tương quan quan sát được, trong khi phân tích thành phần chính tìm cách tìm các thành phần trực giao giải thích phương sai tối đa trong dữ liệu.

Câu hỏi chính trong phân tích nhân tố là liệu dữ liệu có phù hợp với cấu trúc quy định hay không. Nói cách khác, phương pháp này nhằm mục đích xác định xem liệu các mối tương quan quan sát được giữa các biến có thể được giải thích bằng một số yếu tố tiềm ẩn hạn chế hay không. Mỗi ứng dụng của phân tích nhân tố phải được đánh giá dựa trên giá trị riêng của nó để xác định tính hiệu quả và giá trị của nó trong việc khám phá các cấu trúc bên dưới có ý nghĩa.

**Ví dụ 1.1**

Spearman đã khảo sát điểm kiểm tra của học sinh về tiếng Pháp, tiếng Anh, Toán và Âm nhạc đặc trưng cho yếu tố "thông minh"cơ bản của một học sinh, còn nhóm các điểm số khác tương ứng với các nhân tố khác.

**Ví dụ 1.2**

Giả thuyết rằng tồn tại 2 loại trí thông minh là "thông minh văn học" (TMVH) và "thông minh toán học" (TMTH) không được quan sát được và được thể hiện gián tiếp qua điểm kiểm tra của 10 môn học. Khi 1 học sinh được chọn ngẫu nhiên sẽ có 10 điểm số là các biến ngẫu nhiên. Điểm số của mỗi môn học có thể được biểu diễn bằng một tổ hợp tuyến tính của 2 loại trí thông minh trên.

*Ví dụ: Điểm môn "thiên văn học" = 10 * TMVH + 6 * TMTH*

Ở đây ***10*** *và* ***6*** là hệ số tải ứng với môn "thiên văn học", môn học khác nhau thì hệ số tải khác nhau. Ngoài ra, 2 sinh viên có trí thông minh tương đương nhưng điểm vẫn có thể khác nhau do một vài yếu tố chủ quan hoặc khách quan khác.

**Ví dụ 1.3**

Giả thuyết rằng có 3 yếu tố tiềm ẩn ảnh hưởng đến sức khỏe gồm: "lối sống", "di truyền", và "môi trường". Các yếu tố này không thể quan sát trực tiếp mà được biểu hiện qua 15 chỉ số sức khỏe như huyết áp, cholesterol, đường huyết, BMI, nhịp tim,... Mỗi chỉ số có thể được mô tả là tổ hợp tuyến tính của 3 yếu tố này.

*Ví dụ: Chỉ số huyết áp = 4 * "lối sống" + 7 * "di truyền" + 3 x "môi trường".*

Ngoài ra, dù hai người có lối sống, di truyền và môi trường tương đương nhau, các chỉ số sức khỏe của họ vẫn có thể khác biệt do những yếu tố không kiểm soát được, ví dụ như căng thẳng, chế độ ăn uống hiện tại, hoặc thậm chí điều kiện thời tiết. Sai số ở đây chính là sự chênh lệch giữa chỉ số sức khỏe thực tế và giá trị dự đoán từ các yếu tố tiềm ẩn.

## 1.2 Tổng quan

- Phân tích nhân tố có thể được coi là một phương pháp mở rộng của Phân tích thành phần chính, tuy nhiên sự xấp xỉ ma trận hiệp phương sai trên mô hình phân tích nhân tố phức tạp hơn.
- Có thể nói, phân tích nhân tố là mô tả mối quan hệ hiệp phương sai giữa nhiều biến số bằng số lượng ít hơn các nhân tố không quan sát được và giúp hiểu rõ hơn về cấu trúc của dữ liệu.
- Quá trình phân tích nhân tố thường bao gồm các bước sau:
  1. Thu thập dữ liệu: Thu thập dữ liệu về các biến quan sát từ mẫu đối tượng hoặc đơn vị.
  2. Ma trận hiệp phương sai/tương quan: Tính ma trận hiệp phương sai hoặc tương quan của các biến quan sát.
  3. Trích xuất nhân tố: Sử dụng một phương pháp như Phân tích thành phần chính (PCA) hoặc phương pháp ước lượng hợp lý cực đại.
  4.Quay vòng: Quay vòng các yếu tố để cải thiện khả năng diễn giải của kết quả. Các phương pháp quay phổ biến bao gồm: Varimax, Promax,..
  5. Hệ số tải: Khảo sát hệ số tải để xác định mối quan hệ giữa các nhân tố và biến quan sát.
  6. Điểm nhân tố: Diễn giải các nhân tố dựa trên mô hình hệ số tải cao và đặt tên theo các biến có hệ số tải cao trên từng nhân tố.

# Chương 2

# Mô hình nhân tố trực giao

Xét vector ngẫu nhiên có thể quan sát được $X^\top = (X_1, \ldots, X_p)$ có $E(X) = \mu$, $cov(X) = \Sigma$. Mô hình nhân tố giả định rằng mỗi $X_i$ là tổ hợp tuyến tính của một số ít biến ngẫu nhiên không quan sát được $F_1, \ldots, F_m$ (với $m < p$) gọi là các nhân tố chung và $p$ biến cộng thêm $\epsilon_1, \ldots, \epsilon_p$ được gọi là các sai số hoặc các nhân tố xác định.

Ta có mô hình nhân tố trực giao:

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \epsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \epsilon_2 \\ &\vdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \epsilon_p \end{aligned} \tag{2.1}$$

hoặc dưới dạng ma trận:

$$\underset{(p\times 1)}{X - \mu} = \underset{(p\times m)}{L} \times \underset{(m\times 1)}{F} + \underset{(p\times 1)}{\epsilon} \tag{2.2}$$

Trong đó, ma trận $L$ có kích thước $p \times m$ và được gọi là ma trận hệ số tải nhân tố. Các phần tử của ma trận $L$ là $l_{ij}$, là các hệ số tải. $F$ là nhân tố chung, $\epsilon$ là nhân tố xác định.
Việc tìm $F_1, \ldots, F_k$ và $\epsilon_1, \ldots, \epsilon_k$ là một bài toán không giải được. Tuy nhiên nếu thêm giả thiết:

$$E(F) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(m\times 1)} \quad ; \quad \text{cov}(F) = E(FF^\top) = I_{(m\times m)}$$

$$E(\varepsilon) = 0; \quad \text{cov}(\varepsilon) = E(\varepsilon\varepsilon^\top) = \Psi_{p\times p} = \begin{bmatrix} \Psi_1 & 0 & \ldots & 0 \\ 0 & \Psi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Psi_p \end{bmatrix}$$

$F$ và $\varepsilon$ độc lập (không tương quan):

$$\text{cov}(\varepsilon, F) = E(\varepsilon F^\top) = 0_{(p\times m)}$$

thì bài toán sẽ có lời giải.

Như vậy ta cần xét mô hình nhân tố trực giao dưới đây:

$$\underset{(p\times 1)}{X-\mu} = \underset{(p\times m)}{L} \times \underset{(m\times 1)}{F} + \underset{(p\times 1)}{\epsilon} \tag{2.3}$$

với $\mu_i = E(X_i)$
$\epsilon_i$ là nhân tố xác định thứ $i$
$F_j$ là nhân tố chung thứ $j$
$l_{ij}$ là tải trọng của biến $X_i$ đặt lên nhân tố thứ $j$ $(Y_j)$.
$E(F) = 0; E(\epsilon) = 0$

$$cov(F) = I; cov(\epsilon) = \psi = diag(\psi_1, ..., \psi_k) \tag{2.4}$$

$cov(\epsilon, F) = 0$

Ta có:

$$\begin{aligned}(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^\top &= (\mathbf{LF}+\boldsymbol{\varepsilon})(\mathbf{LF}+\boldsymbol{\varepsilon})^\top \\ &= (\mathbf{LF}+\boldsymbol{\varepsilon})((\mathbf{LF})^\top+\boldsymbol{\varepsilon}^\top) \\ &= \mathbf{LF}(\mathbf{LF})^\top+\boldsymbol{\varepsilon}(\mathbf{LF})^\top+\mathbf{LF}\boldsymbol{\varepsilon}^\top+\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \\ \Sigma = \mathrm{cov}(\mathbf{X}) &= E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^\top \\ &= \mathbf{L}E(\mathbf{FF}^\top)\mathbf{L}^\top+\mathbf{L}E(\mathbf{F}\boldsymbol{\varepsilon}^\top)+E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) \\ &= \mathbf{LL}^\top+\boldsymbol{\Psi}\end{aligned}$$

Suy ra nếu có mô hình trực giao trên thì:

$$cov(X) = \Sigma = LL^\top + \psi \tag{2.5}$$

Như vậy việc tìm cấu trúc mô hình nhân tố trực giao đưa về việc tách ma trận hiệp phương sai dưới dạng (2.5)

**Nhận xét:** Nếu $L$ thỏa mãn (2.5) thì $L^* = LT$ cũng thỏa mãn với $T$ là ma trận trực giao.
$\Sigma = LL^\top + \psi = LTT^\top L^\top + \psi = (LT)(T^\top L^\top) + \psi = (L^*)(L^*)^\top + \psi$
Như vậy hệ thức phân tích (2.5) xác định $L$ sai khác một phép biến đổi trực giao.
Từ (2.5) ta có:

$$\begin{aligned} D(X_i) &= l_{i1}^2 + \cdots + l_{im}^2 + \psi_i = \sigma_{ii} \\ cov(X_i, X_j) &= l_{i1}l_{j1} + \cdots + l_{im}l_{jm} = \sigma_{ij} \\ cov(X_i, F_j) &= l_{ij} \end{aligned} \tag{2.6}$$

$h_i^2 = l_{i1}^2 + \cdots + l_{im}^2$ gọi là phương sai chung,
$\psi_i$ được gọi là phương sai xác định.

Ta có: $\underbrace{\sigma_{ii}}_{D(X_i)} = \underbrace{l_{i1}^2 + \cdots + l_{im}^2}_{\text{phương sai chung}} + \underbrace{\psi_i}_{\text{phương sai xác định}}$

Tóm lại, ứng với mô hình nhân tố trực giao, ta có cấu trúc sau đây về hiệp phương sai:

1. $cov(X) = \Sigma = LL^\top + \psi$
   hoặc:

$$\begin{aligned} D(X_i) &= l_{i1}^2 + \cdots + l_{im}^2 + \psi_i \\ cov(X_i, X_j) &= l_{i1}l_{j1} + \cdots + l_{im}l_{jm} \end{aligned} \tag{2.7}$$

2. $cov(X, F) = L$
   hoặc:

$$cov(X_i, F_j) = l_{ij}$$

**Ví dụ 2.1.** Xét ma trận hiệp phương sai sau:

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

Dễ thấy rằng:

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

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

Quy trình phân rã này cũng diễn ra tương tự với các biến khác.

**Nhận xét:**

- Mô hình nhân tố giả định rằng $p + \frac{p(p-1)}{2} = \frac{p(p+1)}{2}$ phương sai và hiệp phương sai cho $X$ có thể được mô phỏng lại từ $pm$ hệ số tải $l_{ij}$ và $p$ phương sai xác định $\psi_i$. Khi $m = p$, bất kỳ ma trận hiệp phương sai $\Sigma$ nào cũng có thể được phân tách chính xác thành $LL^\top$ và $\psi$ có thể là ma trận không. Tuy nhiên, khi $m$ tương đối nhỏ so với $p$ thì phân tích nhân tố mới hữu dụng nhất. Trong trường hợp này, mô hình nhân tố trực giao sẽ đơn giản hóa ma trận hiệp phương sai với số lượng tham số ít hơn $\frac{p(p+1)}{2}$ tham số trong $\Sigma$.

- Ví dụ, nếu $X$ gồm $p = 12$ biến, và mô hình nhân tố với $m = 2$ thỏa mãn, thì khi đó $\frac{p(p+1)}{2} = \frac{12\times13}{2} = 78$ thành phần của $\Sigma$ được biểu diễn bởi $mp + p = 12 \times 2 + 12 = 36$ tham số $l_{ij}$ và $\Psi_i$ của mô hình nhân tố.

- Tuy vậy, hầu hết các ma trận hiệp phương sai không thể phân tích được thành $LL^\top + \Psi$, khi mà số lượng nhân tố $m$ nhỏ hơn nhiều so với $p$. Ví dụ tiếp theo sẽ chỉ ra một trong những vấn đề có thể xảy ra khi cố gắng xác định tham số $l_{ij}$ và $\Psi_i$ từ phương sai và hiệp phương sai của các biến quan sát được.

**Ví dụ 2.2.** (Không tồn tại nghiệm thích hợp) Cho $p = 3$ và $m = 1$, và các biến ngẫu nhiên $X_1, X_2, X_3$ có ma trận hiệp phương sai xác định dương là

$$\Sigma = \begin{bmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{bmatrix}$$

Áp dụng mô hình nhân tố, ta có:

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + \epsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + \epsilon_2 \\ X_3 - \mu_3 &= l_{31}F_1 + \epsilon_3 \end{aligned}$$

Cấu trúc hiệp phương sai cho ta:

$$\Sigma = LL^\top + \Psi \quad \text{hay} \quad \begin{cases} 1 = l_{11}^2 + \psi_1 \\ 0.9 = l_{11}l_{21} \\ 0.7 = l_{11}l_{31} \end{cases}$$

Ta có cặp phương trình:

$$\begin{cases} 0.7 = l_{11}l_{31} \\ 0.4 = l_{21}l_{31} \end{cases}$$

Suy ra $l_{21} = \frac{0.4}{0.7} l_{11}$
thay vào phương trình $0.9 = l_{11}l_{21}$ ta tính ra được: $l_{11}^2 = 1.575$, hay $l_{11} = \pm 1.255$.

Mà do $\text{Var}(F_1) = 1$ (theo giả thiết) và $\text{Var}(X_1) = 1, l_{11} = \text{Cov}(X_1, F_1) = \text{Corr}(X_1, F_1)$. Trị tuyệt đối của hệ số tương quan không thể lớn hơn 1, do đó $|l_{11}| = 1.255$ là quá lớn.

Ta cũng có:

$$1 = l_{11} + \psi_1, \text{hay} \quad \psi_1 = 1 - l_{11}$$

có:

$$\psi_1 = 1 - 1.575 = -0.575$$

Điều này là không thỏa mãn do phương sai không thể bé hơn 0.

Vì vậy, với ví dụ này với $m = 1$ ta có thể nhận được nghiệm duy nhất cho phương trình $\Sigma = LL^\top + \Psi$. Tuy nhiên, nghiệm này không phù hợp với biểu diễn thống kê, vì vậy nó không phải là nghiệm thích hợp. Khi $m > 1$, sẽ luôn có một thuộc tính ẩn gắn liền với mô hình nhân tố. Để thấy điều này, ta lấy $T$ là ma trận trực giao bất kỳ $m \times n$, có tính chất $TT^\top = T^\top T = I$. Khi đó ta có thể viết:

$$X - \mu = LF + \epsilon = LTT^\top F + \epsilon = L^*F^* + \epsilon$$

trong đó:

$$L^* = LT, \quad F^* = T^\top F$$

mà ta có:

$$E(F^*) = T^\top E(F) = 0$$

và:

$$\text{Cov}(F^*) = T^\top \text{Cov}(F)T = T^\top T = I$$

Do đó $F^* = T^\top F$ có các tính chất thống kê tương tự như $F$, mặc dù nhìn chung $L$ và $L^*$ là khác nhau, nhưng chúng đều tạo ra một ma trận hiệp phương sai $\Sigma$:

$$\Sigma = LL^\top + \Psi = LTT^\top L^\top + \Psi = (L^*)(L^*)^\top + \Psi$$

Tính chất ẩn này là nguồn gốc của "phép quay nhân tố", vì ma trận trực giao tương ứng với phép quay hệ tọa độ của $X$. Ta có:

$$L^* = LT \quad \text{và} \quad L^\top$$

đều có biểu diễn giống nhau. Phương sai chung được tạo ra bởi:

$$LL^\top = (L^*)(L^*)^\top$$

cũng không bị ảnh hưởng bởi cách chọn ma trận $T$.

# Chương 3

# Các phương pháp ước lượng

## 3.1 Phương pháp thành phần chính

### 3.1.1 Cở sở lý thuyết

- Giả sử ma trận hiệp phương sai $\underset{(p\times p)}{\sum}$ có các cặp trị riêng và vector riêng là:

$$(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$$

trong đó $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ và hệ vector riêng $e_1, e_2, \ldots, e_p$ trực chuẩn.

- Khi đó:

$$\begin{aligned}\sum &= \lambda_1 e_1 e_1^\top + \lambda_2 e_2 e_2^\top + \cdots + \lambda_p e_p e_p^\top \\ &= \begin{bmatrix}\sqrt{\lambda_1}e_1 & \sqrt{\lambda_2}e_2 & \ldots & \sqrt{\lambda_p}e_p\end{bmatrix} \begin{bmatrix}\sqrt{\lambda_1}e_1^\top \\ \sqrt{\lambda_2}e_2^\top \\ \vdots \\ \sqrt{\lambda_p}e_p^\top\end{bmatrix}\end{aligned} \tag{3.1}$$

$\sum = LL^\top$, trong đó $L = \begin{bmatrix}\sqrt{\lambda_1}e_1 & \sqrt{\lambda_2}e_2 & \ldots & \sqrt{\lambda_p}e_p\end{bmatrix}$

- Điều này phù hợp với cấu trúc hiệp phương sai quy định cho mô hình phân tích nhân tố có nhiều nhân tố là biến $(m = p)$ và phương sai cụ thể $\psi_i = 0$ với mọi $i$. Ma trận tải có cột thứ $j$ được biểu diễn bởi $\sqrt{\lambda_j}e_j$.

- Ta có:

$$\underset{(p\times p)}{\sum} = \underset{(p\times p)}{L}\underset{(p\times p)}{L^\top} + \underset{(p\times p)}{\psi}$$

$$\text{Khi } \psi = 0 \Rightarrow \sum = LL^\top \tag{3.2}$$

$$\rightarrow \text{Đây là mô hình p nhân tố cho } \underset{(p\times 1)}{X}$$

- Mặc dù đại diện cho phân tích nhân tố $\sum$ trong công thức (3.2) là chính xác, nhưng nó không đặc biệt hữu ích. Việc sử dụng càng nhiều nhân tố chung càng có nhiều biến số trong khi đó không cho phép bất kỳ sự thay đổi nào trong các nhân tố cụ thể $\varepsilon$.

- Người ta cố gắng đưa về mô hình nhân tố với $m < p$.
- Xét trường hợp $p-m$ giá trị riêng cuối cùng của $\sum$ là không đáng kể, tức là $\lambda_{m+1}, \ldots, \lambda_p$ gần 0. Chúng ta có thể bỏ qua sự đóng góp của các giá trị riêng $\lambda_{m+1}, \ldots, \lambda_p$ này cho $\sum$.
- Khi đó:

$$\sum \approx \lambda_1 e_1 e_1^\top + \cdots + \lambda_m e_m e_m^\top = \begin{bmatrix} \sqrt{\lambda_1} e_1 & \ldots & \sqrt{\lambda_m} e_m \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} e_1^\top \\ \vdots \\ \sqrt{\lambda_m} e_m^\top \end{bmatrix} = \underset{(p\times m)}{L} \underset{(m\times p)}{L^\top}$$

- Phương sai của các nhân tố cụ thể có thể được coi là các phần tử trên đường chéo của ma trận $\sum - LL^\top$

  Tức là ta có: $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2, i = 1, \ldots, p$

$$\Rightarrow \psi = \begin{pmatrix} \psi_1 & & 0 \\ & \ddots & \\ 0 & & \psi_p \end{pmatrix}$$

  $\Rightarrow \sum \approx LL^\top + \psi$

### 3.1.2 Thuật toán

Sử dụng các khái niệm cụ thể này để ước lượng $L$ và ma trận $\psi$ tương ứng từ dữ liệu. Áp dụng quy trình trên cho một tập dữ liệu nhất định gồm: $\underset{(p\times 1)}{x_1}, \underset{(p\times 1)}{x_2}, \ldots, \underset{(p\times 1)}{x_n}$.

(i) Tính vector trung bình mẫu quan sát được $\bar{x}$.

(ii) Tính các vector độ lệch $x_j - \bar{x} = \begin{bmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jp} \end{bmatrix} - \begin{bmatrix} \bar{x_1} \\ \bar{x_2} \\ \vdots \\ \bar{x_p} \end{bmatrix} = \begin{bmatrix} x_{j1} - \bar{x_1} \\ x_{j2} - \bar{x_2} \\ \vdots \\ x_{jp} - \bar{x_p} \end{bmatrix}, j = 1, \ldots, n$

(iii) Sử dụng các vector độ lệch này để tính ma trận hiệp phương sai mẫu $\underset{(p\times p)}{S}$

Trong trường hợp các đơn vị của các biến không tương xứng, thường là mong muốn thực hiện với các biến được chuẩn hóa $z$ sau:

$$z_j = \begin{bmatrix} \frac{x_{j1} - \bar{x_1}}{\sqrt{s_{11}}} \\ \frac{x_{j2} - \bar{x_2}}{\sqrt{s_{22}}} \\ \vdots \\ \frac{x_{jp} - \bar{x_p}}{\sqrt{s_{pp}}} \end{bmatrix}, j = 1, \ldots, n$$

có ma trận hiệp phương sai là $R$ và đó chính là ma trận tương quan mẫu của các quan sát $x_1, x_2, \ldots, x_n$. Tiêu chuẩn hóa tránh các vấn đề của việc có một biến với phương sai lớn ảnh hưởng quá mức đến việc tải nhân tố. Do các biến sau đều có kì vọng $\mu = 0$ và phương sai $\sigma^2 = 1$.

(iv) Phân tích nhân tố thành phần chính của ma trận hiệp phương sai mẫu $S$ được xác định theo các cặp trị riêng và vector riêng của $S$.
Giả sử các cặp trị riêng và vector riêng đó là $(\hat{\lambda_1}, \hat{e_1}), (\hat{\lambda_2}, \hat{e_2}), \ldots, (\hat{\lambda_p}, \hat{e_p})$, với $\hat{\lambda_1} \geq \hat{\lambda_2} \geq \cdots \geq \hat{\lambda_p}$.

(v) Có $m < p$ là số lượng các nhân tố chung. Khi đó ma trận tải nhân tố được ước lượng dưới dạng: $\hat{L} = \left[\sqrt{\hat{\lambda_1}}\hat{e_1} \quad \sqrt{\hat{\lambda_2}}\hat{e_2} \quad \ldots \quad \sqrt{\hat{\lambda_m}}\hat{e_m}\right]$

(vi) Ước lượng các phương sai cụ thể $\psi_i$ được cho bởi các phần tử trên đường chéo của ma trận $S - \hat{L}\hat{L}^T$, cụ thể là:

$$\hat{\psi_i} = s_{ii} - \sum_{j=1}^{m} \hat{l_{ij}}^2$$

$$\Rightarrow \hat{\psi} = \begin{pmatrix} \hat{\psi_1} & & 0 \\ & \ddots & \\ 0 & & \hat{\psi_p} \end{pmatrix}$$

(vii) Các phương sai chung được ước lượng bởi $\hat{h_i}^2 = \sum_{j=1}^{m} \hat{l_{ij}}^2$

### 3.1.3 Lưu ý

- **Lưu ý 1***: Đối với phân tích thành phần chính, ước lượng của tải trọng cho một nhân tố nhất định không thay đổi khi số lượng nhân tố tăng lên.*
  Ví dụ:

$$m = 1: \quad \hat{L}_{(1)} = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1\right]$$

$$m = 2: \quad \hat{L}_{(2)} = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1 \quad \sqrt{\hat{\lambda}_2}\hat{e}_2\right]$$

  Tổng quát:

$$m = k: \hat{L}_{(k)} = \left[\sqrt{\hat{\lambda_1}}\hat{e_1} \quad \ldots \quad \sqrt{\hat{\lambda}_k}\hat{e}_k\right]$$

$$m = k+1: \hat{L}_{(k+1)} = \left[\hat{L}_{(k)} \quad \vdots \quad \sqrt{\hat{\lambda}_{(k+1)}}\hat{e}_{(k+1)}\right]$$

- **Độ gần đúng của ước lượng**
  Ta có:

$$S \approx \hat{L}\hat{L}^\top + \hat{\psi}$$

  Gọi:

$$\Delta = S - (\hat{L}\hat{L}^\top + \hat{\psi}) = (\Delta_{ij})$$

  Kết quả:

$$\sum_{i,j} \Delta_{ij}^2 = tr(\Delta^2) \leq \sum_{i=m+1}^{p} \hat{\lambda_i}^2$$

  **Chứng minh:**

Bước 1: Ta có các phần tử trên đường chéo của ma trận $\Delta$ là 0.

Bước 2: Tổng bình phương các phần tử của $(S - \hat{L}\hat{L}^\top - \hat{\psi})$ nhỏ hơn hoặc bằng tổng bình phương các phần tử của $(S - \hat{L}\hat{L}^\top)$.
$\Rightarrow tr\Delta^2 (= \sum_{i,j} \Delta_{ij}^2)$ nhỏ hơn hoặc bằng tổng bình phương các phần tử của ma trận $(S - \hat{L}\hat{L}^\top)$

$$S = \begin{bmatrix} \sqrt{\hat{\lambda}_1}\hat{e_1} & \dots & \sqrt{\hat{\lambda}_p}\hat{e_p} \end{bmatrix} \begin{bmatrix} \sqrt{\hat{\lambda}_1}\hat{e_1}^\top \\ \vdots \\ \sqrt{\hat{\lambda}_p}\hat{e_p}^\top \end{bmatrix}$$

$$S = (\hat{\lambda_1}\hat{e_1}\hat{e_1}^\top + \dots + \hat{\lambda_m}\hat{e_m}\hat{e_m}^\top) + (\hat{\lambda_{m+1}}\hat{e_{m+1}}\hat{e_{m+1}}^\top + \dots + \hat{\lambda_p}\hat{e_p}\hat{e_p}^\top)$$

$$S = \hat{L}\hat{L}^\top + \sum_{j=m+1}^{p} \hat{\lambda_j}\hat{e_j}\hat{e_j}^\top$$

$$\Rightarrow \ ( \ S - \hat{L}\hat{L}^\top \ ) \ = \sum_{j=m+1}^{p} \hat{\lambda_j}\hat{e_j}\hat{e_j}^\top \tag{3.3}$$

Bước 3: $tr(\Delta^2) \le tr(S - \hat{L}\hat{L}^\top)^2$

$$\begin{aligned} tr(S - \hat{L}\hat{L}^\top)^2 &= tr[(S - \hat{L}\hat{L}^\top)(S - \hat{L}\hat{L}^\top)] \\ &= tr[(\sum_{j=m+1}^{p} \hat{\lambda_j}\hat{e_j}\hat{e_j}^\top)(\sum_{j=m+1}^{p} \hat{\lambda_j}\hat{e_j}\hat{e_j}^\top)] \\ &= tr[\sum_{j=m+1}^{p} (\hat{\lambda_j}\hat{e_j}\hat{e_j}^\top)(\hat{\lambda_j}\hat{e_j}\hat{e_j}^\top)] \\ &= tr(\sum_{j=m+1}^{p} \hat{\lambda_j}^2\hat{e_j}\hat{e_j}^\top) \\ &= \sum_{j=m+1}^{p} \hat{\lambda_j}^2 tr(\hat{e_j}\hat{e_j}^\top) \\ &= \sum_{j=m+1}^{p} \hat{\lambda_j}^2 tr(\hat{e_j}^\top\hat{e_j}) \\ &= \sum_{j=m+1}^{p} \hat{\lambda_j}^2 \end{aligned}$$

$$\Rightarrow tr\Delta^2 = \sum_{i,j} \Delta_{ij}^2 \le \sum_{j=m+1}^{p} \hat{\lambda_j}^2$$

- **Lưu ý 2:** *Sự đóng góp của các nhân tố vào phương sai mẫu*
  Ta có:
$$s_{ii} = \sum_{j=1}^{m} \hat{l_{ij}}^2 + \hat{\psi}_i$$

  (.) Sự đóng góp của nhân tố đầu tiên cho $s_{ii}$ là $\hat{l}_{i1}^2$.

(.) Sự đóng góp của nhân tố đầu tiên cho tổng phương sai mẫu $tr(S) = s_{11}+s_{22}+\cdots+s_{pp}$ là:

$$\hat{l_{11}} + \hat{l_{21}} + \cdots + \hat{l_{p1}} \tag{3.4}$$

Với: $\underset{(p\times m)}{\hat{L}} = \left[\sqrt{\hat{\lambda_1}}\hat{e_1} \quad \sqrt{\hat{\lambda_2}}\hat{e_2} \quad \ldots \quad \sqrt{\hat{\lambda_m}}\hat{e_m}\right]$ thì ta có cột thứ $j$: $\begin{bmatrix}\hat{l_{1j}}\\ \vdots \\ \hat{l_{pj}}\end{bmatrix} = \sqrt{\hat{\lambda_j}}\hat{e_j}$

Chẳng hạn $j = 1$ thì ta có cột đầu tiên của $\hat{L}$: $\begin{bmatrix}\hat{l_{11}}\\ \vdots \\ \hat{l_{p1}}\end{bmatrix}$

Như vậy:

$$\sum_{i=1}^{p} \hat{l_{ij}}^2 = (\sqrt{\hat{\lambda_j}}\hat{e_j})^\top(\sqrt{\hat{\lambda_j}}\hat{e_j}) = \hat{\lambda_j}\hat{e_j}^\top\hat{e_j} = \hat{\lambda_j}$$

(.) Ví dụ $j = 1 \Rightarrow \sum_{i=1}^{p} \hat{l_{i1}}^2 = \hat{\lambda_1}$

(.) Khi đó tỷ lệ tổng phương sai mẫu được giải thích thông qua nhân tố đầu tiên:

$$\frac{\hat{\lambda_1}}{tr(S)} = \frac{\hat{\lambda_1}}{(\sum_{i=1}^{p} s_{ii})}$$

(.) Tỷ lệ tổng phương sai mẫu được giải thích thông qua k nhân tố đầu tiên:

$$\frac{\sum_{i=1}^{k} \hat{\lambda_i}}{\sum_{i=1}^{p} s_{ii}}$$

**Ví dụ 3.1.** Phân tích nhân tố cho ưu tiên của người tiêu dùng

Trong một nghiên cứu ưu tiên của người tiêu dùng, trong một mẫu ngẫu nhiên của khách hàng đã được yêu cầu đánh giá một số thuộc tính của một sản phẩm mới. Các phản hồi, trên thang phân biệt ngữ nghĩa 7 điểm, đã được lập bảng và ma trận tương quan các thuộc tính được xây dựng.

$$R = \begin{bmatrix}
\text{Thuộc tính} & \text{Mùi vị} & \text{Giá thành} & \text{Hương vị} & \text{Ăn nhanh} & \text{Năng lượng} \\
\text{Mùi vị} & 1.00 & 0.02 & 0.96 & 0.42 & 0.01 \\
\text{Giá thành} & 0.02 & 1.00 & 0.13 & 0.71 & 0.85 \\
\text{Hương vị} & 0.96 & 0.13 & 1.00 & 0.50 & 0.11 \\
\text{Ăn nhanh} & 0.42 & 0.71 & 0.50 & 1.00 & 0.79 \\
\text{Năng lượng} & 0.01 & 0.85 & 0.11 & 0.79 & 1.00
\end{bmatrix}$$

Nhân xét: Chúng ta có thể thấy rằng các biến mùi vị và hương vị có hệ số tương quan khá cao, tương tự với biến giá thành và năng lượng. Từ trên, ta có mong muốn rằng quan hệ giữa các biến sẽ có thể được giải thích chỉ qua hai hoặc ba biến.

Tính toán các giá trị riêng của ma trận $R$, ta thấy hai thành phần đầu tiên $\lambda_1 = 2.85$ và $\lambda_2 = 1.81$ là hai giá trị riêng duy nhất lớn hơn 1. Khi đó:

$$\frac{\lambda_1 + \lambda_2}{(\sum_{i=1}^{5} \lambda_i)} = 0.93$$

Như vậy, tổng $\lambda_1$ và $\lambda_2$ chiếm 93% tổng phương sai của 5 biến đã chuẩn hóa. Ta có ma trận hệ số tải $L$:

$$L = \begin{bmatrix} & \sqrt{\lambda_1}e_1 & \sqrt{\lambda_2}e_2 \\ \text{Mùi vị} & 0.56 & 0.82 \\ \text{Giá thành} & 0.78 & -0.53 \\ \text{Hương liệu} & 0.65 & 0.75 \\ \text{Ăn nhanh} & 0.94 & -0.10 \\ \text{Năng lượng} & 0.80 & -0.54 \end{bmatrix}$$

Tuy nhiên các hệ số tải khá là cao. Ví dụ ở hương liệu có vẻ quan trọng với cả yếu tố 1 và yếu tố 2. Điều này không cung cấp một cách giải thích dữ liệu đơn giản và rõ ràng. Lý tưởng nhất là mỗi biến sẽ xuất hiện như một yếu tố đóng góp đáng kể cho một cột.

## 3.2 Phương pháp ước lượng hợp lý cực đại

- Ước lượng hợp lý cực đại (MLE - Maximum Likelihood Estimation) là một phương pháp trong thống kê dùng để ước lượng giá trị tham số của một mô hình xác suất dựa trên những dữ liệu quan sát được.
- Phương pháp này ước lượng các tham số nói trên bởi những giá trị làm cực đại hóa hàm hợp lý.

Nếu các nhân tố chung $F$ và các nhân tố cụ thể $\epsilon$ có thể được coi là có phân phối chuẩn, khi đó ước lượng hợp lý cực đại của nhân tố tải có thể đạt được.

Khi $F_j$ và $\epsilon_j$ đều có phân phối chuẩn, các quan sát $X_j - \mu = LF_j + \epsilon_j$ cũng là phân phối chuẩn, khi đó ta có hàm hàm hợp lý cực đại là:

$$\begin{aligned} L(\mu, \Sigma) =& (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} e^{-\left(\frac{1}{2}\right) \operatorname{tr}\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j - \overline{\mathbf{x}})(\mathbf{x}_j - \overline{\mathbf{x}})^\top + n(\overline{\mathbf{x}} - \boldsymbol{\mu})(\overline{\mathbf{x}} - \boldsymbol{\mu})^\top\right)\right]} \\ =& (2\pi)^{\frac{-(n-1)p}{2}} |\Sigma|^{-\frac{(n-1)}{2}} e^{-\left(\frac{1}{2}\right) \operatorname{tr}\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j - \overline{\mathbf{x}})(\mathbf{x}_j - \overline{\mathbf{x}})^\top\right)\right]} \\ & \times (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\left(\frac{n}{2}\right)(\overline{\mathbf{x}} - \boldsymbol{\mu})^\top \Sigma^{-1} (\overline{\mathbf{x}} - \boldsymbol{\mu})} \end{aligned} \tag{3.5}$$

phụ thuộc vào $L$ và $\Psi$ do $\Sigma = LL^\top + \Psi$

Mô hình này chưa được xác định cụ thể, do có nhiều phương án cho L có thể thực hiện được bằng các phép biến đổi trực giao. Điều mong muốn là làm cho L được xác định rõ ràng bằng cách áp đặt điều kiện thuận lợi, một trong số đó là điều kiện duy nhất

$$L^\top \Psi^{-1} L = \Delta \text{ là ma trận đường chéo} \tag{3.6}$$

Ước lượng hợp lý cực đại $L$ và $\Psi$ sẽ có được khi ta cực đại hóa hàm hợp lý bên trên. Chúng ta sẽ có những chương trình máy tính có sẵn để giải. Ta sẽ tổng hợp một số kết quả về ước lượng hợp lý cực đại.

**Kết quả 1.** *Với $X_1, X_2, \cdot, X_n$ là một mẫu ngẫu nhiên tuân theo $N_p(\mu, \Sigma)$. Ước lượng hợp lý cực đại $\hat{L}$ và $\hat{\Psi}$ thu được bằng việc tối đa hóa (3.5) với $L$ và $\Psi$ thỏa mãn điều kiện duy nhất (3.6). Các ước lượng này thỏa mãn*

$$\left(\hat{\Psi}^{-\frac{1}{2}} S_n \hat{\Psi}^{-\frac{1}{2}}\right)\left(\hat{\Psi}^{-\frac{1}{2}}\hat{L}\right) = \left(\hat{\Psi}^{-\frac{1}{2}}\hat{L}\right)(I + \hat{\Delta}) \tag{3.7}$$

*với $\hat{\Delta} = \hat{L}^\top \hat{\Psi}^{-1}\hat{L}$.*

Từ công thức (3.7), ta có cột thứ $j$ của ma trận $\hat{\Psi}^{-\frac{1}{2}}\hat{L}$ là vector riêng của ma trận $\hat{\Psi}^{-\frac{1}{2}} S_n \hat{\Psi}^{-\frac{1}{2}}$ tương ứng với giá trị riêng $1 + \hat{\Delta}_i$. Trong đó:

$$S_n = \frac{1}{n}\sum_{j=1}^{n}(X_i - \overline{X})(X_j - \bar{X})^\top = \frac{n-1}{n} S \text{ và } \hat{\Delta}_1 \geq \hat{\Delta}_2 \geq ... \geq \hat{\Delta}_m$$

Ngoài ra:

$$\hat{\psi}_i = \text{ phần tử đường chéo chính thứ } i \text{ của } S_n - \hat{L}\hat{L}^\top$$

và

$$\text{tr}(\hat{\Sigma}^{-1} S_n) = p$$

**Kết quả 2.** *Với $X_1, X_2, ..., X_n$ là mẫu ngẫu nhiên từ $N_p(\mu, \Sigma)$, ở đây $\Sigma = LL^\top + \Psi$ là ma trận hiệp phương sai của mô hình m nhân tố chung . Ước lượng hợp lý cực đại $L, \Psi$ và $\mu = \overline{x}$ với điều kiện*

$$L^\top \Psi^{-1} L = \Delta \text{ là ma trận đường chéo}$$

*Ước lượng hợp lý cực đại cho tính cộng đồng là:*

$$h_i{}^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 + \cdots + \hat{l}_{im}^2 \text{ với } i = 1, 2, \ldots, p$$

(3.8)

*hay*

$$\left(\begin{array}{c}\textit{Tỷ lệ đóng góp của nhân tố } j \\ \textit{trong tổng phương sai}\end{array}\right) = \frac{\widehat{l_{1j}^2} + \widehat{l_{2j}^2} + \ldots + \widehat{l_{pj}^2}}{s_{11} + s_{22} + \ldots + s_{pp}} \tag{3.9}$$

*Chứng minh.* Do $L$ và $\Psi$ thỏa mãn tính chất duy nhất (3.6) nên ước lượng hợp lý cực đại của $L$ và $\Psi$ lần lượt là $\hat{L}$ và $\hat{\Psi}$. Vì vậy, cộng đồng $h_i^2 = l_{i1}^2 + l_{i2}^2 + ... + l_{im}^2$ có ước lượng hợp lý cực đại là $\hat{h_i}^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 + ... + \hat{l}_{im}^2$ □

Ta chuẩn hóa các biến $Z = V^{-\frac{1}{2}}(X - \mu)$. Khi đó, ma trận hiệp phương sai $\rho$ của $Z$ có biểu diễn:

$$\rho = V^{-\frac{1}{2}}\Sigma V^{-\frac{1}{2}} = (V^{-\frac{1}{2}}L)(V^{-\frac{1}{2}}L)^\top + V^{-\frac{1}{2}}\Psi V^{-\frac{1}{2}}$$

Như vậy, ma trận $\rho$ được xác định tương tự như (2.5) với ma trận tải $L_z = V^{-\frac{1}{2}}L$ và ma trận phương sai cụ thể $\Psi_z = V^{-\frac{1}{2}}\Psi V^{-\frac{1}{2}}$.

Theo tính chất bất biến của ước lượng hợp lý cực đại, ước lượng hợp lý cực đại của $\rho$ là:

$$\begin{aligned}\hat{\rho} &= (\hat{V}^{-\frac{1}{2}}\hat{L})(\hat{V}^{-\frac{1}{2}}\hat{L})^\top + \hat{V}^{-\frac{1}{2}}\hat{\Psi}\hat{V}^{-\frac{1}{2}} \\ &= \hat{L}_z\hat{L}_z^\top + \hat{\Psi}_z^\top\end{aligned}$$

Ở đây $\hat{V}^{-\frac{1}{2}}$ và $\hat{L}$ lần lượt là ước lượng hợp lý cực đại của $V^{-\frac{1}{2}}$ và $L$ Như vậy, bất cứ khi nào phân tích hợp lý cực đại của ma trận tương quan, ta gọi:

$$\hat{h}_i^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 + ... + \hat{l}_{im}^2, i = 1, 2, ..., p$$

là ước lượng hợp lý cực đại của các cộng đồng. Ta đánh giá tầm quan trọng của các yếu tố trên cơ sở:

$$\left( \begin{array}{c} \text{Tỷ lệ đóng góp của nhân tố j} \\ \text{trong tổng phương sai} \end{array} \right) = \frac{\widehat{l_{1j}^2} + \widehat{l_{2j}^2} + \ldots + \widehat{l_{pj}^2}}{s_{11} + s_{22} + \ldots + s_{pp}} = \frac{\widehat{l_{1j}^2} + \widehat{l_{2j}^2} + \ldots + \widehat{l_{pj}^2}}{p}$$

Thông thường, các quan sát được chuẩn hóa và ma trận tương quan mẫu được phân tích nhân tố. Ma trận tương quan mẫu $\mathbf{R}$ được thay thế bởi $\frac{n-1}{n}\mathbf{S}$ trong hàm hợp lý và ước lượng hợp lý cực đại và $\hat{L}_z$ và $\hat{\Psi}_z$ được tính toán bằng máy tính.

Với ước lượng tải trọng $\hat{L}_z$ và phương sai cụ thể $\hat{\Psi}_z$ thu được từ $\mathbf{R}$, ta thấy rằng ước lượng hợp lý cực đại cho phân tích nhân tố của ma trận hiệp phương sai $\frac{n-1}{n}\mathbf{S}$ là $\hat{L} = \hat{V}^{\frac{1}{2}}\hat{L}_z$ và $\hat{\Psi} = \hat{V}^{\frac{1}{2}}\hat{\Psi}_z\hat{V}^{\frac{1}{2}}$ hoặc

$$\hat{l}_{ij} = \hat{l}_{z,ij}\sqrt{\hat{\sigma}_{ii}} \text{ và } \hat{\psi}_i = \hat{\psi}_{z,i}\hat{\sigma}_{ii}$$

ở đây $\hat{\sigma}_{ii}$ là phương sai mẫu được tính với ước số $n$.

**Ví dụ 3.2.** Phân tích nhân tố của dữ liệu giá cổ phiếu

Ta có bảng dữ liệu giá cổ phiếu dưới đây:

| Tuần | J P Morgan | Citibank | Wells Fargo | Royal Dutch Shell | Exxon Mobil |
|---|---|---|---|---|---|
| 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 | 0.00495 |
| 5 | 0.01082 | 0.00372 | 0.01103 | 0.02919 | 0.00399 |
| 6 | 0.01107 | 0.00956 | 0.00817 | 0.01296 | 0.00998 |
| 7 | 0.00443 | 0.00807 | 0.00807 | 0.03397 | 0.00376 |
| 8 | 0.03449 | 0.00908 | 0.00843 | 0.00000 | 0.00768 |
| 9 | -0.04449 | -0.00555 | 0.01125 | -0.00857 | 0.00693 |
| 10 | -0.00406 | 0.02099 | -0.00608 | 0.00980 | -0.01267 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 94 | 0.03732 | 0.03593 | 0.02528 | 0.01667 | 0.01733 |
| 95 | 0.0238 | 0.00311 | 0.00621 | 0.01225 | 0.00763 |
| 96 | 0.02568 | 0.01535 | 0.01068 | 0.02013 | 0.02156 |
| 97 | -0.03066 | 0.00332 | 0.02292 | -0.00767 | 0.00343 |
| 98 | -0.01544 | 0.02234 | -0.00692 | -0.00434 | 0.00157 |
| 99 | 0.01519 | 0.00461 | 0.00784 | 0.01563 | 0.00844 |
| 100 | 0.00089 | 0.00656 | 0.00255 | -0.00098 | 0.00408 |
| 101 | -0.00466 | 0.00626 | -0.00288 | 0.00494 | 0.00712 |
| 102 | -0.01079 | 0.00806 | 0.00806 | -0.00049 | 0.00121 |
| 103 | -0.01279 | 0.01469 | -0.00404 | -0.00964 | -0.00234 |

Dữ liệu giá cổ phiếu ở trên đã được phân tích lại với giả định mô hình nhân tố m = 2 và sử dụng ước lượng hợp lý cực đại. Ước lượng hệ số tải nhân tố, tính cộng đồng, phương sai cụ thể và tỷ lệ của tổng phương sai mẫu (chuẩn hóa) được giải thích bởi từng nhân tố được trình bày trong bảng sau:

| **Biến ngẫu nhiên** | **Hợp lý cực đại** | | | **Thành phần chính** | | |
|---|---|---|---|---|---|---|
| | **Ước lượng nhân tố tải** | | **Phương sai cụ thể** $\hat{\psi}_i = 1 - \hat{h}_i^2$ | **Ước lượng nhân tố tải** | | **Phương sai cụ thể** $\tilde{\psi} = 1 - \tilde{h}_i^2$ |
| | **F1** | **F2** | | **F1** | **F2** | |
| 1. J P Morgan | .115 | .755 | .42 | .732 | -.437 | .27 |
| 2. Citibank | .322 | .788 | .27 | .831 | -.280 | .23 |
| 3. Wells Fargo | .182 | .652 | .54 | .726 | -.374 | .33 |
| 4. Royal Dutch Shell | 1.000 | -.000 | .00 | .605 | .694 | .15 |
| 5. Texaco | .683 | -.032 | .53 | .563 | .719 | .17 |
| **Giải thích tỷ lệ tích lũy (chuẩn hóa)** | .323 | .647 | | .487 | .769 | |

Ma trận sai số tương ứng với phương pháp thành phần chính là:

$$R - \tilde{L}\tilde{L}' - \tilde{\Psi} = \begin{bmatrix} 0 & -0.099 & -0.185 & -0.025 & 0.056 \\ -0.099 & 0 & -0.134 & 0.014 & -0.054 \\ -0.185 & -0.134 & 0 & 0.003 & 0.006 \\ -0.025 & 0.014 & 0.003 & 0 & -0.156 \\ 0.056 & -0.054 & 0.006 & -0.156 & 0 \end{bmatrix}$$

Ma trận sai số tương ứng với ước lượng hợp lý cực đại là :

$$R - \hat{L}\hat{L}' - \hat{\Psi} = \begin{bmatrix} 0 & 0.001 & -0.002 & 0.000 & 0.052 \\ 0.001 & 0 & 0.002 & 0.000 & -0.033 \\ -0.002 & 0.002 & 0 & 0.000 & 0.001 \\ 0.000 & 0.000 & 0.000 & 0 & 0.000 \\ 0.052 & -0.033 & 0.001 & 0.000 & 0 \end{bmatrix}$$

## 3.3 Kiểm định mẫu lớn cho số lượng các nhân tố chung

Giả sử ta có mô hình $m$ nhân tố chung giữ nguyên. Trong trường hợp $\Sigma = LL^\top + \Psi$, kiểm định tính đầy đủ của mô hình $m$ nhân tố chung tương đương với kiểm định giả thuyết:

$$H_0 : \underset{(p\times p)}{\Sigma} = \underset{(p\times m)}{L}\,\underset{(m\times p)}{L^\top} + \underset{(p\times p)}{\Psi}$$

và đối thuyết $H_1 : \Sigma$ là bất kì ma trận xác định dương nào khác.
Khi $\Sigma$ không có các điều kiện đặc biệt, giá trị lớn nhất của hàm hợp lý (với $\hat{\Sigma} = \frac{n-1}{n}S = S_n$) tỷ lệ thuận với:

$$|S_n|^{-\frac{n}{2}} e^{-\frac{np}{2}}$$

Theo $H_0, \Sigma = LL^\top + \Psi$, trong trường hợp này, thay $\hat{\mu} = \overline{x}$ vào công thức hàm hợp lý cực đại tổng quát, ở đây $\hat{L}$ và $\hat{\Psi}$ lần lượt là ước lượng hợp lý cực đại của $L$ và $\Psi$, ta có giá trị lớn nhất của hàm hợp lý tỷ lệ thuận với:

$$\begin{aligned}|\hat{\Sigma}|^{-n/2} \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\hat{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j - \overline{\mathbf{x}})(\mathbf{x}_j - \overline{\mathbf{x}})^\top\right)\right]\right\} \\ = \left|\hat{\mathbf{L}}\hat{\mathbf{L}}^\top + \hat{\Psi}\right|^{-n/2} \exp\left\{-\frac{1}{2} n \operatorname{tr}\left[(\hat{\mathbf{L}}\hat{\mathbf{L}} + \hat{\Psi})^{-1}\mathbf{S}_n\right]\right\}\end{aligned}$$

Kết hợp hai điều trên, ta tìm được thống kê tỷ lệ khả năng cho kiểm tra $H_0$ là:

$$\begin{aligned}-2\ln\Lambda &= -2\ln\left[\frac{\text{Cực đại hợp lý ở } H_0}{\text{Cực đại hợp lý}}\right] \\ &= -2\ln\left(\frac{|\hat{\Sigma}|}{|\mathbf{S}_n|}\right)^{-n/2} + n\left[\operatorname{tr}\left(\hat{\Sigma}^{-1}\mathbf{S}_n\right) - p\right]\end{aligned}$$

với bậc tự do:

$$\begin{aligned}v - v_0 &= \frac{1}{2}p(p+1) - [p(m+1) - \frac{1}{2}m(m-1)] \\ &= \frac{1}{2}[(p-m)^2 - p - m]\end{aligned}$$

Do $tr(\hat{\Sigma}^{-1}S_n) - p = 0$ nên $\hat{\Sigma} = \hat{L}\hat{L}^\top + \hat{\Psi}$ là ước lượng hợp lý cực đại của $\Sigma = LL^\top + \Psi$. Như vậy, ta có:

$$-2\ln\Lambda = n\ln\frac{|\hat{\Sigma}|}{|S_n|}$$

Bartlett đã chỉ ra rằng xấp xỉ Khi-bình phương đối với phân phối lấy mẫu của $-2\ln\Lambda$ trong công thức trên có thể được cải thiện bằng cách thay thế $n$ bằng hệ số nhân $\left(n - 1 - \frac{2p+4m+5}{6}\right)$. Sử dụng hiệu chỉnh Bartlett, ta bác bỏ $H_0$ ở mức ý nghĩa $\alpha$ nếu

$$\left(n - 1 - \frac{2p+4m+5}{6}\right)\ln\frac{|\hat{L}\hat{L}^\top + \hat{\Psi}|}{|S_n|} > \chi^2_{\frac{(p-m)^2-p-m}{2}}(\alpha) \tag{3.10}$$

với $n$ và $n-p$ lớn. Do số bậc tự do $\frac{(p-m)^2-p-m}{2} > 0$ nên

$$m < \frac{1}{2}\left(2p + 1 - \sqrt{8p+1}\right)$$

**Bình luận**

Trong việc thực hiện kiểm định ở công thức (3.10), chúng ta đang kiểm tra sự phù hợp của mô hình $m$ nhân tố chung bằng cách so sánh các phương sai tổng quát $|\hat{L}\hat{L}^\top + \hat{\Psi}|$ và $|S_n|$. Nếu $n$ (số lượng mẫu) lớn và $m$ nhỏ so với $p$ (số lượng biến quan sát), giả thuyết $H_0$ thường sẽ bị bác bỏ, dẫn đến việc giữ lại nhiều nhân tố chung hơn.

Tuy nhiên, $\hat{\Sigma} = \hat{L}\hat{L}^\top + \hat{\Psi}$ có thể đủ gần với $S_n$ để việc thêm nhiều nhân tố hơn không cung cấp thêm thông tin chi tiết, mặc dù những nhân tố đó có thể là "đáng kể". Cần phải thận trọng trong việc lựa chọn số nhân tố $m$.

**Ví dụ 3.3.** Kiểm định hai nhân tố chung

Ta xét dữ liệu giá cổ phiếu được trình bày trong ví dụ 3.2. Kiểm tra giả thuyết $H_0 : \Sigma = LL^\top + \Psi$ với $m = 2$ ngưỡng $\alpha = 0.05$.

Thống kê kiểm định trong (3.10) dựa trên tỷ lệ phương sai tổng quát:

$$\frac{|\hat{\Sigma}|}{|S_n|} = \frac{|\hat{L}\hat{L}^\top + \hat{\Psi}|}{|S_n|}$$

Với $\hat{V}^{-\frac{1}{2}}$ là ma trận đường chéo thỏa mãn $\hat{V}^{-\frac{1}{2}} S_n \hat{V}^{-\frac{1}{2}} = R$. Áp dụng các tính chất của định thức, ta có:

$$\begin{aligned}\frac{|\hat{\Sigma}|}{|S_n|} &= \frac{|\hat{V}^{-\frac{1}{2}}||\hat{L}\hat{L}^\top + \hat{\Psi}||\hat{V}^{-\frac{1}{2}}|}{|\hat{V}^{-\frac{1}{2}}||S_n||\hat{V}^{-\frac{1}{2}}|} \\ &= \frac{|\hat{V}^{-\frac{1}{2}}\hat{L}\hat{L}^\top\hat{V}^{-\frac{1}{2}} + \hat{V}^{-\frac{1}{2}}\hat{\Psi}\hat{V}^{-\frac{1}{2}}|}{|\hat{V}^{-\frac{1}{2}}||S_n||\hat{V}^{-\frac{1}{2}}|} \\ &= \frac{|\hat{L}_z\hat{L}_z^\top + \hat{\Psi}_z|}{|R|}\end{aligned}$$

Từ ví dụ 3.2, ta tính được:

$$\frac{\left|\hat{L}_z\hat{L}_z^\top+\hat{\Psi}_z\right|}{|R|} = \frac{\begin{vmatrix} 1.000 & & & & \\ 0.632 & 1.000 & & & \\ 0.513 & 0.572 & 1.000 & & \\ 0.115 & 0.322 & 0.182 & 1.000 & \\ 0.103 & 0.246 & 0.146 & 0.683 & 1.000 \end{vmatrix}}{\begin{vmatrix} 1.000 & & & & \\ 0.632 & 1.000 & & & \\ 0.510 & 0.574 & 1.000 & & \\ 0.115 & 0.322 & 0.182 & 1.000 & \\ 0.154 & 0.213 & 0.146 & 0.683 & 1.000 \end{vmatrix}} = \frac{0.17898}{0.17519} = 1.0216$$

Áp dụng hiệu chỉnh Bartlett, ta đánh giá thống kê kiểm tra trong (3.10):

$$\begin{aligned}&\left[n - 1 - \frac{2p + 4m + 5}{6}\right] \ln \frac{|\hat{L}\hat{L}^\top + \hat{\Psi}|}{|S_n|} \\ &= \left[103 - 1 - \frac{10 + 8 + 5}{6}\right] \ln(1.0216) = 2.10\end{aligned}$$

Do $\frac{1}{2}\left[(p-m)^2 - p - m\right] = \frac{1}{2}\left[(5-2)^2 - 5 - 2\right] = 1$, không vượt qua giá trị tới hạn $\chi_1^2(0.05) = 3.84$ nên giả thuyết $H_0$ không bị bác bỏ. Ta kết luận rằng dữ liệu không mâu thuẫn với mô hình hai nhân tố. Thật vậy, mức ý nghĩa quan sát được, hoặc P-value, $P(\chi_1^2 > 2.10) = 0.15$ ngụ ý rằng $H_0$ sẽ không bị bác bỏ ở bất kì mức ý nghĩa nào.

# Chương 4

# Xoay nhân tố

## 4.1 Đặt vấn đề

Xét ví dụ sau:

Bảng thể hiện các yếu tố ảnh hưởng đến trải nghiệm học tập

$$L = \begin{bmatrix} & F_1 & F_2 \\ \text{Chất lượng giáo viên} & 0.70 & 0.20 \\ \text{Tiện ích cho sinh viên} & 0.40 & -0.40 \\ \text{Dịch vụ hỗ trợ sinh viên} & 0.60 & -0.75 \\ \text{Bài giảng} & 0.90 & 0.30 \\ \text{Cơ sở vật chất} & 0.50 & -0.65 \end{bmatrix}$$

Ta có thấy trong bảng số liệu các hệ số tải của các nhân tố biểu hiện khá cao

Ví dụ: ở yếu tố "Dịch vụ hỗ trợ sinh viên"có vẻ quan trọng với cả yếu tố 1 và 2. Điều này không cung cấp một cách giải thích dữ liệu đơn giản và rõ ràng.

⇒ Do đó có thể sử dụng một phép biến đổi để làm rõ các đặc điểm mà nhân tố được gọi tên và có thể nghiên cứu rõ ràng hơn về các mặt của nhân tố đó.

Lý tưởng nhất, mỗi biến sẽ xuất hiện với hệ số tải cao trong một cột duy nhất, biểu thị mối liên hệ chủ yếu với một nhân tố, các hệ số tải đối với các nhân tố khác chỉ ở mức thấp đến trung bình. Để đạt được điều này, ta có thể áp dụng một phép quay nhằm giúp các hệ số tải của các nhân tố trở nên dễ hiểu hơn, làm cho mỗi biến được liên kết chủ yếu với một nhân tố duy nhất. Phép quay như vậy sẽ giúp tối ưu hóa khả năng diễn giải của các nhân tố và làm rõ hơn sự phân biệt giữa chúng.

## 4.2 Phép xoay nhân tố

Như chúng ta đã chỉ ra trong (2.2) (Mô hình nhân tố trực giao), tất cả các tải nhân tố thu được từ các tải ban đầu bằng một ***phép biến đổi trực giao*** đều có khả năng tái tạo lại ma trận hiệp phương sai (hoặc ma trận tương quan) như nhau (2.3).
Từ đại số ma trận, chúng ta biết rằng một phép biến đổi trực giao thì tương ứng với một phép quay cứng nhắc tọa độ của các trục.

**Định nghĩa 4.1.** Phép xoay nhân tố là phép biến đổi tạo ra ma trận tải nhân tố quay vòng $L^*$ để có một cách giải thích dữ liệu đơn giản và dễ dàng hơn.

Phép xoay nhân tố có 2 loại:

- Phép xoay trực giao: Trong đó các nhân tố chung không tương quan với nhau
- Phép xoay xiên: Trong đó các nhân tố chung có tương quan với nhau

$\Longrightarrow$ Phép xoay xiên thích hợp hơn phép xoay trực giao, bởi nó có xu hướng cung cấp các mẫu tải nhân tố dễ hiểu hơn mà không có các hạn chế phi thực tế rằng các nhân tố phổ biến không tương quan với nhau.

**Định nghĩa 4.2.** Ma trận của các tải xoay Nếu $L$ là ma trận $p \times m$ của hệ số tải ước tính thu được bằng bất kì phương pháp nào (thành phần chính hay ước lượng hợp lý cực đại) thì:

$$L^* = LT \text{ trong đó } TT^\top = T^\top T = 1 \tag{4.1}$$

$L^*$ là một ma trận $p \times m$ của các tải "xoay".

Hơn nữa, ma trận hiệp phương sai (hoặc tương quan) ước tính vẫn không thay đổi vì:

$$\hat{L}\hat{L^\top} + \hat{\Psi} = \hat{L}TT^\top\hat{L} + \hat{\Psi} = \hat{L}^*\hat{L}^{*T} + \hat{\Psi} \tag{4.2}$$

Phương trình (4.2) chỉ ra rằng ma trận dư:

$$S_n - \hat{L}\hat{L}^\top - \hat{\Psi} = S_n - \hat{L}^*\hat{L}^{*T} - \hat{\Psi}$$

không đổi sau khi thực hiện xoay nhân tố. Hơn nữa các phương sai cụ thể $\hat{\Psi}_i$, và cộng đồng $\hat{h}_i$ không thay đổi.
$\Rightarrow$ Ngoài ra, các phương sai riêng $\hat{\Psi}_i$ và các giá trị cộng đồng $\hat{h}_i^2$ cũng vẫn giữ nguyên. Do đó, từ góc độ toán học, việc chọn ma trận tải trọng ban đầu $\hat{L}\hat{L}^T$ hay ma trận xoay $\hat{L}^*\hat{L}^{*T}$ không có sự khác biệt. Tuy nhiên, vì các tải trọng ban đầu có thể khó diễn giải, nên thường người ta sẽ xoay chúng để đạt được một "cấu trúc đơn giản hơn".
$\Longrightarrow$ Lý tưởng nhất là chúng ta muốn thấy một mô hình tải trọng sao cho mỗi biến chịu tải trọng cao đối với một nhân tố duy nhất và có tải trọng nhỏ đến trung bình đối với các nhân tố còn lại.

## 4.3 Phương pháp xoay nhân tố trực giao

Chúng ta sẽ tập trung vào các phương pháp đồ thị và phân tích để xác định vào phép quay trực giao cho một cấu trúc đơn giản.

- Khi $m = 2$ hoặc các thừa số chung được coi là hai nhân tử cùng một lúc, sự biến đổi thành một cấu trúc đơn giản thường có thể được xác định bằng đồ thị.

Các yếu tố không tương quan được coi là các vector đơn vị dọc theo các trục tọa độ vuông góc. Biểu đồ của các cặp hệ số tải $(C)$ tạo thành $p$ điểm mỗi điểm tương ứng với một biến, sau đó tọa độ có thể được quay một cách trực giao thông qua một góc - gọi nó là $\Phi$ và tải trọng mới $l_{ij}^*$ được xác định từ các mối quan hệ:

$$\underset{(p\times 2)}{\hat{L}^*} = \underset{(p\times 2)}{\hat{L}}\underset{(2\times 2)}{T} \tag{4.3}$$

Trong đó:

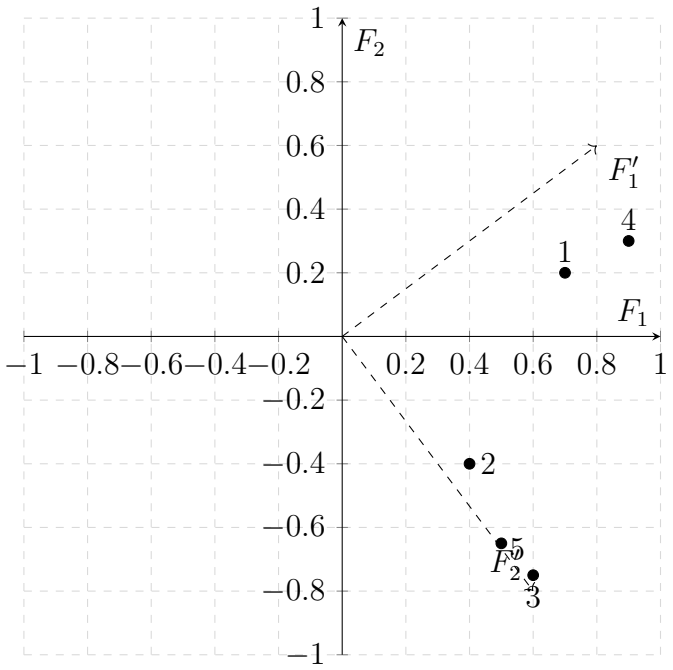
$$\begin{cases} T = \begin{bmatrix} \cos\Phi & \sin\Phi \\ -\sin\Phi & \cos\Phi \end{bmatrix} & \text{xoay theo chiều kim đồng hồ} \\ T = \begin{bmatrix} \cos\Phi & -\sin\Phi \\ \sin\Phi & \cos\Phi \end{bmatrix} & \text{xoay ngược chiều kim đồng hồ} \end{cases}$$

- Đối với $m > 2$, các định hướng không dễ dàng hình dung được và độ lớn của các tải trọng quay phải được kiểm tra để tìm ra cách giải thích có ý nghĩa đối với dữ liệu gốc.
  Việc lựa chọn một ma trận trực giao $T$ thỏa mãn phép đo, phân tích cấu trúc đơn giản sẽ được xem xét ngay sau đây:

**Ví dụ 4.1.** Chúng ta xét lại ví dụ (4.1) để hiểu thêm về phương pháp quay nhân tố:

$$L = \begin{bmatrix} & F_1 & F_2 \\ \text{Chất lượng giáo viên} & 0.70 & 0.20 \\ \text{Tiện ích cho sinh viên} & 0.40 & -0.40 \\ \text{Dịch vụ hỗ trợ sinh viên} & 0.60 & -0.75 \\ \text{Bài giảng} & 0.90 & 0.30 \\ \text{Cơ sở vật chất} & 0.50 & -0.65 \end{bmatrix}$$

$\Longrightarrow$ Chúng ta sẽ tìm cách xoay nhân tố sao cho mỗi biến sẽ xuất hiện như một yếu tố đóng góp đáng kể trong một cột. Để có thể giải thích dữ liệu một cách đơn giản và rõ ràng hơn.




Sau khi thực hiện phép xoay các nhân tố quay trục , chúng ta có thể quan sát được các cặp tải nhân tố trên hệ tọa độ $OF_1, F_2$ và $OF_1', F_2'$ , hệ tọa độ sau khi xoay

$$L = \begin{bmatrix} & F_1 & F_2 \\ \text{Chất lượng giáo viên} & 0.68 & 0.26 \\ \text{Tiện ích cho sinh viên} & 0.08 & 0.56 \\ \text{Dịch vụ hỗ trợ sinh viên} & 0.03 & 0.96 \\ \text{Bài giảng} & 0.90 & 0.30 \\ \text{Cơ sở vật chất} & 0.02 & 0.86 \end{bmatrix}$$

Yếu tố 1 có tải trọng cao với Chất lượng giáo viên (0.68) và Bài giảng (0.90), cho thấy rằng những yếu tố này đóng vai trò quan trọng trong việc nâng cao trải nghiệm học tập của sinh viên. Có thể đặt tên nhân tố F1 là "Chất lượng giảng dạy".

Yếu tố 2 có tải trọng cao với Tiện ích sinh viên (0.56), Cơ sở vật chất (0.86) và Dịch vụ hỗ trợ sinh viên (0.96). Có thể đặt tên yếu tố F2 là "Cơ sở hạ tầng cho sinh viên"

Từ đó, ta rút được chiều của ma trận từ 5 yếu tố còn 2 yếu tố.

**Ví dụ 4.2.** Ma trận tương quan mẫu về các nguyên nhân gây ra tình trạng tắc đường p = 6 :

$$\mathbf{R} = \begin{bmatrix} 1.0 & 0.439 & 0.410 & 0.288 & 0.329 & 0.248 \\ 0.439 & 1.0 & 0.351 & 0.354 & 0.320 & 0.329 \\ 0.410 & 0.351 & 1.0 & 0.164 & 0.190 & 0.181 \\ 0.288 & 0.354 & 0.164 & 1.0 & 0.595 & 0.470 \\ 0.329 & 0.320 & 0.190 & 0.595 & 1.0 & 0.464 \\ 0.248 & 0.329 & 0.181 & 0.470 & 0.464 & 1.0 \end{bmatrix}$$

1. Tín hiệu giao thông: Hệ thống đèn đường biển bảng chưa hợp lý
2. Mật độ dân cư: Khu vực đông dân/thưa dân quyết định việc tắc khá nhiều, VD: ở các vùng đô thị/thành phố sẽ dễ tắc hơn ở vùng quê/nông thôn
3. Đường hẹp: Đường xá nhỏ do thiếu quy hoạch/bố trí đường chưa hiệu quả
4. Thời tiết: Ảnh hưởng của thời tiết (như mưa, tuyết) dẫn đến việc không di chuyện được gây tắc đường.
5. Tai nạn, sự cố: Một số tình huống không may xảy ra phải quây lại khu vực hiện trường
6. Sự kiện lớn: Các ngày lễ, sự kiện có khả năng làm gia tăng lưu lượng giao thông, chẳng hạn như lễ hội hoặc buổi hòa nhạc, đi bão,...

Thực hiện phương pháp ước lượng hợp lý cực đại với $m = 2$ yếu tố theo dữ liệu đã cho thu được các ước tính trong bảng đưới đây:

Tải trọng ước lượng và giá trị cộng đồng

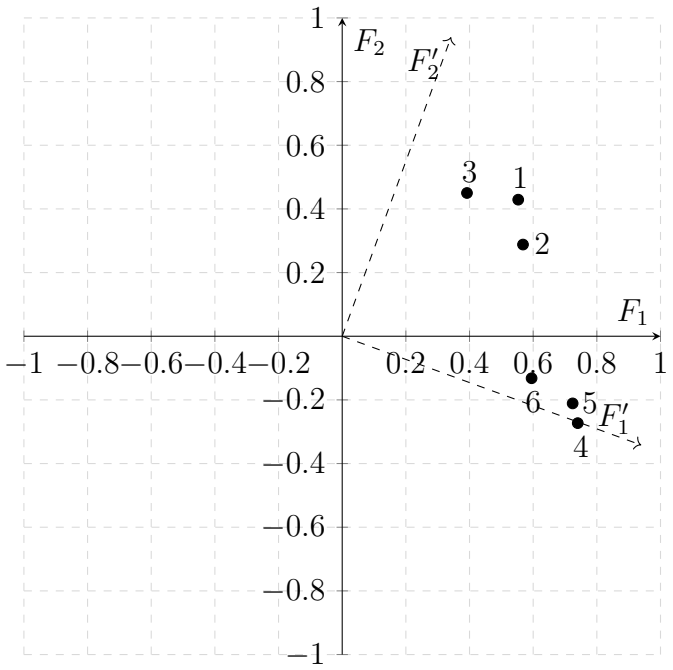
| Biến | Tải trọng ước lượng | | Giá trị cộng đồng |
|---|---|---|---|
| | $F_1$ | $F_2$ | $h^2$ |
| Tín hiệu giao thông | 0.553 | 0.429 | 0.490 |
| Mật độ dân cư | 0.568 | 0.288 | 0.406 |
| Đường hẹp | 0.392 | 0.450 | 0.356 |
| Thời tiết | 0.740 | -0.273 | 0.623 |
| Tai nạn, sự cố | 0.724 | -0.211 | 0.569 |
| Sự kiện lớn | 0.595 | -0.132 | 0.372 |

Từ kết quả của bảng dữ liệu ta có các nhận xét sau:

- Tất cả các biến đổi ở F1 đều có tải dương lên nhân tố đầu tiên. Ta có thể dự đoán yếu tố này ảnh hưởng chung tới việc tắc đường nhưng ta khó đặt tên cho yếu tố này.
- Đối với nhân tố thứ 2 ta thấy một nửa số tải là dương và một nửa là âm. Yếu tố dạng này là tùy ý bởi vì các dấu hiệu của tải được gọi là một yếu tố "lưỡng cực".

Phép xoay được chọn theo dữ liệu gốc là phép quay trực giao theo chiều kim đồng hồ của các trục tọa độ qua góc $\Phi = 20^o$ sao cho một trong các trục mới đi qua $(\hat{l_{41}}, \hat{l_{42}})$

$\Rightarrow$ Tất cả các điểm đều nằm trong góc phần tư thứ nhất và hai nhóm điểm riêng biệt được định nghĩa rõ ràng hơn.




Bảng 4.1: Tải trọng ước lượng và giá trị cộng đồng đã được xoay

| Biến | Tải trọng ước lượng đã xoay | | Giá trị cộng đồng |
|---|---|---|---|
| | $F_1^*$ | $F_2^*$ | $h\hat{*}_i^2 = \hat{h_i^2}$ |
| Tín hiệu giao thông | 0.369 | 0.594 | 0.490 |
| Mật độ dân cư | 0.433 | 0.467 | 0.406 |
| Đường hẹp | 0.211 | 0.558 | 0.356 |
| Thời tiết | 0.789 | 0.001 | 0.623 |
| Tai nạn, sự cố | 0.752 | 0.054 | 0.569 |
| Sự kiện lớn | 0.604 | 0.083 | 0.372 |

**Nhận xét kết quả:**

- Các biến "Thời tiết","Tại nạn, sự cố"và "Sự kiện lớn"có xu hướng có tải cao hơn trên $F1^*$ và có tải không đáng kể trên $F2^*$; có thể đặt F1* là yếu tố ngoại cảnh biến động
- Trong khi các biến "Tín hiệu giao thông","Mật độ dân cư","Đường hẹp"có xu hướng có tải cao hơn trên $F2^*$; có thể gọi $F2^*$ là yếu tố cố định (phụ thuộc vào mức quy hoạch đô thị)

# 4.4 Các phương pháp xoay nhân tố phân tích

## 4.4.1 Giới thiệu

- Phương pháp xoay nhân tố khách quan đầu tiên được đưa ra bởi Carroll (1953), mặc dù một số phương pháp đã được đưa ra ngay sau đó, trong trường hợp trực giao, được cho là tương đương với giải pháp của Carroll.Thuật ngữ chung cho các giải pháp tương đương này là "Phương pháp Quartimax".
- Thủ tục Varimax của Kaiser (1958) là một sửa đổi của thủ tục Quartimax và có lẽ là phương pháp phân tích xoay nhân tố được sử dụng phổ biến nhất.
- Có thể lưu ý rằng Quartimax và Varimax là những trường hợp đặc biệt của lớp tiêu chuẩn trực giao cho phép xoay trực giao.
- Một số lượng lớn các phương pháp xoay nhân tố là kết quả của việc mở rộng một số thủ tục trực giao sang trường hợp tổng quát hơn (xiên). Các phương pháp xoay xiên đầu tiên được gọi là các phương pháp gián tiếp vì chúng liên quan đến sử dụng các cấu trúc tham chiếu.

## 4.4.2 Phương pháp xoay Quartimax

Cách tiếp cận của Ferguson (1954) được phát triển từ những cân nhắc về lý thuyết thông tin. Lý thuyết về "phương pháp xoay Quartimax"được trình bày một cách tổng quát:

**Định nghĩa 4.3.** (Phương pháp xoay Quartimax) Một phương pháp quay không đơn lẻ không làm thay đổi lượng phương sai được giải thích và tính cộng đồng của từng biến cũng không thay đổi. Nghĩa là:

$$\left(\sum_{j=1}^{q}\lambda_{ij}^2\right)^2 = \sum_{j=1}^{q}\lambda_{ij}^4 + \sum_{j=1}^{q}\sum_{j\neq k}^{q}\lambda_{ij}^2\lambda_{ik}^2 = const \tag{4.4}$$

chỉ số $j = 0$ đã được bỏ qua để biểu thức được rõ ràng hơn. Tổng hợp phương trình (4.4) trên tất cả các biến $p$ đã cho:

$$\sum_{i=1}^{p}\sum_{j} = 1^q\lambda_{ij}^4 + \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{j\neq k}^{q}\lambda_{ij}^2\lambda_{ik}^2 = const \tag{4.5}$$

Một ví dụ về phân tích nhân tố được đưa ra dưới đây chỉ ra rằng phương pháp Quartimax có xu hướng giữ lại một phần nhân tố đầu tiên quan trọng. Đây là đặc điểm của phương pháp luân chuyển nhân tố và xảy ra bởi vì quartimax về cơ bản cố gắng đơn giản hóa các hàng của ma trận mẫu thông qua việc giảm thiểu số hạng tích chéo trong phương trình (4.5).

## 4.4.3 Phương pháp xoay Varimax

Phương pháp Varimax cố gắng đơn giản hóa các cột thay vì các hàng của ma trận mẫu. Do đó, nó ngăn cản việc duy trì một yếu tố đầu tiên khá chung chung.
Kaiser (1958) định nghĩa tính đơn giản của một nhân tố là bình phương của phương sai của tải trọng của nó. Đối với yếu tố này người ta biểu diễn bằng biến $v_j^*$, trong đó

$$v_j^* = \frac{1}{p}\left[\sum_{i=1}^{p}\lambda_{ij}^4 - \frac{1}{p}\left(\sum_{i=1}^{p}\lambda_{ij}^2\right)^2\right], j = 1, \ldots, q. \tag{4.6}$$

Phương pháp Varimax liên quan đến việc tối đa hóa tổng số đơn giản tức là tối đa hóa

$$V^* = \sum_{j=1}^{q} v_j^* \tag{4.7}$$

Trong thực tế, các hệ số tải nhân tố thường được chuẩn hóa bởi các cộng đồng tương ứng của chúng, không thay đổi bởi một vòng quay không đơn lẻ. Do đó tiêu chí Varimax được đưa ra bởi:

$$V = \frac{1}{p}\sum_{j=1}^{q}\left[\sum_{i=1}^{p}\frac{\lambda_{ij}^4}{h_i^4} - \frac{1}{p}\left(\sum_{i=1}^{p}\frac{\lambda_{ij}^2}{h_i^2}\right)\right], \tag{4.8}$$

trong đó

$$h_i^2 = \sum_{j=1}^{q}\lambda_{ij}^2, i = 1, \ldots, p, \tag{4.9}$$

là tính cộng đồng của biến $x$. Như với phép xoay quartimax, quy trình tính toán liên quan đến phép xoay theo cặp của các thừa số.

Kiểm tra tỷ lệ phương sai tương đối được giải thích bởi từng yếu tố chỉ ra rằng Varimax có xu hướng làm các tải lớn trên các cột của ma trận mẫu ở mức độ lớn hơn so với quartimax. Điều này là do, để $V$ đạt cực đại, số hạng thứ hai trong phương trình (4.9) phải nhỏ, tức là:

$$\sum_{i=1}^{p}\frac{\lambda_{ij}^2}{h_i^2}$$

phải tương đối ổn định giữa các yếu tố.

**Ví dụ 4.3.** Tải luân phiên đối với dữ liệu giá cổ phiếu
Bảng trình bày các ước tính khả năng ban đầu và luân phiên tối đa của hệ số tải đối với dữ liệu giá cổ phiếu

| **Cổ phiếu** | **Hệ số tải ước tính $F_1$** | **Hệ số tải ước tính $F_2$** | **Hệ số tải ước tính đã xoay $F_1^*$** | **Hệ số tải ước tính đã xoay $F_2^*$** | **Phương sai cụ thể $\hat{\psi}_i^2 = 1 - \hat{h}_i^2$** |
|---|---|---|---|---|---|
| **FRT-Bán lẻ kĩ thuật số FPT** | 0.401 | 0.823 | 0.114 | 0.908 | 0.16 |
| **MSN-Tập đoàn Masan** | 0.358 | 0.799 | 0.080 | 0.872 | 0.23 |
| **SHB-NH TMCP Sài Gòn** | 0.523 | 0.654 | 0.706 | -0.450 | 0.30 |
| **VCB-Vietcombank** | 0.700 | -0.112 | 0.699 | 0.120 | 0.50 |
| **TCB-Techcombank** | 0.873 | 0.035 | 0.815 | 0.315 | 0.24 |
| **Giải thích tỷ lệ tích lũy** | 0.363 | 0.714 | 0.334 | 0.714 | |

Bảng hệ số tải nhân tố ước tính và đã xoay cùng phương sai cụ thể

**Nhận xét kết quả**

- Trước khi xoay
  - Trước khi xoay nhân tố F1 ảnh hưởng tới cả 5 cổ phiếu; có thể được tạm gọi là nhân tố chung mang tên là Tác động của lãi suất
- Sau khi xoay

  - Cổ phiếu ngành bán lẻ (FRT, MSN ) chịu ảnh hưởng cao ở yếu tố $F2^*$.
  - Cổ phiếu ngành ngân hàng (SHB, VCB và TCB) chịu ảnh hưởng cao ở yếu tố $F1^*$.
  - Tải trọng luân chuyển thu được từ giải pháp thành phần chính không được hiển thị.
  - Hai yếu tố luân chuyển cùng nhau phân biệt các ngành công nghiệp.
  - Yếu tố 1 đại diện cho những lực lượng kinh tế độc đáo khiến cổ phiếu bán lẻ di chuyển cùng nhau.
  - Yếu tố 2 dường như đại diện cho các điều kiện kinh tế ảnh hưởng đến cố phiếu ngân hàng.

- Một nhân tố chung (tức là nhân tố mà tất cả các biến đều tải cao) có xu hướng "bị triệt tiêu sau khi xoay".

- Một số chương trình phân tích nhân tố mục đích chung cho phép khắc phục các tải liên quan đến các yếu tố nhất định và quay vòng các yếu tố còn lại.

**Ví dụ 4.4.** Tải luân phiên cho dữ liệu mười môn phối hợp Olympic

- Hệ số tải ước tính và phương sai cụ thể cho dữ liệu mười môn phối hợp Olympic được trình bày trong Ví dụ Mười môn phối hợp.

- Các đại lượng này được lấy từ mô hình nhân tố $m = 4$, sử dụng cả phương pháp giải thành phần chính và phương pháp giải hợp lý tối đa.

- Phép quay Varimax thực hiện để xem liệu các tải nhân tố đã quay có cung cấp thêm thông tin chuyên sâu hay không.Tải trọng quay Varimax cho các giải pháp nhân tố $m = 4$ được hiển thị trong Bảng, cùng với các phương sai cụ thể.

Kết quả nghiên cứu phân tích dữ liệu các môn phối hợp Olympic cho tất cả 160 lần xuất phát. Trong đó, $n = 280$ kết quả từ 1960 đến 2004. Kết quả phân tích ma trận tương quan dựa trên 280 trường hợp:

$$R = \begin{pmatrix} 1.000 & 0.6386 & 0.4752 & 0.3227 & 0.5520 & 0.3262 & 0.3509 & 0.4008 & 0.1821 & -0.0352 \\ 0.6386 & 1.0000 & 0.4953 & 0.5668 & 0.4706 & 0.3520 & 0.3998 & 0.5167 & 0.3102 & 0.1012 \\ 0.4752 & 0.4953 & 1.0000 & 0.4357 & 0.2539 & 0.2812 & 0.7926 & 0.4728 & 0.4682 & -0.0120 \\ 0.3227 & 0.5668 & 0.4357 & 1.0000 & 0.3449 & 0.3503 & 0.3657 & 0.6040 & 0.2344 & 0.2380 \\ 0.5520 & 0.4706 & 0.2539 & 0.3449 & 1.0000 & 0.1546 & 0.2100 & 0.4213 & 0.2116 & 0.4125 \\ 0.3262 & 0.3520 & 0.2812 & 0.3503 & 0.1546 & 1.0000 & 0.2553 & 0.4163 & 0.1712 & 0.0002 \\ 0.3509 & 0.3998 & 0.7926 & 0.3657 & 0.2100 & 0.2553 & 1.0000 & 0.4036 & 0.4179 & 0.0109 \\ 0.4008 & 0.5167 & 0.4728 & 0.6040 & 0.4213 & 0.4163 & 0.4036 & 1.0000 & 0.3151 & 0.2395 \\ 0.1821 & 0.3102 & 0.4682 & 0.2344 & 0.2116 & 0.1712 & 0.4179 & 0.3151 & 1.0000 & 0.0983 \\ -0.0352 & 0.1012 & -0.0120 & 0.2380 & 0.4125 & 0.0002 & 0.0109 & 0.2395 & 0.0983 & 1.0000 \end{pmatrix}$$

Ước lượng hệ số tải và phương sai cụ thể cho phương pháp thành phần chính

| | Thành phần chính | | | | | Thành phần chính đã xoay | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Môn thể thao | $F_1$ | $F_2$ | $F_3$ | $F_4$ | $\tilde{\psi}_i = 1 - \tilde{h}_i^2$ | $F_1^*$ | $F_2^*$ | $F_3^*$ | $F_4^*$ | $\tilde{\psi}_i = 1 - \tilde{h}_i^2$ |
| Điền kinh | 0.696 | 0.022 | -0.468 | -0.416 | 0.12 | 0.182 | 0.885 | 0.205 | -0.139 | 0.12 |
| Bơi lội | 0.793 | 0.075 | -0.255 | -0.115 | 0.29 | 0.291 | 0.664 | 0.429 | 0.055 | 0.29 |
| Cầu lông | 0.771 | -0.434 | 0.197 | -0.112 | 0.17 | 0.819 | 0.302 | 0.252 | -0.097 | 0.17 |
| Bóng đá | 0.711 | 0.181 | 0.005 | 0.367 | 0.33 | 0.267 | 0.221 | 0.683 | 0.293 | 0.33 |
| Bóng chuyền | 0.605 | 0.549 | -0.045 | -0.397 | 0.17 | 0.086 | 0.747 | 0.068 | 0.507 | 0.17 |
| Tennis | 0.513 | -0.083 | -0.372 | 0.561 | 0.28 | 0.048 | 0.108 | 0.826 | -0.161 | 0.28 |
| Thể dục nghệ thuật | 0.690 | -0.456 | 0.289 | -0.078 | 0.23 | 0.832 | 0.185 | 0.204 | -0.076 | 0.23 |
| Karate | 0.761 | 0.162 | 0.018 | 0.304 | 0.30 | 0.324 | 0.278 | 0.656 | 0.293 | 0.30 |
| Taekwondo | 0.518 | -0.252 | 0.519 | -0.074 | 0.39 | 0.754 | 0.024 | 0.054 | 0.188 | 0.39 |
| Bóng rổ | 0.220 | 0.746 | 0.493 | 0.085 | 0.15 | -0.002 | 0.019 | 0.075 | 0.921 | 0.15 |
| Tỷ lệ tích lũy giải thích | 0.42 | 0.56 | 0.67 | 0.76 | | 0.22 | 0.43 | 0.62 | 0.76 | |

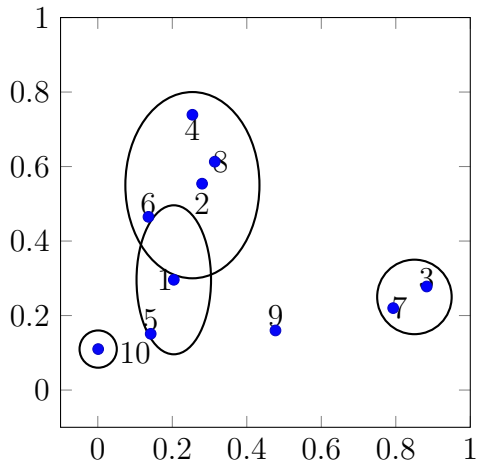
Ước lượng hệ số tải và phương sai cụ thể cho phương pháp ước lượng cực đại

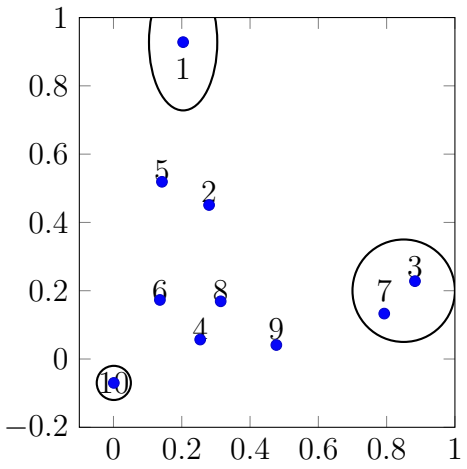
| | Ước lượng cực đại | | | | | Ước lượng cực đại đã xoay | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Môn thể thao | $F_1$ | $F_2$ | $F_3$ | $F_4$ | $\tilde{\psi}_i = 1 - \hat{h}_i^2$ | $F_1^*$ | $F_2^*$ | $F_3^*$ | $F_4^*$ | $\tilde{\psi}_i = 1 - \hat{h}_i^2$ |
| Điền kinh | 0.993 | -0.069 | -0.021 | 0.002 | 0.01 | 0.204 | 0.296 | 0.928 | -0.005 | 0.01 |
| Bơi lội | 0.665 | 0.252 | 0.239 | 0.220 | 0.39 | 0.280 | 0.554 | 0.451 | 0.155 | 0.39 |
| Cầu lông | 0.530 | 0.777 | -0.141 | -0.079 | 0.09 | 0.883 | 0.278 | 0.228 | -0.045 | 0.09 |
| Bóng đá | 0.363 | 0.428 | 0.421 | 0.424 | 0.33 | 0.254 | 0.739 | 0.057 | 0.242 | 0.33 |
| Bóng chuyền | 0.571 | 0.019 | 0.620 | -0.305 | 0.20 | 0.142 | 0.151 | 0.519 | 0.700 | 0.20 |
| Tennis | 0.343 | 0.189 | 0.090 | 0.323 | 0.73 | 0.136 | 0.465 | 0.173 | -0.033 | 0.73 |
| Thể dục nghệ thuật | 0.402 | 0.718 | -0.102 | -0.095 | 0.30 | 0.793 | 0.220 | 0.133 | -0.009 | 0.30 |
| Karate | 0.440 | 0.407 | 0.390 | 0.263 | 0.42 | 0.314 | 0.613 | 0.169 | 0.279 | 0.42 |
| Taekwondo | 0.218 | 0.461 | 0.084 | -0.085 | 0.73 | 0.477 | 0.160 | 0.041 | 0.139 | 0.73 |
| Bóng rổ | -0.016 | 0.091 | 0.609 | -0.145 | 0.60 | 0.001 | 0.110 | -0.070 | 0.619 | 0.60 |
| Tỷ lệ tích lũy giải thích | 0.27 | 0.45 | 0.57 | 0.62 | | 0.30 | 0.37 | 0.51 | 0.62 | |

**Nhận xét:**

- Nhân tố $F_1^*$: Dường như đại diện cho các môn thể thao có tính đòi hỏi kỹ năng Kỹ thuật và khéo léo: Cầu lông, thể dục nghệ thuật , Taekwondo
- Nhân tố $F_2^*$: Dường như đại diện Sức bền và thể lực : Điền kinh,Bơi lội, Bóng chuyền
- Nhân tố $F_3^*$ Dường như đại diện cho Sức mạnh, phản xạ và chiến thuật: Bóng đá, Tennis,Karate
- Nhân tố $F_4^*$ Dường như đại diện cho Phối hợp nhóm và chiến thuật: Bóng rổ và bóng chuyền
- Ngoài các tải ước tính, phép quay sẽ chỉ ảnh hưởng đến phân phối tỷ lệ của tổng phương sai mẫu được giải thích bởi từng yếu tố.
- Tỷ lệ tích lũy của tổng phương sai mẫu giải thích cho tất cả các nhân tố không thay đổi.
- Các hệ số tải nhân tố xoay vòng cho cả hai phương pháp giải pháp đều chỉ ra các thuộc tính cơ bản giống nhau, mặc dù các yếu tố 1 và 2 không theo cùng một thứ tự.
- Biểu đồ tải trọng khả năng tối đa xoay vòng cho các cặp nhân tố (1,2) và (1,3) bằng Phương pháp MLE

  Các điểm thường được nhóm dọc theo các trục nhân tố.



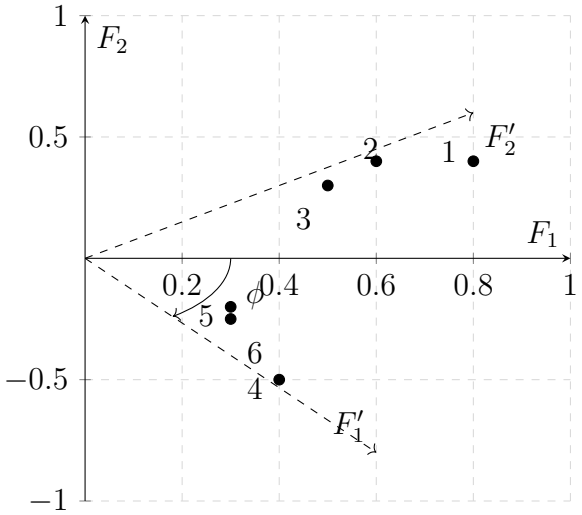
## 4.5 Phương pháp xoay xiên

**Định nghĩa 4.4.** "Oblimin trực tiếp" Là thủ tục xoay xiên được giới thiệu bởi Jennrich và Sampson (1966) không liên quan đến việc sử dụng các trục tham chiếu. Tiêu chí oblimin trực tiếp có dạng như sau:

$$\sum\sum_{j<k}^{q}\left[\sum_{i=1}^{p}\lambda_{ij}^2\lambda_{ik}^2-\frac{\zeta}{p}\sum_{i=1}^{p}\lambda_{ij}^2\sum_{i=1}^{p}\lambda_{ik}^2\right], \tag{4.10}$$

trong đó các chỉ số 0 đã được bỏ qua cho rõ ràng. $\zeta$ là một tham số kiểm soát mức độ tương quan giữa các yếu tố.

- Các phép xoay trực giao phù hợp với mô hình nhân tố trong đó các nhân tố chung được giả định là độc lập.
- Nhiều nhà nghiên cứu trong khoa học xã hội xem xét phép xoay xiên (không trực giao).
- Nếu coi $m$ thừa số chung là các trục tọa độ thì điểm có $m$ tọa độ $(\hat{l}_{i1}, \hat{l}_{i2}, \ldots, \hat{l}_{im})$ biểu thị vị trí của biến thứ $i$ trong không gian nhân tử.
- Giả sử rằng các biến được nhóm thành các cụm không chồng chéo, một phép xoay trực giao đến một cấu trúc đơn giản tương ứng với một phép xoay cứng nhắc của các trục tọa độ sao cho các trục, sau khi quay, đi càng gần các cụm càng tốt.
- Một phép xoay xiên đối với một cấu trúc đơn giản tương ứng với một phép quay không cứng nhắc của hệ tọa độ sao cho các trục quay (không còn vuông góc) đi qua (gần như) qua các cụm.
- Một phép xoay xiên tìm cách thể hiện từng biến theo thuật ngữ của một số yếu tố tối thiểu tốt nhất là một yếu tố duy nhất.




$\Longrightarrow$ Kết luận: Trong thực tế người ta sử dụng phương pháp quay xiên phổ biến hơn vì nó sẽ cho kết quả một cách trực quan và dễ hiểu hơn và các nhân tố chung tương quan với nhau.

# Chương 5

# Điểm nhân tố

## 5.1 Điểm nhân tố là gì?

Trong phân tích nhân tố, ta thường quan tâm vào các tham số trong mô hình nhân tố. Tuy nhiên, các giá trị ước lượng của các nhân tố chung, gọi là điểm nhân tố, cũng có thể được yêu cầu. Những đại lượng này thường được sử dụng cho mục đích chuẩn đoán, cũng như đầu vào cho phân tích tiếp theo.

Điểm nhân tố không phải là ước lượng của các tham số chưa biết theo nghĩa thông thường. Thay vào đó, chúng là ước lượng của các giá trị cho các vector nhân tố ngẫu nhiên không quan sát được. Tức là điểm nhân tố

$$\hat{f_j} = \text{ ước lượng của giá trị } f_j \text{ đạt được bởi } F_j(\text{thành phần thứ } j).$$

Việc ước lượng này rất phức tạp bởi vì số lượng các đại lượng không quan sát được $f_j$ và $\epsilon_j$ nhiều hơn số đại lượng quan sát được $x_i$. Để vượt qua trở ngại này, ta có một số cách giải quyết. Chương này sẽ đề cập đến 2 cách tiếp cận chủ yếu là phương pháp bình phương tối thiểu có trọng số và phương pháp hồi quy.

Hai cách tiếp cận đều có 2 yếu tố chung:

- Ta coi ước lượng của các hệ số tải $\hat{l}_{ij}$ và phương sai cụ thể $\hat{\Psi}_j$ là các giá trị thực.
- Chúng liên quan đến các phép biến đổi tuyến tính của dữ liệu gốc, có thể đã được chuẩn hóa.

## 5.2 Phương pháp bình phương tối thiểu có trọng số

Trước hết, ta giả sử rằng vector trung bình $\mu$ , ma trận tải $L$ và vector phương sai cụ thể $\Psi$ đã biết với mô hình nhân tố:

$$\underset{(p\times 1)}{X} - \underset{(p\times 1)}{\mu} = \underset{(p\times m)}{L}\underset{(m\times 1)}{F} + \underset{(p\times 1)}{\epsilon}$$

Hơn nữa, ta coi các sai số $\epsilon^\top = [\epsilon_1, \epsilon_2, \ldots, \epsilon_p]$ là các nhân tố xác định. Vì $Var(\epsilon_i) = \Psi_i, i = 1, 2, \ldots p$, không cần phải bằng nhau, Bartlett đã gợi ý rằng phương pháp bình phương tối thiểu có trọng số có thể được sử dụng để ước lượng giá trị các nhân tố chung.

Tổng bình phương của các nhân tố xác định, có trọng số bằng nghịch đảo các phương sai của chúng là:

$$\sum_{i=1}^{p} \frac{\epsilon_i^2}{\Psi_i} = \epsilon^\top \Psi^{-1} \epsilon = (x - \mu - Lf)^\top \Psi^{-1} (x - \mu - Lf) \tag{5.1}$$

Bartlett đề xuất chọn các ước lượng của $f$ để cực tiểu hóa (5.1). Ta thu được nghiệm là

$$\hat{f} = (L^\top \Psi^{-1} L)^{-1} L^\top \Psi^{-1} (x - \mu) \tag{5.2}$$

Từ công thức (5.2), ta lấy các ước lượng $\hat{L}, \hat{\Psi},$ và $\hat{\mu} = \bar{x}$ là các giá trị thực và thu được điểm nhân tố cho thành phần thứ $j$ là

$$\hat{f}_j = (\hat{L}^\top \hat{\Psi}^{-1} \hat{L})^{-1} \hat{L}^\top \hat{\Psi}^{-1} (x_j - \bar{x}) \tag{5.3}$$

Khi $\hat{L}$ và $\hat{\Psi}$ được xác định bằng phương pháp ước lượng hợp lí cực đại thì nó phải thỏa mãn điều kiện duy nhất, $\hat{L}^{-1\top} \hat{\Psi}^{-1} \hat{L} = \hat{\Delta}$ là một ma trận đường chéo. Từ đó ta có kết quả sau đây:

**Điểm nhân tố thu được bằng phương pháp bình phương tối thiểu có trọng số từ ước lượng hợp lí cực đại**

$$\begin{aligned} \hat{f}_j &= (\hat{L}^\top \hat{\Psi}^{-1} \hat{L})^{-1} \hat{L}^\top \hat{\Psi}^{-1} (x_j - \hat{\mu}) \\ &= \hat{\Delta}^{-1} \hat{L}^\top \hat{\Psi}^{-1} (x_j - \bar{x}), \quad j = 1, 2, \ldots, n \end{aligned} \tag{5.4}$$

hoặc

$$\begin{aligned} \hat{f}_j &= (\hat{L}_z^\top \hat{\Psi}_z^{-1} \hat{L}_z)^{-1} \hat{L}_z^\top \hat{\Psi}_z^{-1} z_j \\ &= \hat{\Delta}_z^{-1} \hat{L}_z^\top \hat{\Psi}_z^{-1} z_j, \quad j = 1, 2, \ldots, n \end{aligned}$$

Trong đó $z_j = D^{-1/2}$ và $\hat{\rho} = \hat{L}_z \hat{L}_z^\top + \hat{\Psi}_z$.
Điểm nhân tố được xác định trong công thức (5.4) có vector trung bình và ma trận hiệp phương sai bằng 0.

Nếu ma trận tải sau khi quay $\hat{L}^* = \hat{L}T$ được sử dụng để thay thế ma trận tải ban đầu thì các điểm nhân tố mới $\hat{f}_j^* = T^\top \hat{f}_j, \quad j = 1, 2, \ldots, n$.

Nếu ma trận tải được ước lượng bằng phương pháp thành phần chính, thông thường ta sẽ tạo điểm nhân tố bằng cách sử dụng bình phương tối thiểu không có trọng số. Một cách ngầm định, điều này có nghĩa rằng chúng giống nhau hoặc gần giống nhau. Khi đó các điểm nhân tố:

$$\hat{f}_j = (\tilde{L}^\top \tilde{L})^{-1} \tilde{L}^\top (x_j - \bar{x})$$

hoặc

$$\hat{f}_j = (\tilde{L}^\top \tilde{L})^{-1} \tilde{L}^\top z_j$$

với dữ liệu chuẩn hóa.
Vì $\tilde{L} = \left[\sqrt{\lambda_1} e_1 \;\middle|\; \sqrt{\lambda_2} e_2 \;\middle|\; \ldots \;\middle|\; \sqrt{\lambda_m} e_m\right]$ nên ra ta có:

$$\hat{f}_j = \begin{bmatrix} \frac{1}{\sqrt{\hat{\lambda}_1}} \hat{e}_1^\top (x_j - \bar{x}) \\ \frac{1}{\sqrt{\hat{\lambda}_2}} \hat{e}_2^\top (x_j - \bar{x}) \\ \vdots \\ \frac{1}{\sqrt{\hat{\lambda}_m}} \hat{e}_m^\top (x_j - \bar{x}) \end{bmatrix} \tag{5.5}$$

Với các điểm nhân tố này:

$$\frac{1}{n}\sum_{j=1}^{n}\hat{f}_j = 0 \quad \text{(trung bình mẫu)}$$

$$\frac{1}{n-1}\sum_{j=1}^{n}\hat{f}_j\hat{f}_j^\top = I \quad \text{(phương sai mẫu hiệu chỉnh)}$$

## 5.3 Phương pháp hồi quy

Ta xuất phát từ mô hình nhân tố ban đầu

$$X - \mu = LF + \epsilon$$

ta coi như đã biết ma trận tải $L$ và ma trận phương sai $\Psi$. Khi các nhân tố chung $F$ và các nhân tố xác định $\epsilon$ có cùng phân phối chuẩn với trung bình và hiệp phương sai được cho bởi mô hình thì $X - \mu = LF + \epsilon$ có phân phối chuẩn $N_p(0, LL^\top + \Psi)$.
Hơn nữa, phân phối chung của $(X - \mu)$ và $F$ là $N_{m+p}(0, \Sigma^*)$, trong đó:

$$\underset{(m+p)\times(m+p)}{\Sigma^*} = \begin{bmatrix} \underset{(p\times p)}{\Sigma = LL^\top + \Psi} & \underset{(p\times m)}{L} \\ \underset{(m\times p)}{L^\top} & \underset{(m\times m)}{I} \end{bmatrix} \tag{5.6}$$

và 0 là vector không cỡ $(m + p) \times 1$.

**Nhắc lại kết quả ở chương 4**:
Cho $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ có phân phối chuẩn $N_p(\mu, \epsilon)$ với $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, và $|\Sigma_{22}| > 0$. Khi đó, phân phối có điều kiện của $X_1$ với điều kiện $X_2 = x_2$ là phân phối chuẩn và có:

$$\text{Trung bình } = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

và

$$\text{Hiệp phương sai } = \Sigma 11 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$

hơn nữa, hiệp phương sai này không phụ thuộc vào giá trị $x_2$ của biến điều kiện.

Sử dụng kết quả này, ta tìm được phân phối có điều kiện của $F|x$ là phân phối chuẩn đa biến với:

$$E(F|x) = L^\top\Sigma^{-1}(x - \mu) = L^\top(LL^\top + \Psi)^{-1}(x - \mu) \tag{5.7}$$

$$Cov(F|x) = I - L^\top\Sigma^{-1}L = I - L^\top(LL^\top + \Psi)^{-1}L \tag{5.8}$$

Các đại lượng $L^\top(LL^\top + \Psi)^{-1}$ là các hệ số hồi quy của các nhân tố trên các biến. Ước lượng của các hệ số này tạo ra các điểm nhân tố tương tự như ước lượng của các giá trị trung bình trong phân tích hồi quy đa biến.
Do đó, với bất kì vector quan sát được $x_j$ và lấy các ước lượng hợp lí cực đại của $\hat{L}$ và $\hat{\Psi}$ là các giá trị thực, ta thấy vector điểm nhân tố thứ $j$ được xác định bởi

$$\hat{f}_j = \hat{L}^\top\hat{\Sigma}^{-1}(x_j - \bar{x}) = \hat{L}^\top(\hat{L}\hat{L}^\top + \hat{\Psi})^{-1}(x_j - \bar{x}), \quad j = 1, 2, \ldots, n \tag{5.9}$$

Kết quả của $\hat{f}_j$ trong (5.9) có thể được rút gọn bằng cách sử dụng ma trận đơn vị:

$$\underset{(m\times p)}{\hat{L}^\top}\underset{(p\times p)}{(\hat{L}\hat{L}^\top + \hat{\Psi})^{-1}} = \underset{(m\times m)}{(I + \hat{L}^\top\hat{\Psi}^{-1}\hat{L})^{-1}}\underset{(m\times p)}{\hat{L}^\top}\underset{(p\times p)}{\hat{\Psi}^{-1}} \tag{5.10}$$

Ta có thể so sánh các điểm nhân tố trong công thức (5.9), được tạo bởi phương pháp hồi quy với các điểm nhân tố được tạo bởi phương pháp bình phương tối thiểu ở công thức (5.4). Ta kí hiệu điểm nhân tố được tạo bởi hai phương pháp lần lượt là $\hat{f}_j^R$ và $\hat{f}_j^{LS}$. Từ công thức (5.10), ta có:

$$\hat{f}_j^{LS} = (\hat{L}^\top \hat{\Psi}^{-1} \hat{L})^{-1}(I + \hat{L}^\top \hat{\Psi}^{-1} \hat{L})\hat{f}_j^R = (I + (\hat{L}^\top \hat{\Psi}^{-1} \hat{L})^{-1})\hat{f}_j^R \tag{5.11}$$

Với các ước lượng hợp lí cực đại $(\hat{L}^\top \hat{\Psi}^{-1} \hat{L})^{-1} = \hat{\Delta}^{-1}$ và nếu các phần tử của ma trận đường chéo này gần bằng 0 thì hai phương pháp sẽ cho ra các điểm nhân tố gần bằng nhau.
Để giảm thiểu ảnh hưởng của một quyết định sai về số lượng các nhân tố, ta thường có xu hướng sử dụng ma trận $S$ (ma trận hiệp phương sai mẫu ban đầu) thay cho $\hat{\Sigma} = \hat{L}\hat{L}^\top + \hat{\Psi}$. Do đó ta có kết quả sau:

**Điểm nhân tố thu được bằng phương pháp hồi quy**

$$\hat{f}_j = \hat{L}^\top S^{-1}(x_j - \bar{x}), \quad j = 1, 2, \ldots, n \tag{5.12}$$

hoặc

$$\hat{f}_j = \hat{L}_z^\top R^{-1} z_j$$

trong đó, $z_j = D^{-1/2}(x_j - \bar{x})$ và $\hat{\rho} = \hat{L}_z \hat{L}_z^\top + \hat{\Psi}_z$
Nếu ma trận tải sau khi quay $\hat{L}^* = \hat{L}T$ được sử dụng để thay thế ma trận tải ban đầu thì điểm nhân tố mới $\hat{f}_j = T^\top \hat{f}_j, \quad j = 1, 2, \ldots, n$.

**Ví dụ 5.1. (Tính toán điểm nhân tố)** Quay trở lại với ví dụ 4.3 về giá cổ phiếu trong sách , ta đã có ước lượng của ma trận tải sau khi quay và phương sai cụ thể bằng phương pháp ước lượng hợp lí cực đại:

$$\hat{L}_z^* = \begin{bmatrix} .763 & .024 \\ .821 & .227 \\ .669 & .104 \\ .118 & .993 \\ .113 & .675 \end{bmatrix} \text{ và } \hat{\Psi}_z = \begin{bmatrix} .42 & 0 & 0 & 0 & 0 \\ 0 & .27 & 0 & 0 & 0 \\ 0 & 0 & .54 & 0 & 0 \\ 0 & 0 & 0 & .00 & 0 \\ 0 & 0 & 0 & 0 & .53 \end{bmatrix}$$

Vector của các quan sát đã chuẩn hóa:

$$z^\top = \begin{bmatrix} .50 & -1.40 & -.20 & -.70 & 1.40 \end{bmatrix}$$

Ta có điểm nhân tố cho các nhân tố 1 và 2:
Phương pháp bình phương tối thiểu có trọng số (5.4):

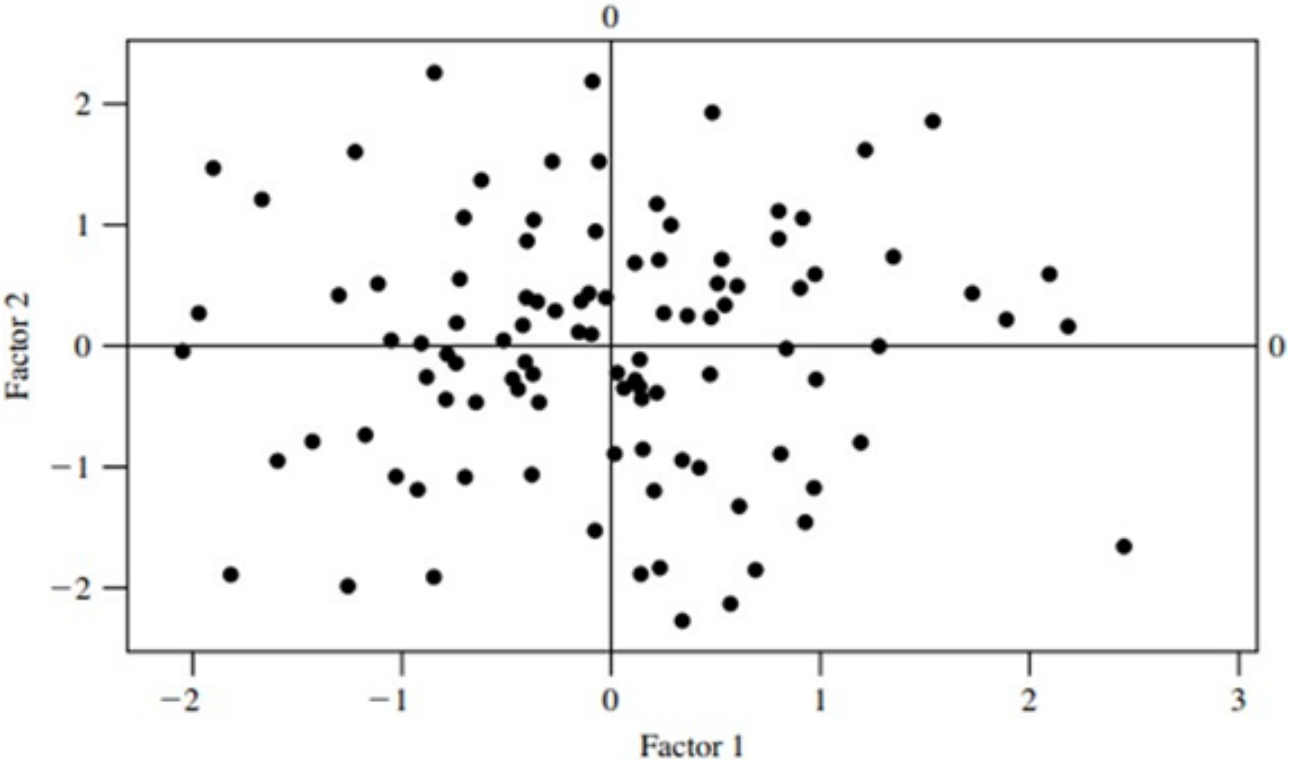
$$\hat{f} = (\hat{L}_z^{*\top} \hat{\Psi}_z^{-1} \hat{L}_z^*)^{-1} \hat{L}_z^{*\top} \hat{\Psi}_z^{-1} z = \begin{bmatrix} -.61 \\ -.61 \end{bmatrix}$$

Phương pháp hồi quy (5.12):

$$\hat{f} = \hat{L}_z^{*\top} R^{-1} z = \begin{bmatrix} .331 & .526 & .221 & -.137 & .011 \\ -.040 & -.063 & -.026 & 1.023 & -.001 \end{bmatrix} \begin{bmatrix} .50 \\ -1.40 \\ -.20 \\ -.70 \\ 1.40 \end{bmatrix}$$

$$= \begin{bmatrix} -.50 \\ -.64 \end{bmatrix}$$

Trong ví dụ này, cả 2 phương pháp đều cho kết quả khá giống nhau. Tất cả các điểm nhân tố thu được bằng phương pháp hồi quy được biểu diễn ở hình bên dưới:




Điểm nhân tố sử dụng cho nhân tố 1 và 2

Điểm nhân tố ứng với các thuộc tính trực quan có thể được xây dựng rất đơn giản. Ta nhóm các biến có hệ số tải cao (có giá trị tuyệt đối lớn hơn 0.40) trên một nhân tố. Điểm nhân tố cho nhân tố 1 được xây dựng bằng cách tính tổng các giá trị quan sát được (đã chuẩn hóa) của các biến trong nhóm. Điểm nhân tố cho nhân tố 2 là tổng của các quan sát chuẩn hóa ứng với các biến có hệ số tải cao trên nhân tố 2,... Việc giảm lượng dữ liệu được thực hiện bằng cách thay thế dữ liệu đã chuẩn hóa bằng các điểm nhân tố rút gọn này. Điểm nhân tố rút gọn thường có mối tương quan cao với các điểm nhân tố thu được bằng phương pháp hồi quy và bình phương tối thiểu.

**Ví dụ 5.2. (Xây dựng các điểm nhân tố rút gọn từ việc nhóm các nhân tố phân tích)** Phương pháp phân tích thành phần chính của dữ liệu giá cổ phiếu ở ví dụ 3.2 đã cho ta các ước lượng:

$$\tilde{L} = \begin{bmatrix} .732 & -.437 \\ .831 & -.280 \\ .726 & -.374 \\ .605 & .694 \\ .563 & .719 \end{bmatrix} \text{ và } \tilde{L}^* = \tilde{L}T = \begin{bmatrix} .852 & .030 \\ .851 & .214 \\ .813 & .079 \\ .133 & .911 \\ .084 & .909 \end{bmatrix}$$

Với mỗi nhân tố, lấy các tải trọng có giá trị tuyệt đối lớn nhất và bỏ qua các tải trọng nhỏ hơn, ta có các tổ hợp tuyến tính:

$$\hat{f}_1 = x_1 + x_2 + x_3 + x_4 + x_5$$
$$\hat{f}_2 = x_4 + x_5 - x_1$$

Trong thực tế, ta sẽ chuẩn hóa các biến mới này.
Nếu thay vì $\tilde{L}$, ta bắt đầu với ma trận $\tilde{L}^*$, ta sẽ thu được các điểm nhân tố rút gọn:

$$\hat{f}_1 = x_1 + x_2 + x_3$$
$$\hat{f}_2 = x_4 + x_5$$

Việc xác định các tải cao và tải không đáng kể thực sự rất chủ quan. Các tổ hợp tuyến tính làm cho các chủ đề có ý nghĩa được ưu tiên hơn.

# Chương 6

# Ví dụ thực tiễn

Chúng em đưa ra tình huống thực tế như sau
Giả sử ta làm việc ở bộ phận HR (Human Resources) và ta cần tìm ra lý do ẩn nhân viên nghỉ việc. Có 16 yếu tố được đánh giá:

| Tên yếu tố | Tên yếu tố (tiếng anh) |
|---|---|
| Tuổi | Age |
| Giới tính | Gender |
| Trình độ học vấn | Edu |
| Cấp bậc | Level(lvl) |
| Vị trí công việc | Position |
| Đơn vị (phòng) | Depart |
| Khoảng cách từ nhà tới chỗ làm | Distance |
| Thu nhập | Income |
| OT(làm thêm giờ) | OT |
| Sự thỏa mãn trong công việc | Satisfaction(satis) |
| KPI | KPI |
| Mức độ tham gia công việc | ParticipantLevel(partlvl) |
| Cân bằng cuộc sống | balance |
| Tần suất đi du lịch | travelfreq |
| Tình trạng kết hôn | status |
| Số năm trong công ty | years |

Tên yếu tố

Ta muốn nghiên cứu các nhân tố ẩn đóng góp trong việc đánh giá này.
Sau đây sẽ là các bước cần làm để đạt được mục tiêu trên.

**Bài làm:**
**Bước 1. Tìm số lượng nhân tố ẩn**
**a) Tiền xử lý dữ liệu**
Để có thể xử lý được dữ liệu, ta cần đảm bảo dữ liệu đã "sạch". Tiến hành loại bỏ đi các dữ liệu bị trống, dư thừa, không đúng định dạng, cú pháp và vô lý.

- Bị trống

- Dư thừa (Các quan sát trùng lặp)
- Không đúng định dạng, cú pháp
- Vô lý

| Tuổi | Giới Tính | Trình độ học vấn | Cấp bậc | Vị trí công việc | Đơn vị | ... | Tần suất đi du lịch | Tình trạng kết hôn | Số năm trong công ty |
|---|---|---|---|---|---|---|---|---|---|
| 41 | Female | 2 | 2 | Sales Executive | Sales | ... | Travel rarely | Single | 6 |
| 49 | Male | 1 | 2 | Research Scientist | Research Development | ... | Travel Frequently | Married | 10 |
| 37 | Male | 2 | 1 | Laboratory Technician | Research Development | ... | Travel Rarely | Single | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 27 | Male | 3 | 2 | Manufacturing Director | Research Development | ... | Travel Rarely | Married | 6 |
| 49 | Male | 3 | 2 | Sales Executive | Sales | ... | Travel Frequently | Married | 9 |
| 34 | Male | 3 | 2 | Laboratory Technician | Research Development | ... | Travel Rarely | Married | 4 |

Bảng dữ liệu tiền xử lý

**b) Chuẩn hóa dữ liệu**

Dữ liệu được thu thập với nhiều biến khác nhau, bao gồm cả biến số và biến phân loại. Để thực hiện phân tích nhân tố, việc chuẩn hóa ,chuyển đổi các biến phân loại dạng chuỗi sang dạng số là rất cần thiết.

Dữ Liệu Thu Thập Bảng khảo sát của chúng tôi bao gồm các biến sau:

- Biến Số: Tuổi, Thu nhập, Khoảng cách tới công ty, Số năm trong công ty, ...
- Biến Phân Loại (Chuỗi): Giới tính, Tình trạng kết hôn, Đơn vị (Phòng), Đi du lịch (xả hơi), ...

- Cần chuẩn hóa dữ liệu là để đảm bảo chúng có cùng một thang đo, vì có thể xảy ra trường hợp các yếu tố được đo bằng các đơn vị khác nhau (ví dụ số giờ làm việc là giờ, thu nhập là triệu đồng,...) khiến cho quá trình phân tích không được chính xác.
- Với các biến số , chuẩn hóa bằng phương pháp chuẩn tắc hay còn gọi là Z-score :

$$Z\_score = \frac{X-\mu}{\sigma}$$

- Áp dụng phương pháp Label Encoding :là một kỹ thuật trong tiền xử lý dữ liệu, được sử dụng để chuyển đổi các biến phân loại (categorical variables) thành các giá trị số nguyên (integer values). Mỗi nhãn (label) duy nhất trong cột sẽ được ánh xạ thành một số nguyên duy nhất, giúp dữ liệu có thể được xử lý bởi các thuật toán học máy không làm việc trực tiếp với dữ liệu dạng văn bản.

  Để chuyển đổi các biến phân loại sang dạng số và vẫn giữ nguyên tên cột.

Dữ liệu sau khi được chuyển đổi sẽ có dạng:

| Tuổi | Giới Tính | Trình độ học vấn | Cấp bậc | Vị trí công việc | Đơn vị | ... | Tần suất đi du lịch | Tình trạng kết hôn | Số năm trong công ty |
|---|---|---|---|---|---|---|---|---|---|
| 0.44635 | -1.224745 | -0.891688 | -0.057788 | 1.032716 | 1.401512 | ... | 0.590048 | 1.236820 | -0.164613 |
| -0.913194 | 0.816497 | -1.868426 | -0.057788 | 0.626374 | -0.493817 | ... | -0.913194 | -0.133282 | 0.488508 |
| 0.590048 | 0.816497 | -0.891688 | -0.961486 | -0.998992 | -0.493817 | ... | 0.590048 | 1.236820 | -1. |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 0.590048 | 0.816497 | 0.085049 | -0.057788 | -0.186309 | -0.493817 | ... | 0.590048 | -0.133282 | -0.164613 |
| -0.913194 | 0.816497 | 0.085049 | -0.057788 | 1.032716 | 1.401512 | ... | -0.913194 | -0.133282 | 0. |
| 0.590048 | 0.816497 | 0.085049 | -0.057788 | -0.998992 | -0.493817 | ... | 0.590048 | -0.133282 | -0.491174 |

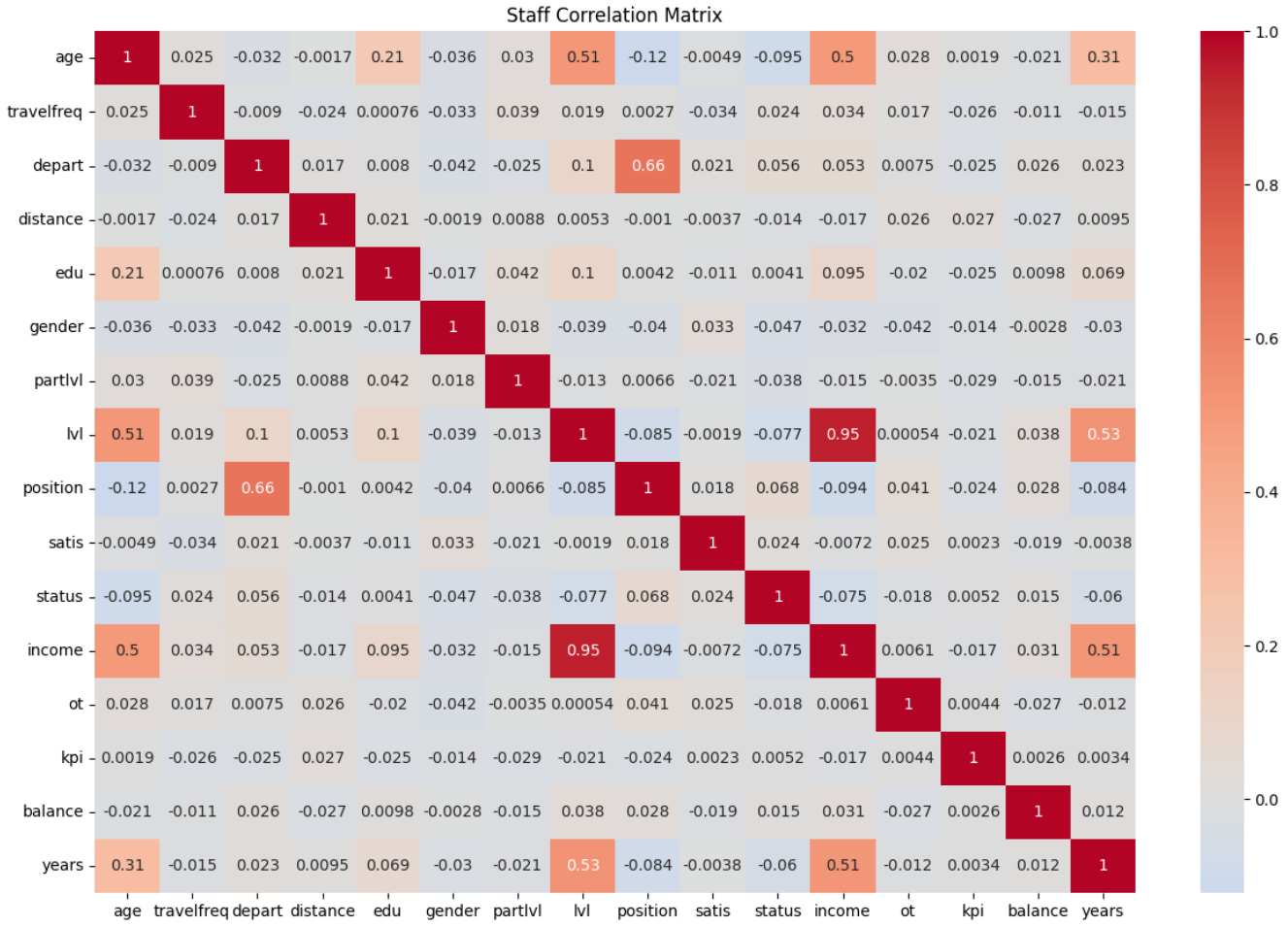
Bảng dữ liệu sau khi đã chuẩn hóa

**c) Tính ma trận hệ số tương quan A**

Ma trận hệ số tương quan của $a$ yếu tố là một ma trận cỡ $a \times a$, với $a_{ij}$ là hệ số tương quan của yếu tố $a_i$ và yếu tố $a_j$, được tính theo công thức:

$$a_{ij} = \frac{cov(a_i, a_j)}{\sigma_i \times \sigma_j}$$

với $cov(a_i, a_j)$ là hiệp phương sai của yếu tố $a_i$ và $a_j$, $\sigma_i$ là độ lệch chuẩn của yếu tố $a_i$.
Ta có kết quả ma trận hệ số tương quan:




Ma trận hệ số tương quan A

**d) Dựa vào các giá trị riêng của ma trận A, tìm số lượng nhân tố**

Từ ma trận hệ số tương quan A, ta có thể tính được các giá trị riêng của nó. Như ở phần 3.1, sách giáo khoa có giới thiệu cách tìm mức độ thể hiện của giá trị riêng theo công thức

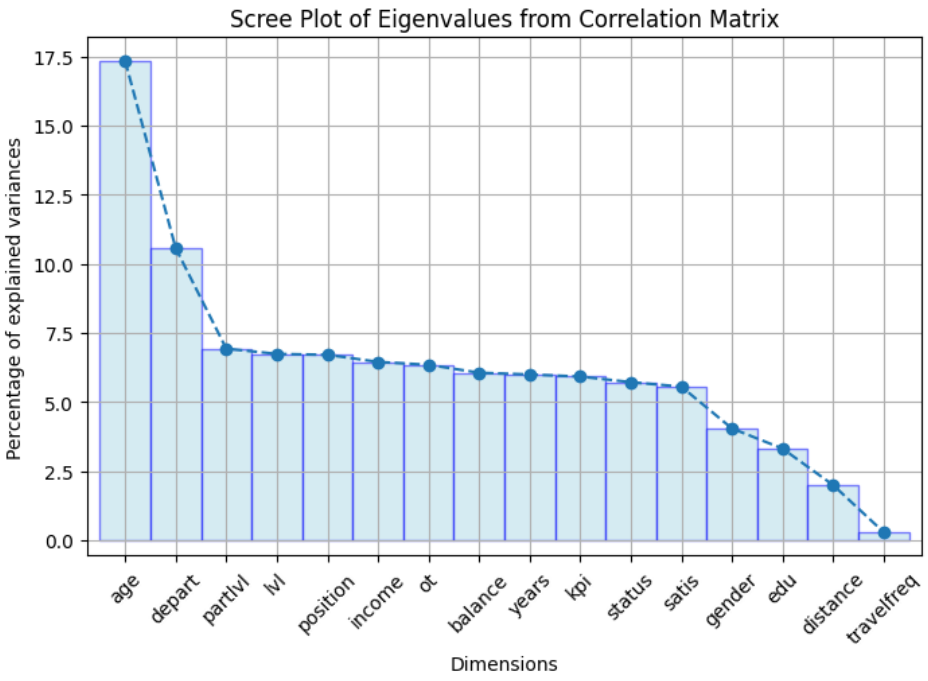
$$\text{Mức độ thể hiện} = \frac{\sum_{i=1}^{n} \lambda_i}{p}$$

với $p$ là số trị riêng của ma trận A, $n < p$ là số nhân tố được sử dụng.

Áp dụng vào dữ liệu, ta tìm được các giá trị riêng là:

$$\lambda_i = [2.77492837, 1.6902668, \ldots, 0.53047563, 0.31958415, 0.04687273]$$

Chọn $n = 10$, được mức độ thể hiện là 0.8479, hay với số lượng nhân tố phân tích là 10 thì có thể diễn giải được tới 84.79% toàn bộ dữ liệu.
Tương tự như phương pháp đã được sách trình bày,biểu đồ Scree cũng cho biết mức độ thể hiện của giá trị riêng
Scree plot của ví dụ này là:




Scree plot

**Bước 2. Xác định các hệ số tải của nhân tố**
**a) Chọn lựa phương pháp**

Trong ví dụ này, chúng em lựa chọn sử dụng phương pháp nhân tố chính PCA. Các công thức được sử dụng đã được nêu trong mục 3.1.

| | Factor 1 | Factor 2 | Factor 3 | ⋯ | Factor 8 | Factor 9 | Factor 10 |
|---|---|---|---|---|---|---|---|
| age | −0.692551 | 0.060418 | 0.151640 | ⋯ | −0.057257 | 0.049676 | 0.107920 |
| travelfreq | −0.031206 | −0.004405 | 0.467223 | ⋯ | 0.248630 | 0.286835 | −0.364654 |
| depart | −0.011319 | −0.906409 | −0.009013 | ⋯ | 0.069352 | 0.020776 | 0.037888 |
| distance | 0.000139 | −0.014463 | −0.000042 | ⋯ | 0.175172 | −0.242326 | −0.624331 |
| edu | −0.210927 | −0.029569 | 0.424581 | ⋯ | −0.207189 | −0.017774 | 0.198712 |
| gender | 0.058445 | 0.126707 | −0.116289 | ⋯ | 0.142376 | 0.172768 | −0.237699 |
| partlvl | 0.003548 | 0.041807 | 0.614101 | ⋯ | −0.049617 | 0.409723 | −0.073221 |
| lvl | −0.930621 | −0.107602 | −0.072232 | ⋯ | 0.044012 | 0.003207 | −0.062375 |
| position | 0.172692 | −0.886465 | 0.054087 | ⋯ | 0.013956 | 0.047415 | 0.059518 |
| satis | 0.014929 | −0.048959 | −0.326943 | ⋯ | −0.147886 | 0.287036 | −0.175056 |
| status | 0.143687 | −0.158880 | −0.093719 | ⋯ | 0.200401 | 0.014138 | −0.256965 |
| income | −0.922149 | −0.074958 | −0.068349 | ⋯ | 0.045191 | 0.013125 | −0.054940 |
| ot | −0.005914 | −0.061568 | 0.048201 | ⋯ | −0.550116 | −0.001356 | −0.039130 |
| kpi | 0.018990 | 0.062899 | −0.356440 | ⋯ | 0.138192 | 0.714197 | 0.185867 |
| balance | −0.029506 | −0.082478 | −0.141675 | ⋯ | −0.639852 | 0.147548 | −0.407207 |
| years | −0.691309 | −0.030452 | −0.137856 | ⋯ | 0.078263 | −0.028480 | −0.035749 |

Ma trận hệ số tải L

**b) Xoay nhân tố**
Trong bài ví dụ này, chúng em đã sử dụng xoay nhân tố với ma trận xoay như sau:

| | Factor 1 | Factor 2 | Factor 3 | ⋯ | Factor 8 | Factor 9 | Factor 10 |
|---|---|---|---|---|---|---|---|
| age | 0.62484 | −0.14329 | −0.11943 | ⋯ | −0.102601 | −0.02472 | −0.04993 |
| travelfreq | 0.033921 | −0.04435 | 0.749151 | ⋯ | −0.04832 | −0.08079 | −0.04729 |
| depart | 0.079778 | 0.889862 | −0.078804 | ⋯ | −0.01395 | −0.05275 | 0.003766 |
| distance | −0.01241 | −0.01915 | −0.07709 | ⋯ | −0.0399 | −0.02125 | 0.9758 |
| edu | 0.101248 | −0.01124 | −0.16172 | ⋯ | 0.004415 | −0.0791 | 0.020665 |
| gender | −0.19202 | −0.19071 | −0.30345 | ⋯ | −0.07142 | −0.19988 | −0.04504 |
| partlvl | −0.018642 | −0.09945 | 0.14068 | ⋯ | −0.0913 | −0.0634 | −0.09872 |
| lvl | 0.918371 | −0.00995 | −0.07051 | ⋯ | −0.00186 | −0.0746 | −0.01992 |
| position | −0.11058 | 0.888546 | −0.059 | ⋯ | −0.00178 | −0.04562 | −0.02065 |
| satis | −0.05278 | −0.06301 | −0.19333 | ⋯ | −0.10034 | −0.05922 | −0.09551 |
| status | −0.02791 | 0.145139 | 0.531144 | ⋯ | 0.132492 | 0.056673 | 0.121507 |
| income | 0.909057 | −0.04044 | −0.05276 | ⋯ | −0.0056 | −0.07033 | −0.04193 |
| ot | 0.021266 | 0.023608 | −0.13314 | ⋯ | 0.023477 | −0.04595 | 0.083401 |
| kpi | 0.007824 | −0.01855 | −0.0738 | ⋯ | 0.003695 | 0.970122 | 0.024354 |
| balance | 0.017554 | −0.0018 | −0.07062 | ⋯ | 0.978133 | −0.02992 | −0.0294 |
| years | 0.693373 | −0.04512 | −0.07955 | ⋯ | −0.01691 | −0.01216 | 0.020481 |

Ma trận hệ số tải đã được xoay L*

Tiến trình xoay được thực hiện vào phương pháp Varimax.
**c) Tính điểm nhân tố (nếu cần thiết phân tích sâu hơn về sau)**

Trong khuôn khổ ví dụ thực tiễn nay , chúng em chỉ dừng lại ở xác định tên nhân tố nên chưa tính toán điểm nhân tố.

**Bước 3. Xác định nhân tố**

**a) Dựa vào hệ số tải, nhóm các yếu tố có hệ số tải lớn của cùng 1 nhân tố**

Sau khi áp dụng phương pháp, ta thu được kết quả về hệ số tải như sau:

| | Factor 1 | Factor 2 | Factor 3 | ⋯ | Factor 8 | Factor 9 | Factor 10 |
|---|---|---|---|---|---|---|---|
| age | 0.62484 | −0.14329 | −0.11943 | ⋯ | −0.102601 | −0.02472 | −0.04993 |
| travelfreq | 0.033921 | −0.04435 | 0.749151 | ⋯ | −0.04832 | −0.08079 | −0.04729 |
| depart | 0.079778 | 0.889862 | −0.078804 | ⋯ | −0.01395 | −0.05275 | 0.003766 |
| distance | −0.01241 | −0.01915 | −0.07709 | ⋯ | −0.0399 | −0.02125 | 0.9758 |
| edu | 0.101248 | −0.01124 | −0.16172 | ⋯ | 0.004415 | −0.0791 | 0.020665 |
| gender | −0.19202 | −0.19071 | −0.30345 | ⋯ | −0.07142 | −0.19988 | −0.04504 |
| partlvl | −0.018642 | −0.09945 | 0.14068 | ⋯ | −0.0913 | −0.0634 | −0.09872 |
| lvl | 0.918371 | −0.00995 | −0.07051 | ⋯ | −0.00186 | −0.0746 | −0.01992 |
| position | −0.11058 | 0.888546 | −0.059 | ⋯ | −0.00178 | −0.04562 | −0.02065 |
| satis | −0.05278 | −0.06301 | −0.19333 | ⋯ | −0.10034 | −0.05922 | −0.09551 |
| status | −0.02791 | 0.145139 | 0.531144 | ⋯ | 0.132492 | 0.056673 | 0.121507 |
| income | 0.909057 | −0.04044 | −0.05276 | ⋯ | −0.0056 | −0.07033 | −0.04193 |
| ot | 0.021266 | 0.023608 | −0.13314 | ⋯ | 0.023477 | −0.04595 | 0.083401 |
| kpi | 0.007824 | −0.01855 | −0.0738 | ⋯ | 0.003695 | 0.970122 | 0.024354 |
| balance | 0.017554 | −0.0018 | −0.07062 | ⋯ | 0.978133 | −0.02992 | −0.0294 |
| years | 0.693373 | −0.04512 | −0.07955 | ⋯ | −0.01691 | −0.01216 | 0.020481 |

- Nhận xét: Đối với nhân tố thứ nhất (Factor 1), top 4 hệ số tải lần lượt là 0.9183 ứng với yếu tố *Cấp bậc*, 0.9091 ứng với yếu tố *Thu nhập* , 0.6934 ứng với yếu tố *Số năm trong công ty* và 0.6248 ứng với yếu tố *Tuổi*.
- Thực hiện tìm những hệ số tải lớn (tương tự với nhận xét ở trên), chúng ta có thể nhóm các yếu tố lại như sau:

| **Variable** | **Factor 1** | **Factor 2** | **Factor 3** | **Factor 4** | **Factor 5** | **Factor 6** | **Factor 7** | **Factor 8** | **Factor 9** | **Factor 10** |
|---|---|---|---|---|---|---|---|---|---|---|
| age | 0.6248 | −0.1433 | −0.1194 | 0.2455 | 0.2178 | −0.0713 | −0.0112 | −0.1026 | −0.0247 | −0.0499 |
| travelfreq | 0.0339 | −0.0443 | 0.7492 | 0.3830 | −0.2059 | −0.1230 | −0.0418 | −0.0483 | −0.0808 | −0.0473 |
| depart | 0.0798 | 0.8899 | −0.0788 | 0.0839 | −0.0754 | −0.1438 | 0.0334 | −0.0140 | −0.0528 | 0.0038 |
| distance | −0.0124 | −0.0192 | −0.0771 | 0.0985 | −0.0453 | −0.1065 | −0.0111 | −0.0399 | −0.0213 | 0.9758 |
| edu | 0.1012 | −0.0112 | −0.1617 | 0.2658 | 0.8106 | −0.1453 | 0.0631 | 0.0044 | −0.0791 | 0.0207 |
| gender | −0.1920 | −0.1907 | −0.3034 | −0.0451 | −0.3393 | −0.5483 | 0.0915 | −0.0714 | −0.1999 | −0.0450 |
| partlvl | −0.1864 | −0.0995 | 0.0141 | 0.6314 | 0.0790 | −0.3541 | −0.2074 | −0.0913 | −0.0634 | −0.0987 |
| lvl | 0.9183 | −0.0099 | −0.0705 | 0.1004 | −0.0750 | −0.1572 | −0.0005 | −0.0019 | −0.0746 | −0.0199 |
| position | −0.1106 | 0.8885 | −0.0590 | 0.1285 | −0.0475 | −0.0916 | 0.0288 | −0.0018 | −0.0456 | −0.0207 |
| satis | −0.0528 | −0.0630 | −0.1933 | 0.1268 | −0.1264 | −0.0810 | 0.8465 | −0.1003 | −0.0592 | −0.0955 |
| status | −0.0279 | 0.1451 | 0.5311 | −0.2162 | 0.2588 | −0.0752 | 0.4770 | 0.1325 | 0.0567 | 0.1225 |
| income | 0.9091 | −0.0404 | −0.0528 | 0.0990 | −0.0826 | −0.1475 | −0.0024 | −0.0056 | −0.0703 | −0.0419 |
| ot | 0.0213 | 0.0236 | −0.1331 | 0.5208 | −0.2017 | 0.6722 | 0.1727 | 0.0235 | −0.0460 | 0.0834 |
| kpi | 0.0078 | −0.0186 | −0.0738 | 0.1131 | −0.0760 | −0.1203 | 0.0536 | 0.0037 | 0.9701 | 0.0244 |
| balance | 0.0176 | −0.0018 | −0.0706 | 0.0897 | −0.0520 | −0.1162 | −0.0016 | 0.9781 | −0.0299 | −0.0294 |
| years | 0.6934 | −0.0451 | −0.0795 | −0.0149 | −0.0766 | −0.1130 | −0.0076 | −0.0169 | −0.0122 | 0.0205 |

**b) Gọi tên nhân tố dựa vào yếu tố**

Sau khi đã nhóm các yếu tố dựa vào hệ số tải ứng với nhân tố, ta có thể gọi tên các nhân tố như sau:

| STT | Tên nhân tố | Ứng với các yếu tố |
|---|---|---|
| 1 | Cơ hội thăng tiến | Tuổi, Cấp bậc, Thu nhập, Số năm trong công ty |
| 2 | Môi trường làm việc | Đơn vị(Phòng) , Vị trí công việc |
| 3 | Hoạt động giải trí và xã hội | Tần suất đi du lịch, Tình trạng hôn nhân |
| 4 | Cam kết công việc | Mức độ tham gia công việc, OT |
| 5 | Trình độ học vấn | Trình độ học vấn |
| 6 | Công bằng trong thời gian làm việc | Giới tính , OT |
| 7 | Thỏa mãn trong công việc | Mức độ thỏa mãn công việc |
| 8 | Cân bằng cuộc sống với công việc | Cân bằng cuộc sống |
| 9 | Áp lực doanh số | KPI |
| 10 | Tiện ích đi lại | Khoảng cách từ nhà tới công ty |

Tên nhân tố

**Mã nguồn cho bước tìm số lượng nhân tố**

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.patches as patches
from sklearn.preprocessing import StandardScaler, LabelEncoder

df = pd.read_csv('PTSL - VD6.csv')
df.rename(columns={'Tuoi':'age','Di du lich (xa h i)':'travelfreq','Don vi (Phong )':'depart','Khoang cach toi cty':'distance','Trinh do hoc van':'edu','Gioi tinh':'gender','Muc do tham gia':'partlvl','Cap bac ( 5 : CEO: Quan ly cap cao,....)':'lvl','Vi tri cong viec':'position','Su thoa man trong cviec(moi truong, dong nghiep, sep)':'satis','Tinh trang ket hon':'status','Thu nhap':'income','OT':'ot','KPI':'kpi','Can bang cuoc song':'balance','So nam trong cong ty':'years'}, inplace=True)
columns_to_encode = {'travelfreq', 'depart', 'gender', 'position', 'status', 'ot'}

for col in columns_to_encode:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])

columns_to_scale = df.columns
scaler = StandardScaler()
df[columns_to_scale] = scaler.fit_transform(df[columns_to_scale])
df_corr = df.corr()

# Plot the heatmap
plt.figure(figsize=(15, 10))
sns.heatmap(df_corr, annot=True, cmap='coolwarm', center=0)

# Add title and show the plot
plt.title('Staff Correlation Matrix')
plt.show()

def eigenvalues(df):
    """
    Calculate the eigenvalues of a square matrix represented by a DataFrame.

    Parameters:
    df (pd.DataFrame): A square DataFrame.

    Returns:
    np.ndarray: The eigenvalues of the matrix.
    """
    if df.shape[0] != df.shape[1]:
        raise ValueError("Input DataFrame must represent a square matrix.")

    matrix = df.to_numpy()
    eigenvalues = np.linalg.eigvals(matrix)
    sorted_indices = np.argsort(eigenvalues)[::-1]
    return sorted_indices, eigenvalues[sorted_indices]

df_indices, df_eigenval = eigenvalues(df_corr)

for i in range(len(df_eigenval)):
    proportion = np.sum(df_eigenval[:i+1]) / len(df_eigenval) * 100
    factors = ', '.join(df.columns[df_indices[:i]].tolist())
    print(f'explained variances of {i} factors ({factors}) is {proportion:.2
```

```
        f}%.')

plt.figure(figsize=(8, 5))
plt.plot(range(1, len(df_indices) + 1),
         np.array([df_eigenval[k] / len(df_eigenval) * 100 for k in range(
             len(df_eigenval))]),
         marker='o', linestyle='--')
plt.title('Scree Plot of Eigenvalues from Correlation Matrix')
plt.xlabel('Dimensions')
plt.ylabel('Percentage of explained variances')
plt.xticks(ticks=range(1, len(df_indices) + 1), labels=df.columns[df_indices
    ], rotation=45)

for x in range(1, 17):
    y = df_eigenval[x-1] / len(df_eigenval)
    width = 1
    height = y * 100
    rectangle = patches.Rectangle((x - 0.5, 0), width, height, linewidth=1,
        edgecolor='blue', facecolor='lightblue', alpha=0.5)
plt.gca().add_patch(rectangle)
plt.grid()
plt.show()
```

**Mã nguồn cho bước lựa chọn phương pháp tìm ma trận hệ số tải : PCA**

```

def factor_loadings_matrix(corr_matrix, n_factors):
    """
    Tinh toan ma tran he so tai bang phuong phap Phan tich Nhan to Chinh (
        PCA).
    """
    eigenvalues, eigenvectors = np.linalg.eigh(corr_matrix)
    sorted_indices = np.argsort(eigenvalues)[::-1]
    eigenvalues = eigenvalues[sorted_indices]
    eigenvectors = eigenvectors[:, sorted_indices]
    eigenvalues = eigenvalues[:n_factors]
    eigenvectors = eigenvectors[:, :n_factors]
    loadings = eigenvectors * np.sqrt(eigenvalues)
    return loadings

n_factors = 10
factor_loadings = factor_loadings_matrix(df_corr, n_factors)
factor_loadings_df = pd.DataFrame(factor_loadings, index=df_corr.index,
columns=[f'Factor{i+1}' for i in range(n_factors)])
print("Ma tran he so tai:\n", factor_loadings_df)
output_path = 'C:/Users/User/factor_loadings.xlsx'
factor_loadings_df.to_excel(output_path)
```

**Mã nguồn cho bước xoay nhân tố Varimax**

```

import numpy as np
import pandas as pd

def calculate_varimax_Q(loadings):
    """
    Tinh gia tri ham muc tieu Varimax Q.
    """
    p, m = loadings.shape
    Lambda_squared = loadings**2
    Q = 0
    for j in range(m):
```

```
        sum_lij_squared = np.sum(Lambda_squared[:, j])
        sum_lij_squared_squared = np.sum(Lambda_squared[:, j]**2)
        Q_j = sum_lij_squared_squared - (1 / p) * (sum_lij_squared**2)
        Q += Q_j
    Q /= p
    return Q

def varimax(loadings, gamma=1.0, max_iter=100, tol=1e-6):
    """
    Thuc hien xoay Varimax de toi uu hoa tai trong.
    """
    p, m = loadings.shape
    T = np.eye(m)
    Q_old = calculate_varimax_Q(loadings)
    for iteration in range(max_iter):
        Lambda_rotated = np.dot(loadings, T)
        Lambda_squared = Lambda_rotated**2
        col_sums = np.sum(Lambda_squared, axis=0)
        V = np.dot(Lambda_rotated.T, Lambda_squared - (gamma / p) * col_sums
            )
        U, _, Vt = np.linalg.svd(V)
        T_new = np.dot(U, Vt)
        T = np.dot(T, T_new)
        Q_new = calculate_varimax_Q(np.dot(loadings, T))
        if np.abs(Q_new - Q_old) < tol:
            print(f"Varimax  h i   t   sau {iteration + 1} v ng  l p . Q: {
                Q_new}")
            break
        Q_old = Q_new
    return np.dot(loadings, T)

input_path = 'C:/Users/User/factors_loadings.xlsx'
factor_loadings_df = pd.read_excel(input_path, index_col=0)
loadings = factor_loadings_df.values
rotated_loadings = varimax(loadings)
rotated_loadings_df = pd.DataFrame(rotated_loadings,  columns=[f'Factor{i+1}
    ' for i in range(rotated_loadings.shape[1])],index=factor_loadings_df.
    index)
output_path = 'C:/Users/User/rotated_factor_loadings_with_Q.xlsx'
rotated_loadings_df.to_excel(output_path)
```

# Tổng kết

Trên đây là toàn văn báo cáo của nhóm 6 về chủ đề Phân tích nhân tố . Báo cáo đã đưa ra được mô hình nhân tố trực giao, trình bày các phương pháp ước lượng (phân tích nhân tố chính - PCA, phương pháp hợp lý cực đại - Maximum likelihood method), xoay nhân tố điểm nhân tố và ứng dụng vào một ví dụ thực tế.

Một lần nữa, nhóm chúng em xin gửi lời cảm ơn đến giảng viên, ThS. Lê Xuân Lý đã giảng dạy, đồng hành cùng chúng em trong suốt học phần Phân tích số liệu. Báo cáo của nhóm vẫn còn nhiều thiếu sót, vậy nên chúng em rất mong được thầy và các bạn cùng góp ý, nhận xét để báo cáo trở nên hoàn thiện hơn.

Chúng em xin chân thành cảm ơn!

**NHÓM 6**

# Tài liệu tham khảo


[1] R. A. Johnson, D. W. Wichern (2015) *Applied Multivariate Statistical Analysis*, 6th Edition, Pearson Education Inc.

[2] Factor Analysis Guide with an Example, trang web *statisticsbyjim.com/*

[3] Darton, R. A. (1980). *Rotation in Factor Analysis*. The Statistician, 29(3), 167.

[4] Ertel, S. (2011). *Exploratory factor analysis revealing complex structure*. Personality and Individual Differences, 50(2), 196–200.

[5] Zhang, G., Preacher, K. J. (2015). *Factor Rotation and Standard Errors in Exploratory Factor Analysis*. Journal of Educational and Behavioral Statistics, 40(6), 579–603.