ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

KHOA TOÁN - TIN
Faculty of Applied Mathematics and Informatics

PHÂN TÍCH NHÂN TỐ

(Factor Analysis)

Giảng viên hướng dẫn: ThS. Lê Xuân Lý
Nhóm sinh viên: Nhóm 2
Học phần: MI4024
Mã lớp: 146157

ĐẠI HỌC BÁCH KHOA HÀ NỘI

KHOA TOÁN - TIN

PHÂN TÍCH NHÂN TỐ

HỌC PHẦN: PHÂN TÍCH SỐ LIỆU

Giảng viên hướng dẫn: ThS. Lê Xuân Lý

Sinh viên thực hiện: Nhóm 2 - Lớp 146157

Lê Thị Hằng	20216924
Đào Ngọc Trinh	20216963
Lê Anh Việt	20216967
Nguyễn Mai Phương	20216955
Vũ Quỳnh Anh	20216912
Nguyễn Trường Giang	20216921
Lê Ngọc Hà	20216922
Nguyễn Thị Linh Chi	20216913
Lê Đức Việt	20216968
Vũ Danh Trung Hiếu	20216925
Hà Khánh Linh	20216938

HÀ NỘI - 2024

ĐÁNH GIÁ THÀNH VIÊN

STT	HỌ VÀ TÊN	MSSV	NHIỆM VỤ	ĐÁNH GIÁ
1	Đào Ngọc Trinh	20216963	1 + 2	1.5
2	Lê Anh Việt	20216967	3.1.1 + 3.1.2 + 3.1.3	1.5
3	Nguyễn Mai Phương	20216955	3.1.4 + 3.1.5	1.5
4	Vũ Quỳnh Anh	20216912	3.2	1.5
5	Lê Thị Hằng	20216924	3.3 + 3.4	1.5
6	Nguyễn Trường Giang	20216921	4.1 + 4.2 + 4.3 + 4.4	1.5
7	Lê Ngọc Hà	20216922	5.1 + 5.2 + 5.4	1.5
8	Nguyễn Thị Linh Chi	20216913	5.3 + 5.4	1.5
9	Lê Đức Việt	20216968	6.1	1.5
10	Vũ Danh Trung Hiếu	20216925	6.2	1.5
11	Hà Khánh Linh	20216938	4.5 + 6.3	1.5

LỜI CẢM ƠN

Lời đầu tiên, nhóm báo cáo xin gửi lời cảm ơn đến Khoa Toán - Tin, Đại học Bách khoa Hà Nội đã tạo cơ hội để chúng em thực hiện báo cáo môn học trong một môi trường tốt nhất nhằm đáp ứng yêu cầu hoàn thành báo cáo của chúng em.

Đặc biệt chúng em xin bày tỏ lòng biết ơn sâu sắc đến thầy Lê Xuân Lý, người đã trực tiếp giảng dạy và hướng dẫn chúng em trong học phần "Phân tích số liệu". Đối với chúng em, điều may mắn là đã được thầy tận tình chỉ bảo rất nhiều điều, từ những điều nhỏ nhặt như cách đọc tài liệu chuyên ngành, dịch thuật, cách soạn thảo bằng Latex tới những kiến thức quan trọng trong bài học.

Trong quá trình tìm hiểu cũng có gặp một vài khó khăn về mặt kiến thức song chúng em đã nỗ lực để hoàn thành tốt nhất có thể. Đây là bản báo cáo chúng em tổng hợp lại các kiến thức đã tìm hiểu và thực hành được. Trong báo cáo không tránh khỏi những sai sót, nhóm rất mong những sự góp ý từ thầy để chúng em rút kinh nghiệm và hoàn thiện báo cáo tốt hơn. Chúng em chân thành cảm ơn thầy đã hướng dẫn cho chúng em ở bộ môn này!

Nhóm báo cáo

Nhóm 2

Mục lục

CHƯƠNG 1

GIỚI THIỆU VỀ PHÂN TÍCH NHÂN TỐ

Ý tưởng đầu tiên về phân tích nhân tố đã dược nêu ra bởi nhà toán học người Anh Karl Pearson (1857-1936) và nhà tâm lý học Chales Spearman (1863-1945) và đầu thế kỉ XX. Ban đầu, chỉ dùng để định nghĩa và đo lường trí thông minh, với sự phát triển không ngừng của máy tính hiện đại giúp cho việc tính toán ngày càng dễ dàng hơn và nghiên cứu về phân tích nhân tố cũng được quan tâm hơn.

Mục đích: Phân tích nhân tố là mô tả (nếu được) những mối quan hệ tương quan giữa nhiều biến thông qua ít biến, các biến này không thể quan sát được, được gọi là các yếu tố.

Mối quan hệ tương quan: Mối quan hệ tương quan giữa hai biến thường được biểu diễn qua hai khái niệm chính là Hiệp phương sai và Hệ số tương quan.

- Hiệp phương sai thể hiện mối quan hệ giữa hai biến với nhau, có thể là đồng biến hoặc nghịch biến. Nếu X, Y độc lập thì Cov (X, Y) = 0.
- Hệ số tương quan được sử dụng để đo lường mức độ quan hệ tuyến tính giữa 2 biến.

Giả sử tất cả các biến số trong một nhóm cụ thể có mối quan hệ tương quan cao với nhau, nhưng có các mối quan hệ tương quan tương đối nhỏ với các biến số trong một nhóm khác. Sau đó, tưởng tượng rằng mỗi nhóm biến số đại diện cho một cấu trúc cơ bản duy nhất.

Ví dụ 1.1

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

Ví dụ 1.2

Khảo sát về độ tiêu dùng của một gia đình, ta có thể quan sát được mức tiêu thụ X các tài nguyên khác nhau của các thành viên của hộ gia đình trong vòng khoảng thời gian một tháng (nước, gas, điện,...). Các thành phần đo bị chi phối bởi một số hành vi xã hộ từ các thành viên của hộ gia đình. Các nhân tố chưa được quan sát này được chúng ta quan tâm nhiều hơn so với các định lượng có thể đo đạc được.

Ví dụ 1.3

Giả thuyết rằng tồn tại 2 loại trí thông minh là “thông minh văn học” và “thông minh toán học” 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 như môn “thiên văn học” có điểm số bằng 10 lần lượng thông minh văn học + 6 lần lượng thông minh toán học. 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. Từ đó ta có định nghĩa sai số, là sự chênh lệch giữa điểm thực tế so với điểm dự đoán.

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.

Trong các phần tiếp theo, chúng ta sẽ thảo luận về cách phân tích nhân tố đã được sử dụng trong các lĩnh vực khác nhau như tâm lý học, khoa học xã hội, nghiên cứu thị trường... Hiểu về phân tích nhân tố sẽ cung cấp cho các nhà nghiên cứu và nhà phân tích một công cụ mạnh mẽ để khám phá các mẫu ẩn, giảm độ phức tạp của dữ liệu và thu được thông tin chi tiết có giá trị từ dữ liệu đa biến.

CHƯƠNG 2

MÔ HÌNH PHÂN TÍCH NHÂN TỐ

Nhắc lại

Nếu X là một ma trận mxn, giá trị kỳ vọng của X là một ma trận của các giá trị kỳ vọng:

$$\mathrm{E(X)} = \mathrm{E}\begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{3,1} & x_{3,2} & \dots & x_{3,n} \end{bmatrix} = \begin{bmatrix} E(x_{1,1}) & E(x_{1,2}) & \dots & E(x_{1,n}) \\ E(x_{2,1}) & E(x_{2,2}) & \dots & E(x_{2,n}) \\ \vdots & \vdots & \ddots & \vdots \\ E(x_{3,1}) & E(x_{3,2}) & \dots & E(x_{3,n}) \end{bmatrix}$$

$$\Sigma_{ij} = cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$$

Trong đó $\mu_i = E(X_i)$ là giá trị kỳ vọng của thành phần thứ i của vecto X. Nói cách khác:

$$\Sigma = \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & \dots & E[(X_1 - \mu_1)(X_n - \mu_n)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & \dots & E[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \ddots & \vdots \\ E[(X_n - \mu_n)(X_1 - \mu_1)] & \dots & E[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}$$

Xét vector ngẫu nhiên có thể quan sát được $X^{'} = (X_1, \ldots, X_p)$ 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 $\varepsilon_1, \ldots, \varepsilon_p$ được gọi là các sai số hoặc các nhân tố xác định.

Ta có mô hình nhân tố trực giao:

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2 \\ &\ldots. \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p \end{aligned} \tag{2.1}$$

Hoặc dưới dạng ma trận:

$$\begin{array}{ccccccc} X - \mu & = & L & \times & F & + & \varepsilon \\ (p \times 1) & & (p \times m) & & (m \times 1) & & (p \times 1) \end{array} \tag{2.2}$$

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.

Việc tìm $F_1, \ldots, F_k$ và $\varepsilon_1, \ldots, \varepsilon_k$ là một bài toán không giải được. Tuy nhiên nếu thêm giả thiết:

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

$$E(\varepsilon) = 0;\ \text{cov}(\varepsilon) = E(\varepsilon\varepsilon^{'}) = \Psi_{p\times p} = \begin{bmatrix} \Psi_1 & 0 & \dots & 0 \\ 0 & \Psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Psi_p \end{bmatrix} \tag{2.3}$$

F và ε độc lập (không tương quan):

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

Thì bài toán sẽ có lời giải.

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

$$\begin{array}{ccccccc} X-\mu & = & L & \times & F & + & \varepsilon \\ (p\times 1) & & (p\times m) & & (m\times 1) & & (p\times 1) \end{array} \tag{2.4}$$

Trong đó:

- $\mu_i = E(X_i)$.
- ε_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 (F_j).

Các vector ngẫu nhiên không thể quan sát được F và ε thỏa mãn những điều kiện sau:

$$cov(F) = I;\ cov(\varepsilon) = \psi = diag(\psi_1, ..., \psi_p)$$

- F và ε độc lập.
- $E(F) = 0;\ E(\varepsilon) = 0$.
- $cov(\varepsilon, F) = 0$.

Ta có:

$$\begin{aligned}(\mathbf{X}-\mu)(\mathbf{X}-\mu)^{'} &= (\mathbf{LF}+\varepsilon)(\mathbf{LF}+\varepsilon)^{'} \\ &= (\mathbf{LF}+\varepsilon)((\mathbf{LF})^{'}+\varepsilon^{'}) \\ &= \mathbf{LF}(\mathbf{LF})^{'}+\varepsilon(\mathbf{LF})^{'}+\mathbf{LF}\varepsilon^{'}+\varepsilon\varepsilon^{'}\end{aligned}$$

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

Suy ra nếu có mô hình trực giao trên thì:

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

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.

$$\Sigma = LL^{'}+\psi = LTT^{'}L^{'}+\psi = (LT)(T^{'}L^{'})+\psi = (L^{*})(L^{*})^{'}+\psi$$

Như vậy hệ thức phân tích (2.5) xác định L sai khác một phép biến đổi trực giao.

Từ (2.5) ta có:

$$var(X_i) = l_{i1}^2+\ldots+l_{im}^2+\psi_i = \sigma_{ii}$$

$$cov(X_i,X_j) = l_{i1}l_{j1}+\ldots+l_{im}l_{jm} = \sigma_{ij}$$

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

Với $h_i^2 = l_{i1}^2+\ldots+l_{im}^2$ (2.6) gọi là *phương sai chung*, ψ_i được gọi là *phương sai xác định*.

Ta có:

$$\underbrace{\sigma_{ii}}_{var(X_i)} = \underbrace{l_{i1}^2 + ... + l_{im}^2}_{\text{phương sai chung}} + \underbrace{\psi_i}_{\text{phương sai xác định}} \quad , i = 1, 2, ..., p$$

$$(X - \mu).F^T = (LF + \varepsilon).F^T = LFF^T + \varepsilon.F^T$$

$$Cov(X, F) = E(X - \mu).F^T = L.E(F.F^T) + E(\varepsilon.F^T) = L$$

Tóm lại

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

1. $cov(X) = \Sigma = LL^{'} + \psi$

   Hoặc:

$$Var(X_i) = l_{i1}^2 + \cdots + l_{im}^2 + \psi_i$$
$$cov(X_i, X_j) = l_{i1}l_{j1} + \cdots + l_{im}l_{jm} \tag{2.7}$$

2. $cov(X, F) = L$

   Hoặc:

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

Ví dụ 2.1

Xét ma trận hiệp phương sai:

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

Dễ thấy rằng:

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

$$= \begin{pmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{pmatrix} \begin{pmatrix} 4 & 7 & -1 & 1 \\ 1 & 2 & 6 & 8 \end{pmatrix} + \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}$$

$$= LL' + \psi$$

Do đó Σ được cấu thành từ mô hình nhân tố trực giao m = 2:

$$L = \begin{pmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{pmatrix} \qquad \psi = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}$$

Hiệp phương sai của X_1 có thể được phân ra như sau:

$$\underbrace{19}_{\text{phương sai}} = \underbrace{4^2 + 1^2}_{\text{phương sai chung}} + \underbrace{2}_{\text{phương sai xác định}} = 17 + 2$$

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 ψ_i. Khi $m=p$, bất kỳ ma trận hiệp phương sai Σ nào cũng có thể được phân tách chính xác thành $LL^{'}$ và ψ 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 Σ.
- 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 Σ được biểu diễn bởi $mp+p=12\times2+12=36$ tham số l_{ij} và Ψ_i của mô hình nhân tố.
- Tuy vậy, hầu hết các ma trận hiệp phương sai không thể phân tích được thành $LL^{'}+\Psi$, khi mà số lượng nhân tố m nhỏ hơn nhiều so với p.

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:

$$\Sigma=\begin{pmatrix}1 & .9 & .7\\ .9 & 1 & .4\\ .7 & .4 & 1\end{pmatrix}$$

Áp dụng mô hình nhân tố, ta có:

$$X_1-\mu_1=l_{11}F_1+\varepsilon_1$$

$$X_2-\mu_2=l_{21}F_1+\varepsilon_2$$

$$X_3-\mu_3=l_{31}F_1+\varepsilon_3$$

Cấu trúc hiệp phương sai cho ta:

$$\Sigma = LL^{'} + \psi$$

Hoặc:

$$\begin{array}{lll} 1 = l_{11}^2 + \psi_1 & .90 = l_{11}l_{21} & .70 = l_{11}l_{31} \\ & 1 = l_{21}^2 + \psi_2 & .40 = l_{21}l_{31} \\ & & 1 = l_{31}^2 + \psi_3 \end{array}$$

Ta có cặp phương trình:

$$.70 = l_{11}l_{31}$$

$$.40 = l_{21}l_{31}$$

Suy ra:

$$l_{21} = (\frac{.40}{.70})l_{11}$$

Thay vào phương trình:

$$.90 = l_{11}l_{21}$$

Ta tính ra được: $l_{11}^2 = 1.575$, hay $l_{11} = \pm 1.225$.

Mà do $Var(F_1) = 1$ (theo giả thiết) và $Var(X_1) = 1, l_{11} = Cov(X_1, F_1) = Corr(X_1, F_1)$. Trị tuyệt đối của hệ số tương quan không thể lớn hơn 1, do đó $|l_{11}| = 1.225$ là quá lớn. Ta có:

$$1 = l_{11}^2 + \psi_1 \text{hoặc } \psi_1 = 1 - l_{11}^2$$

Có:

$$\psi_1 = 1 - 1.575 = -.575$$

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

Vì vậy, với ví dụ này khi $m = 1$ ta có thể nhận được nghiệm duy nhất cho phương trình $\Sigma = LL^{'} + \psi$. 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 kì m x n, có tính chất $TT^{'} = T^{'}T = I$. Khi đó ta có thể viết:

$$X - \mu = LF + \varepsilon = LTT^{'}F + \varepsilon = L^{*}F^{*} + \varepsilon \tag{2.8}$$

Trong đó:

$$L^{*} = LT$$

$$F^{*} = T^{'}F$$

Mà ta có:

$$E(F^{*}) = T^{'}E(F) = 0$$

Và:

$$Cov(F^{*}) = T^{'}Cov(F)T = T^{'}T = I$$

Do đó $F^{*} = T^{'}F$ 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 Σ:

$$\Sigma = LT^{'} + \Psi = LTT^{'}L^{'} + \Psi = (L^{*})(L^{*})^{'} + \Psi \tag{2.9}$$

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:

$$L^{*} = LT, L \tag{2.10}$$

Đều có biểu diễn giống nhau. Phương sai chung được tạo ra bởi $LL^{'} = (L^{*})(L^{*})^{'}$ cũng không bị ảnh hưởng bởi cách chọn ma trận T.

CHƯƠNG 3

CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Cho các quan sát $x_1, x_2, ...x_n$ trên các biến tương quan chung p, phân tích nhân tố tìm cách trả lời câu hỏi "Liệu mô hình nhân tố với một số lượng nhân tố nhỏ, có đại điện đầy đủ cho dữ liệu hay không?". Về bản chất, chúng ta giải quyết vấn đề xây dựng mô hình thống kê này bằng cách cố gắng xác minh mối quan hệ hiệp phương sai trong (2.7).

- Ma trận hiệp phương sai mẫu S là công cụ ước lượng của ma trận hiệp phương sai tổng thể chưa biết Σ. Nếu các phần tử nằm ngoài đường chéo của S nhỏ hoặc của ma trận tương quan mẫu R gần bằng không, thì các biến không liên quan với nhau và phân tích nhân tố sẽ không hữu ích.
- Chúng ta sẽ xét hai trong số các phương pháp ước lượng tham số phổ biến nhất, phương pháp thành phần chính và phương pháp hợp lý cực đại. Giải pháp từ một trong hai phương pháp có thể được xoay vòng để đơn giản hóa việc giải thích các yếu tố, việc này sẽ được đề cập tại phần quay nhân tố.

3.1 Phương pháp dựa trên phân tích thành phần chính.

3.1.1 Cơ sở lý thuyết.

Với phương pháp PCA, giả sử chúng ta có ma trận cov(X) = Σ có p cặp trị riêng (Eigenvalue) $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ và hệ vector riêng (Eigenvector) trực chuẩn $e_1, e_2, \ldots, e_p$ là $(\lambda_1 e_1), ..., (\lambda_p e_p)$. Khi đó, ta sẽ có:

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

$\Rightarrow \sum = LL^{'}$, trong đó $L = \begin{bmatrix} \sqrt{\lambda_1} e_1 & \sqrt{\lambda_2} e_2 & \cdots & \sqrt{\lambda_p} e_p \end{bmatrix}$.

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:

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

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

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

Mặc dù đại diện phân tích nhân tố của $\sum$ 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 ε. Người ta cố gắng đưa về mô hình m nhân tố với $m < p$.

Xét trường hợp $p-m$ giá trị riêng cuối cùng của ma trận $\sum$ 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:

$$\sum \approx \lambda_1 e_1 e_1' + \ldots + \lambda_m e_m e_m' = \begin{bmatrix} \sqrt{\lambda_1} e_1 & \ldots & \sqrt{\lambda_m} e_m \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \vdots \\ \sqrt{\lambda_m} e_m' \end{bmatrix} = \underset{(p\times m)}{L} \underset{(m\times p)}{L'} \quad (3.3)$$

Phương sai của các nhân tố cụ thể có thể được coi là các phần tử trên đường chéo của ma trận $\sum - LL'$. Tức là ta có: $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2$, $i = 1, ..., p$

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

$$\Rightarrow \sum \approx LL' + \psi \quad (3.4)$$

3.1.2 Thuật toán.

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

1. Tính vector trung bình mẫu quan sát được $\bar{x}$.

2. Tính các vector độ lệch:

$$x_j - \bar{x} = \begin{bmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jp} \end{bmatrix} - \begin{bmatrix} \bar{x_1} \\ \bar{x_2} \\ \vdots \\ \bar{x_p} \end{bmatrix} = \begin{bmatrix} x_{j1} - \bar{x_1} \\ x_{j2} - \bar{x_2} \\ \vdots \\ x_{jp} - \bar{x_p} \end{bmatrix}, j = 1, \ldots, n \tag{3.5}$$

3. 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 $\sigma^2 = 1$.

4. 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 $\hat{\lambda_1} \geq \hat{\lambda_2} \geq \cdots \geq \hat{\lambda_p}$.

5. Có $m < p$ là số lượng các nhân tố chung. Khi đó ma trận tải nhân tố được ước lượng dưới dạng:

$$\hat{L} = \begin{bmatrix} \sqrt{\hat{\lambda_1}}\hat{e_1} & \sqrt{\hat{\lambda_2}}\hat{e_2} & \ldots & \sqrt{\hat{\lambda_m}}\hat{e_m} \end{bmatrix} \tag{3.6}$$

6. Ước lượng các phương sai cụ thể ψ_i được cho bởi các phần tử trên đường chéo của ma trận $S - LL^{'}$, cụ thể là:

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

$$\Rightarrow \hat{\psi} = \begin{pmatrix} \hat{\psi}_1 & & 0 \\ & \ddots & \\ 0 & & \hat{\psi}_p \end{pmatrix} \tag{3.7}$$

7. Các phương sai chung được ước lượng bởi:

$$\hat{h_i}^2 = \sum_{j=1}^{m} \hat{l_{ij}}^2 \tag{3.8}$$

3.1.3 Một phương pháp cải tiến cho nghiệm nhân tố chính.

Đôi khi, một sửa đổi của phương pháp tiếp cận thành phần chính được xem xét, chúng ta sẽ thực hành phương pháp này với R. Nếu mô hình nhân tố $P = LL' + \Psi$ được xác định, thì m nhân tố chung phải tính đến các phần tử nằm ngoài đường chéo của P, cũng như các phần chung của các phần tử đường chéo:

$$\rho_{ii} = 1 = h_i^2 + \psi_i$$

Giả sử đã có sẵn các ước lượng khởi đầu cho phương sai riêng ψ_i^*. Sau đó thay đường chéo thứ i của R bằng $h_i^{*2} = 1 - \psi_i^*$, chúng ta thu được ma trận tương quan mẫu "giảm":

$$R_r = \begin{bmatrix} h_1^{*2} & r_{12} & \dots & r_{1p} \\ r_{12} & h_2^{*2} & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \dots & h_p^{*2} \end{bmatrix}$$

Chúng ta sẽ phân tách ma trận R_r này:

$$R_r = L_r^* L_r^{*\prime} \tag{3.9}$$

Phương pháp nhân tố chính cho chúng ta ước lượng:

$$L_r^* = \left[\sqrt{\hat{\lambda_1}^*}\hat{e_1}^* \quad \sqrt{\hat{\lambda_2}^*}\hat{e_2}^* \quad \dots \quad \sqrt{\hat{\lambda_m}^*}\hat{e_m}^*\right] \tag{3.10}$$

$$\psi_i^* = 1 - \sum_{j=1}^{m} \overset{*2}{l}_{ij}$$

Và ta cũng có ước lượng lại của phần chung là:

$$\widetilde{h_i}^{*2} = \sum_{j=i}^{m} l_{ij}^{*2} \tag{3.11}$$

Tiếp tục như vậy,đến khi nào sai khác giữa hai ma trận tải trước và sau là không đáng kể thì dừng. Ta có rất nhiều cách chọn khởi đầu cho phương sai của nhân tố riêng. Một cách phổ biến nhất là dựa trên ma trận nghịch đảo của ma trận tương quan mẫu R, là $\Psi_i^* = 1/r^{ii}$ với r^{ii} là phần tử thứ i của R^{-1}. Ước lượng khởi đầu cho tính cộng đồng là:

$$h_i^{*2} = 1 - \psi_i^* = 1 - \frac{1}{r^{ii}} \tag{3.12}$$

Cái này cũng chính là hệ số tương quan bội giữa X_i và $p-1$ biến còn lại.

3.1.4 Một số lưu ý.

Lưu ý 1

Đối với giải pháp thành phần chính, ước lượng của tải trọng cho một nhân tố nhất định không thay đổi khi số lượng các nhân tố được tăng lên.

Giả sử có một mô hình m nhân tố.

Ví dụ:

$$m = 1 : \widehat{L}_{(1)} = \left[\sqrt{\widehat{\lambda}_1}\widehat{e_1}\right]$$
$$m = 2 : \widehat{L}_{(2)} = \left[\sqrt{\widehat{\lambda}_1}\widehat{e_1} \quad \sqrt{\widehat{\lambda}_2}\widehat{e_2}\right]$$

Tổng quát:

$$m = k : \widehat{L}_{(k)} = \left[\sqrt{\widehat{\lambda}_1}\widehat{e_1} \quad \dots \quad \sqrt{\widehat{\lambda}_k}\widehat{e_k}\right]$$
$$m = k+1 : \widehat{L}_{(k+1)} = \left[\ \widehat{L}_{(k)} \ \vdots \ \sqrt{\widehat{\lambda_{k+1}}}\widehat{e_{k+1}}\ \right]$$

Độ gần đúng của ước lượng

Ta có:

$$S - (\widehat{L}\widehat{L}' + \widehat{\psi}) \tag{3.13}$$

Gọi:

$$\Delta = S - \left(\widehat{L}\widehat{L}' + \widehat{\psi}\right) = (\Delta_{ij})$$

Hệ quả:

$$\sum_{i,j}\Delta_{ij}^2 = tr\,\Delta^2 \leq \sum_{i=m+1}^{p}\widehat{\lambda_i}^2 \tag{3.14}$$

Chứng minh:

1. Ta có các phần tử trên đường chéo của ma trận Δ là 0.
2. Tổng bình phương các phần tử của $(S - \hat{L}\hat{L}' - \hat{\psi})$ nhỏ hơn hoặc bằng tổng bình phương các phần tử của $(S - \hat{L}\hat{L}')$.

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

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

$$S = (\hat{\lambda_1}\hat{e_1}\hat{e_1}' + \dots + \hat{\lambda_m}\hat{e_m}\hat{e_m}') + (\hat{\lambda_{m+1}}\hat{e_{m+1}}\hat{e_{m+1}}' + \dots + \hat{\lambda_{m+p}}\hat{e_{m+p}}\hat{e_{m+p}}')$$

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

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

3. Kết hợp thu được kết quả là: $tr(\Delta^2) \leq tr(S - \hat{L}\hat{L}')^2$. Thực hiện biến đổi vế phải ta được:

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

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

Lưu ý 2

Sự đóng góp của các nhân tố vào tổng phương sai mẫu.

Ta có:

$$s_{ii} = \sum_{j=1}^{m} \widehat{l_{ij}}^2 + \widehat{\psi}_i$$

- Sự đóng góp của nhân tố đầu tiên cho s_{ii} là $\widehat{l_{i1}}^2$.
- Sự đóng góp của nhân tố đầu tiên cho tổng phương sai mẫu $tr(S) = s_{11} + s_{22} + ... + s_{pp}$ là $\widehat{l_{11}}^2 + \widehat{l_{21}}^2 + ... + \widehat{l_{p1}}^2$.
- Ta có:

$$\underset{(p\times m)}{\hat{L}} = \left[\sqrt{\widehat{\lambda_1}}\widehat{e_1} \quad \sqrt{\widehat{\lambda_2}}\widehat{e_2} \quad \cdots \quad \sqrt{\widehat{\lambda_m}}\widehat{e_m}\right]$$

- Cột thứ j: $\begin{bmatrix} \widehat{l_{1j}} \\ \vdots \\ \widehat{l_{pj}} \end{bmatrix} = \sqrt{\widehat{\lambda_j}}\widehat{e_j}$
- Ví dụ với $j = 1 \Rightarrow$ Cột đầu tiên của $\widehat{L}$: $\begin{bmatrix} \widehat{l_{11}} \\ \vdots \\ \widehat{l_{p1}} \end{bmatrix}$

Như vậy:

$$\sum_{i=1}^{p} l_{ij}^2 = \left(\sqrt{\widehat{\lambda_j}}\widehat{e_j}\right)' \left(\sqrt{\widehat{\lambda_j}}\widehat{e_j}\right) = \widehat{\lambda_j}\widehat{e_j}'\widehat{e_j} = \widehat{\lambda_j}$$

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

Ngoài ra:

- Tỷ lệ của tổng phương sai mẫu được giải thích thông qua nhân tố đầu tiên:

$$\frac{\widehat{\lambda_1}}{trS} = \frac{\widehat{\lambda_1}}{\sum_{i=1}^{p} s_{ii}}$$

- Tỷ lệ của tổng phương sai mẫu được giải thích thông qua k nhân tố đầu tiên:

$$\frac{\sum_{i=1}^{k} \widehat{\lambda}_i}{\sum_{i=1}^{p} s_{ii}}$$

3.1.5 Chạy code và ví dụ.

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, 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 thuộc tính được xây dựng.

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

Nhận xét: Chúng ta có thể dễ thấy rằng biến 1 và 3 tạo thành một nhóm, biến 2 và 5 tạo thành thành một nhóm khác. Biến 4 gần với nhóm (2, 5) hơn (1, 3). Từ những suy luận trên chúng ta mong muốn dự đoán rằng quan hệ tuyến tính giữa các biến có thể được giải thích bằng 2 đến 3 biến. Hai thành phần đầu tiên $\lambda_1 = 2,85;\ \lambda_2 = 1,81$ là hai giá trị riêng duy nhất lớn hơn 1. Khi đó:

$$\frac{\lambda_1 + \lambda_2}{k} = 0.93$$

Như vậy, hai yếu tố đầu tiên chiếm khoảng 93% tổng phương sai của 5 biến đã chuẩn hóa.

Ta có ma trận tải L:

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

Do ma trận tải có các hệ số tải cao ở cả hai yếu tố ví dụ như ở thuộc tính "hương vị"cao ở cả yếu tố 1 và 2 . Việc đi vào giải thích các nhân tố lúc này là chưa thực sự hiệu quả.

Ma trận phương sai xác định:

$$\psi = \begin{bmatrix} 0.02 & 0 & 0 & 0 & 0 \\ 0 & 0.12 & 0 & 0 & 0 \\ 0 & 0 & 0.02 & 0 & 0 \\ 0 & 0 & 0 & 0.11 & 0 \\ 0 & 0 & 0 & 0 & 0.07 \end{bmatrix}$$

Khi đó ta có kết quả $R = L * L^T + \psi$.

MÃ NGUỒN

```
import pandas as pd
import numpy as np
from numpy import linalg as LA

inputMatrix = np.matrix([
              [1.00, 0.02, 0.96, 0.42, 0.01],
              [0.02, 1.00, 0.13, 0.71, 0.85],
              [0.96, 0.13, 1.00, 0.50, 0.11],
              [0.42, 0.71, 0.50, 1.00, 0.79],
              [0.01, 0.85, 0.11, 0.79, 1.00]])

data = {
  "Taste":                  [1.00, 0.02, 0.96, 0.42, 0.01],
  "Good buy for money":     [0.02, 1.00, 0.13, 0.71, 0.85],
  "Flavor":                 [0.96, 0.13, 1.00, 0.50, 0.11],
  "Suitable for snack":     [0.42, 0.71, 0.50, 1.00, 0.79],
  "Provides lots of energy": [0.01, 0.85, 0.11, 0.79, 1.00]
}

df = pd.DataFrame(data, index = ["1","2","3","4","5"])
df1 = df.T
print(df1)
matrix = inputMatrix
print("Ma tran dau vao:")
print(inputMatrix)
w, v = LA.eig(matrix)
v = -v.T
print('E-value: \n', w)
print('E-vector: \n', v)
```

```
L = np.zeros((1, 5))
print("So cac tri rieng > 1: ", sum(w > 1))
for i in range(0, len (w)):
    if w[i] > 1:
        print("Ung voi tri rieng", w[i], "Ta co vecto rieng", v[i])
        L = np.r_[L, np.sqrt(w[i])*v[i]]
L = np.delete(L, 0, 0)
E = L.T
print("Ma tran L: \n", E)

# Tinh ma tran Communalities
C = np.zeros((5, 1))
for i in range(0, len (E)):
    C[i] = np.power(E[i,0],2) + np.power(E[i,1],2)
print("Ma tran Communalities: \n", C)

# Tinh ma tran PSI (Specific variances)
PSI = np.zeros((5, 1))
for i in range(0, len (C)):
    PSI[i] = 1 - C[i]
PPSI = np.zeros((5, 5))
for i in range(0, len (PSI)):
    PPSI[i,i] = PSI[i]
print("Ma  trn  PSI: \n", PPSI)

# Ma tran Residue
RESIDUE = np.subtract(np.subtract(df1, np.matmul(L.T,L)),PPSI)
print("Ma tran Residue: \n", RESIDUE)
```

Ví dụ 3.2

(Phân tích nhân tố dữ liệu giá cổ phiếu). Tỷ suất lợi nhuận hàng tuần của 5 cổ phiếu (JPMorgan, Citibank, Wells Fargo, Royal Dutch Shell và ExxonMobil) niêm yết trên Sở giao dịch chứng khoán New York được xác định trong khoảng thời gian từ tháng 1 năm 2004 đến tháng 12 2005. Tỷ suất lợi nhuận hàng tuần được điều chỉnh theo việc chia cổ phiếu và cổ tức.Các quan sát trong 103 tuần liên tiếp dường như được phân bổ độc lập, nhưng tỷ suất lợi nhuận giữa các cổ phiếu có mối tương quan với nhau, bởi vì như người ta mong đợi, các cổ phiếu có xu hướng biến động cùng nhau để đáp ứng với các điều kiện kinh tế chung.

Ta có bảng dữ liệu giá cổ phiếu như sau:

Table 8.4 Stock-Price Data (Weekly Rate Of Return)

Week	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	−0.00695
5	0.01082	0.00372	−0.01013	0.02919	0.04098
6	0.01017	−0.01220	−0.00838	0.01371	0.00299
7	0.01113	0.02800	0.00807	0.03054	0.00323
8	0.04848	−0.00515	0.01825	0.00633	0.00768
9	−0.03449	−0.01380	−0.00805	−0.02990	−0.01081
10	−0.00466	0.02099	−0.00608	−0.02039	−0.01267
⋮	⋮	⋮	⋮	⋮	⋮
94	0.03732	0.03593	0.02528	0.05819	0.01697
95	0.02380	0.00311	−0.00688	0.01225	0.02817
96	0.02568	0.05253	0.04070	−0.03166	−0.01885
97	−0.00606	0.00863	0.00584	0.04456	0.03059
98	0.02174	0.02296	0.02920	0.00844	0.03193
99	0.00337	−0.01531	−0.02382	−0.00167	−0.01723
100	0.00336	0.00290	−0.00305	−0.00122	−0.00970
101	0.01701	0.00951	0.01820	−0.01618	−0.00756
102	0.01039	−0.00266	0.00443	−0.00248	−0.01645
103	−0.01279	−0.01437	−0.01874	−0.00498	−0.01637

Gọi x_1, x_2, x_3, x_4, x_5 lần lượt biểu thị tỷ suất lợi nhuận quan sát được hàng tuần của JP Morgan, Citibank, Wells Fargo, Royal Dutch Shell và ExxonMobil. Ta có ma trận R:

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

Tính toán các giá trị riêng của ma trận R, ta thấy hai thành phần đầu tiên $\hat{\lambda_1} = 2.437,\ \hat{\lambda_2} = 1.407$ là hai giá trị riêng duy nhất lớn hơn. Khi đó:

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

Hệ số ma trận tải L:

$$\text{L} = \begin{bmatrix} & \sqrt{\lambda_1}e_1 & \sqrt{\lambda_2}e_2 \\ \text{JP Morgan} & 0.732 & -0.437 \\ \text{Citibank} & 0.831 & -0.280 \\ \text{Wells Fargo} & 0.726 & -0.374 \\ \text{Royal Dutch Shell} & 0.605 & 0.694 \\ \text{ExxonMobil} & 0.563 & 0.719 \end{bmatrix}$$

Nhận xét:

- Có vẻ khá rõ ràng rằng yếu tố đầu tiên, đại diện cho các điều kiện kinh tế chung và có thể được gọi là yếu tố thị trường. Tất cả các cổ phiếu đều có mức độ ảnh hưởng lớn đến yếu tố này và mức độ này là như nhau.
- Yếu tố thứ hai so sánh cổ phiếu ngân hàng với cổ phiếu dầu mỏ. Theo yếu tố này, các ngân hàng có tải trọng âm tương đối lớn và dầu có tải trọng dương lớn. Do đó, yếu tố thứ 2 dường như phân biệt các cổ phiếu trong các ngành khác nhau và có thể được gọi là yếu tố ngành.
- Tóm lại, tỷ suất lợi nhuận dường như được xác định bởi các điều kiện và hoạt động thị trường chung dành riêng cho các ngành khác nhau, cũng như yếu tố dư thừa hoặc yếu tố cụ thể của công ty.

MÃ NGUỒN

```
import pandas as pd
import numpy as np
from numpy import linalg as LA

inputMatrix = np.matrix([
              [1.000, 0.632, 0.511, 0.115, 0.155],
              [0.632, 1.000, 0.574, 0.322, 0.213],
              [0.511, 0.574, 1.000, 0.183, 0.146],
              [0.115, 0.322, 0.183, 1.000, 0.683],
              [0.155, 0.213, 0.146, 0.683, 1.000]])
data = {
  "JP Morgan":        [1.000, 0.632, 0.511, 0.115, 0.155],
  "Citibank":         [0.632, 1.000, 0.574, 0.322, 0.213],
  "Wells Fargo":      [0.511, 0.574, 1.000, 0.183, 0.146],
  "Royal Dutch Shell": [0.115, 0.322, 0.183, 1.000, 0.683],
  "Texaco":           [0.155, 0.213, 0.146, 0.683, 1.000]}

df = pd.DataFrame(data, index = ["1","2","3","4","5"])
df1 = df.T
print(df1)
matrix = []
matrix = inputMatrix
print("Ma tran dau vao:")
print(inputMatrix)
w, v = LA.eig(matrix)
v = v.T
print('E-value: \n', w)
print('E-vector: \n', v)

L = np.zeros((1, 5))
```

```
print("So cac tri rieng > 1: ", sum(w > 1))
for i in range(0, len (w)):
    if w[i] > 1:
        print("Ung voi tri rieng", w[i], "Ta co vecto rieng", v[i])
        L = np.r_[L, np.sqrt(w[i])*v[i]]
L = np.delete(L, 0, 0)
E = L.T
print("Ma tran L: \n", E)

#Tinh ma tran Communalities
C = np.zeros((5, 1))
for i in range(0, len (E)):
    C[i] = np.power(E[i,0],2) + np.power(E[i,1],2)
print("Ma tran Communalities: \n", C)

#Tinh ma tran PSI (Specific variances)
PSI = np.zeros((5, 1))
for i in range(0, len (C)):
    PSI[i] = 1 - C[i]
PPSI = np.zeros((5, 5))
for i in range(0, len (PSI)):
    PPSI[i,i] = PSI[i]
print("Ma  trn  PSI: \n", PPSI)

#Ma tran Residue
RESIDUE = np.subtract(np.subtract(df1, np.matmul(L.T,L)),PPSI)
print("Ma tran Residue: \n", RESIDUE)
```

3.2 Phương pháp ước lượng hợp lý cực đại.

Bài toán đặt ra

Ước lượng tham số từ số liệu.

Ví dụ: Tuổi thọ của 30 người bất kỳ trong 1 viện dưỡng lão $(x_1, x_2, ..., x_n)$

Giả sử phân phối tuổi thọ là phân phối chuẩn $N(\mu, \sigma^2)$. Vậy phải làm thế nào để ước lượng μ, σ^2 từ các số liệu đó?

$\Longrightarrow$ Có 3 phương pháp chính:

1. Phương pháp momen.
2. Phương pháp ước lượng hợp lý cực đại.
3. Phương pháp Bayes.

$\Longrightarrow$ Trong phần này chúng ta sẽ phân tích về 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 likelihood function. Các ước lượng thu được cũng được viết tắt là MLE (Maximum Likelihood Estimates).

Xây dựng công thức ước lượng tham số:

Giả sử có n biến ngẫu nhiên $X_1, X_2, ..., X_n$ với phân phối chuẩn $N(\mu, \sigma^2)$:

$$f(X|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(X-\mu)^2}{2\sigma^2}\right]$$

Và hàm hợp lý:

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

Nếu nhân tố chung F và nhân tố riêng ε được cho là có phân phối nhiều chiều, thì khi đó chúng ta có thể ước lượng ma trận trọng tải L và ma trận phương sai xác định ψ bằng hợp lý cực đại.

Khi F_j và ε_j có phân phối chuẩn, quan sát $X_j - \mu = LF_j + \varepsilon_j; j = \overline{1,n}$ cũng là phân phối chuẩn. Do đó ta có hàm hợp lý cực đại:

$$\begin{aligned} L(\mu,\Sigma) =& (2\pi)^{\frac{-np}{2}} |\Sigma|^{\frac{-n}{2}} e^{-(\frac{1}{2})tr[\Sigma^{-1}(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})^{+}n(\bar{x}-\mu)(\bar{x}-\mu)']} \\ =& (2\pi)^{\frac{-(n-1)p}{2}} |\Sigma|^{\frac{-(n-1)}{2}} e^{-(\frac{1}{2})tr[\Sigma^{-1}(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})']} \\ & x(2\pi)^{\frac{-p}{2}} |\Sigma|^{\frac{-1}{2}} e^{-(\frac{n}{2})[(\bar{x}-\mu)\Sigma^{-1}(\bar{x}-\mu)} \end{aligned} \tag{3.15}$$

Hàm này phụ thuộc vào L, ψ qua $\Sigma = LL' + \psi$. Mô hình chưa xác định vì L có thể sai khác bằng cách nhân với ma trận trực giao. Do đó, để tiện tính toán, người ta thêm điều kiện:

$$L'\psi^{-1}L = \triangle \text{ là ma trận chéo.} \tag{3.16}$$

Ước lượng hợp lý cực đại $\widehat{L}$ và $\widehat{\psi}$ sẽ có được khi ta cực đại hóa hàm Likelihood 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 (3.15) với L và Ψ thỏa mãn điều kiện duy nhất (3.16). Các ước lượng này thỏa mãn:

$$\left(\hat{\Psi}^{-\frac{1}{2}} S_n \hat{\Psi}^{-\frac{1}{2}}\right)\left(\hat{\Psi}^{-\frac{1}{2}} \hat{L}\right) = \left(\hat{\Psi}^{-\frac{1}{2}} \hat{L}\right)(I + \hat{\Delta}) \tag{3.17}$$

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

Từ công thức (3.17), 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:

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

Ngoài ra:

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

Và:

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

Kết quả 2

Cho $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên từ $N_p(\mu, \Sigma)$, trong đó $\Sigma = LL' + \psi$ là ma trận hiệp phương sai cho m nhân tố chung, Ước lượng hợp lý cực đại $\widehat{L}$ và $\widehat{\psi}$ và $\widehat{\mu} = \bar{x}$ với điều kiện $L'\psi^{-1}L = \triangle$ là ma trận chéo. Ước lượng hợp lý cực đại cho tính cộng đồng là:

$$\widehat{h_i^2} = \widehat{l_{i1}^2} + \widehat{l_{i2}^2} + \ldots + \widehat{l_{im}^2}; (i = 1, 2, \ldots, p) \tag{3.18}$$

Hoặc:

$$\begin{pmatrix} \text{Tỷ lệ đóng góp của nhân} \\ \text{tố j trong tổng phương sai} \end{pmatrix} = \frac{\widehat{l_{1j}^2} + \widehat{l_{2j}^2} + \ldots + \widehat{l_{pj}^2}}{s_{11} + s_{22} + \ldots + s_{pp}} \tag{3.19}$$

Chứng minh:

Do L và Ψ thỏa mãn tính chất duy nhất (13) nên ước lượng hợp lý cực đại của L và Ψ 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 ρ của Z có biểu diễn:

$$\begin{aligned} \rho &= V^{-1/2}\Sigma V^{-1/2} \\ &= (V^{-1/2}L)(V^{-1/2}L)' + V^{-1/2}\Psi V^{-1/2} \end{aligned} \tag{3.20}$$

Do đó, ρ có phân tách tương tự như (5) với ma trận hệ số tải $L_z = V^{-1/2}L$ và ma trận phương sai riêng $\Psi_z = V^{-1/2}\Psi V^{-1/2}$. Bằng tính chất bất biến của ước lượng hợp lý cực đại, ước lượng hợp lý cực đại của ρ là:

$$\begin{aligned} \hat{\rho} &= (\hat{V}^{-1/2}\hat{L})(\hat{V}^{-1/2}\hat{L})' + \hat{V}^{-1/2}\hat{\Psi}\hat{V}^{-1/2} \\ &= \hat{L}_z\hat{L'}_z + \hat{\Psi}_z \end{aligned} \tag{3.21}$$

Với $\hat{V}^{-1/2}$ và $\hat{L}$ là ước lượng hợp lý cực đại của $V^{-1/2}$ và L, lần lượt. Như một hệ quả ta cũng có ước lượng hợp lý cực đại của tính cộng đồng tương tự như trên. Ngoài ra ta cũng có ước lượng độ quan trọng của các nhân tố dựa trên:

$$\begin{pmatrix} \text{Tỷ lệ trên tổng (quy chuẩn)} \\ \text{phương sai mẫu theo nhân tố } j \end{pmatrix} = \frac{\hat{l}_{1j}^2 + \hat{l}_{2j}^2 + \cdots + \hat{l}_{pj}^2}{p} \tag{3.22}$$

Để tránh sự có nhiều hơn những kí hiệu nhàm chán, $\hat{L}_{ij}$ sau biểu thị phần tử của $\hat{L}_z$. 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 3.3

Ví dụ để so sánh hai phương pháp phân tích thành phần chính và phương pháp ước lượng hợp lý cực đại.

Các số liệu về chứng khoán gồm $n = 103$ lãi suất chứng khoán hàng tuần trên $k = 5$ loại cổ phiếu được niêm yết trên sàn giao dịch chứng khoán New York.

Sau khi phân tích thành phần chính ma trận tương quan mẫu R người ta giữ lại 2 thành phần chính để giải bài toán phân tích nhân tố. Kết quả được cho trong bảng.

Lời giải với m = 2 nhân tố.

Các biến	ƯL tải trọng của F_1	F_2	ƯL phương sai xác định $\hat{\psi}_i = 1 - \hat{h}_i^2$
1. J P Morgan	0.732	-0.437	0.27
2. Citibank	0.831	-0.280	0.23
3. Wells Fargo	0.726	-0.374	0.33
4. Royal Dutch Shell	0.603	0.694	0.15
5. ExxonMobil	0.563	0.719	0.17
Tỷ lệ tích lũy phương sai mẫu của các thành phần chính với tổng phương sai	0.487	0.769	

Bảng tính bằng ước lượng hợp lý cực đại:

Các biến(công ty)	ƯL tải trọng của F_1	F_2	ƯL phương sai xác định $\hat{\psi}_i = 1 - \hat{h}_i^2$
1. J P Morgan	0.115	0.755	0.42
2. Citibank	0.322	0.788	0.27
3. Wells Fargo	0.182	0.652	0.54
4. Royal Dutch Shell	1.000	-0.000	0.00
5. ExxonMobil	0.683	-0.032	0.53
Tỷ lệ tích lũy phương sai mẫu của các thành phần chính với tổng phương sai	0.323	0.647	

Với phân tích 2 nhân tố, ta tính được ma trận phần dư:

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

Bằng phương pháp hợp lí cực đại, ta tính được ma trận phần dư:

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

Nhận xét:

- Các phần tử của ma trận dư dùng ước lượng hợp lí cực đại nhỏ hơn đáng kể các phần tử của ma trận dư dùng phương pháp phân tích thành phần chính.
- Tỷ lệ tích lũy của các phương sai giải thích bởi các thành phần chính lớn hơn tỷ lệ tương ứng nhân xét bằng phương pháp hợp lý cực đại. Điều đó không ngạc nhiên vì phương pháp phân tích thành phần chính đảm bảo tính chất tối ưu về phương sai.
- Ta thấy rằng các biến có trọng tại dương và xấp xỉ bằng nhau trên nhân tố F1, ta gọi nó là nhân tố thị trường, bởi vì các công ty vào những năm lấy số liệu đều là những cty lớn và có vị thế vững mạnh, F1 thể hiện giá trị cổ phiếu với đà thị trường.
- F2 được gọi là biến tố công nghiệp, lí do là vì các công ty hoạt động trên các lĩnh vực khác nhau và có các yếu tố để sản xuất các sản phẩm khác nhau, tạo lên doanh thu của từng công ty lên có sự chênh lệch như thế.

Ví dụ 3.4

Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên độc lập từ phân phối Poisson với tham số $\lambda > 0$. Tìm ước lượng hợp lý cực đại của λ.

Phân phối của X_i là:

$$P[X_i = x_i] = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \quad \text{với } x_i = 0, 1, 2, ...$$

Hàm hợp lý:

$$lnL(X, \lambda) = (ln\lambda) \sum_{i=1}^{n} X_i - n\lambda - ln \prod_{i=1}^{n} X_i!$$

$$\frac{\delta lnL(X, \lambda)}{\delta \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i$$

Vậy nếu:

$$\frac{\delta lnL(X, \lambda)}{\delta \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i = 0$$

Thì:

$$\lambda = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Mặt khác:

$$\frac{\delta^2 lnL(X, \lambda)}{\delta \lambda^2} = -\frac{\sum_{i=1}^{n} X_i}{\lambda^2} < 0$$

Do đó, tại $\lambda = \frac{1}{n} \sum_{i=1}^{n} X_i$ thì $\frac{\delta^2 lnL(X, \lambda)}{\delta \lambda^2} < 0$ tức là hàm $L(X, \lambda)$ đạt cực đại.

Từ đó suy ra:

$$\lambda = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Là ước lượng hợp lý cực đại của λ.

Ví dụ 3.5

Giả sử cân nặng của các nữ sinh viên đại học chọn ngẫu nhiên tuân theo phân phối chuẩn với giá trị trung bình μ chưa biết và độ lệch chuẩn σ. Một mẫu ngẫu nhiên gồm 10 sinh viên có trọng lượng như sau :

115 122 130 127 149 160 152 138 149 180

Dựa trên định nghĩa hãy xác định hàm hợp lý xảy ra và ước lượng khả năng hợp lý cực đại của μ (trọng lượng trung bình của tất cả nữ sinh viên đại học).

Ta có các quan sát tuân theo phân phối chuẩn $X_1, X_2, X_3, ...X_n \sim N(\mu, \sigma^2)$ với $n = \overline{1,10}$

Hàm mật độ xác suất của X_i là:

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

Cho $-\infty < x < \infty$. Không gian tham số là:

$$\Omega = \{(\mu, \sigma) : -\infty < \mu < \infty; 0 < \sigma < \infty\}$$

Do đó hàm hợp lý là :

$$L(\mu, \sigma^2) = \sigma^{-n}(2\pi)^{-\frac{n}{2}} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

Với $-\infty < \mu < \infty$ và $0 < \sigma < \infty$. Ta thấy khi tối đa hàm khả năng đối với biến μ thì hàm ước lượng hợp lý cực đại của μ là:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

Dựa trên mẫu đa cho, ước lượng hợp lý cực đại của μ là :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{10}(115 + ... + 180) = 142.2$$

3.3 Kiểm định mẫu lớn cho số lượng của nhân tố chung.

Giả định về tính chuẩn của mẫu cho phép ta kiểm định tính chính xác của mô hình. Chúng ta xét mô hình có m nhân tố. Trong trường hợp $\Sigma = LL' + \psi$, ta kiểm định:

$$\begin{array}{cccccccc} H_0: & \Sigma & = & L & \times & L' & + & \psi \\ & (p\times p) & & (p\times m) & & (m\times p) & & (p\times p) \end{array} \tag{3.23}$$

Đối thuyết H_1 : Σ là một ma trận xác định dương bất kì (không phân tích được như H_0).

Khi Σ không có các điều kiện đặc biệt, giá trị lớn nhất của hàm hợp lý:

$$L(\hat{\mu},\hat{\Sigma}) = \frac{1}{(2\pi)^{\frac{np}{2}}} e^{-\frac{np}{2}} \frac{1}{|\hat{\Sigma}|^{\frac{n}{2}}}$$

Theo $H_0, \Sigma = LL^{'} + \Psi$, trong trường hợp này, thay $\hat{\mu} = \bar{x}$ 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 Ψ, ta có giá trị lớn nhất của hàm hợp lý tỷ lệ thuận với:

$$\begin{aligned} &|\hat{\Sigma}|^{-n/2} \exp\left\{ -\frac{1}{2}\text{tr}\left[\hat{\Sigma}^{-1} \left(\sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^{'} \right) \right] \right\} \\ &\quad = \left| \hat{\mathbf{L}}\hat{\mathbf{L}}^{'} + \hat{\Psi} \right|^{-n/2} \exp\left\{ -\frac{1}{2} n \,\text{tr}\left[(\hat{\mathbf{L}}\hat{\mathbf{L}} + \hat{\Psi})^{-1} \mathbf{S}_n \right] \right\} \end{aligned} \tag{3.24}$$

Dựa vào hai kết quả trên, ta có tỷ lệ hợp lý cho kiểm định H_0 là:

$$\begin{aligned} -2\ln(\wedge) &= -2\ln\left[\frac{\text{cực đại hợp lý với } H_0}{\text{cực đại hợp lý}} \right] \\ &= -2\ln\left(\frac{|\widehat{\Sigma}|}{|S_n|} \right)^{-n/2} + n[tr(\widehat{\Sigma}^{-1} S_n) - p] \end{aligned} \tag{3.25}$$

Với bậc tự do:

$$v - v_0 = \frac{1}{2}p(p+1) - [p(m+1) - \frac{1}{2}m(m-1)] = \frac{1}{2}[(p-m)^2 - p - m] \tag{3.26}$$

Ngoài ra, nếu $tr(\widehat{\Sigma}^{-1}S_n) - p = 0$ thì $\widehat{\Sigma} = \widehat{L}\widehat{L}' + \widehat{\Psi}$ là ước lượng hợp lý cực đại của $\Sigma = LL' + \Psi$

Do vậy, ta chỉ còn:

$$-2\ln(\wedge) = n\ln\left(\frac{|\widehat{\Sigma}|}{|S_n|}\right) \tag{3.27}$$

Bartlett còn chỉ ra rằng: Xấp xỉ Chi bình phương cho $-2\ln(\wedge)$ có thể được cải tiến bằng cách thay thế n ở trên bằng $(n-1-(2p+4m+5)/6)$. Sử dụng chỉnh sửa của Bartlett, ta bác bỏ H_0 với mức ý nghĩa α nếu:

$$\left(n-1-\frac{2p+4m+5}{6}\right)\ln\frac{|\hat{L}\hat{L}' + \hat{\Psi}|}{|S_n|} > \chi^2_{\frac{(p-m)^2-p-m}{2}}(\alpha) \tag{3.28}$$

Với n và n - p lớn. Mặt khác, do số bậc tự do $\frac{1}{2}[(p-m)^2 - p - m]$ phải là một số nguyên dương nên:

$$m < \frac{1}{2}(2p+1-\sqrt{8p+1}) \tag{3.29}$$

Để có thể áp dụng kiểm định.

Ví dụ 3.6

Kiểm tra hai nhân tố.

Kiểm tra giả thuyết $H_0 : \Sigma = LL' + \psi$ với $m = 2, \alpha = 0.05$.

	Principal component			Maximum likelihood		
Các biến	ƯL tải trọng của F_1	ƯL tải trọng của F_2	ƯL phương sai xác định $\hat{\psi}_i = 1 - \hat{h}_i^2$	ƯL tải trọng của F_1	ƯL tải trọng của F_2	ƯL phương xác định sai $\hat{\psi}_i = 1 - \hat{h}_i^2$
1. J P Morgan	0.732	-0.437	0.27	0.115	0.755	0.42
2. Citibank	0.831	-0.280	0.23	0.322	0.788	0.27
3. Wells Fargo	0.726	-0.374	0.33	0.182	0.652	0.54
4. Royal Dutch Shell	0.603	0.694	0.15	1.000	-0.000	0.00
5. ExxonMobil	0.563	0.719	0.17	0.683	-0.032	0.53
Tỷ lệ tích lũy phương sai mẫu của các thành phần chính với tổng phương sai	0.487	0.769		0.323	0.647	

Thống kê thử nghiệm trong (3.14), dựa trên tỷ lệ của các phương sai tổng quát:

$$\frac{|\widehat{\Sigma}|}{|S_n|} = \frac{|\widehat{L}\widehat{L}' + \widehat{\psi}|}{|S_n|}$$

Gọi $\widehat{V}^{-1/2}$ là ma trận đường chéo sao cho $\widehat{V}^{-1/2} S_n \hat{V}^{-1/2} = R$. Theo tính chất của định thức, ta có:

$$|\widehat{V}^{-1/2}||\widehat{L}\widehat{L}' + \widehat{\psi}||\widehat{V}^{-1/2}| = |\widehat{V}^{-1/2}\widehat{L}\widehat{L}'\widehat{V}^{-1/2} + \widehat{V}^{-1/2}\widehat{\psi}\widehat{V}^{-1/2}|$$

Và:

$$|\widehat{V}^{-1/2}||S_n||\widehat{V}^{-1/2}| = |\widehat{V}^{-1/2} S_n \widehat{V}^{-1/2}|$$

Do đó:

$$\begin{aligned}\frac{|\widehat{\Sigma}|}{S_n} &= \frac{|\widehat{V}^{-1/2}|}{|\widehat{V}^{-1/2}|}\frac{|\widehat{L}\widehat{L}'+\widehat{\psi}|}{|S_n|}\frac{|\widehat{V}^{-1/2}|}{|\widehat{V}^{-1/2}|} \\ &= \frac{|\widehat{V}^{-1/2}\widehat{L}\widehat{L}'\widehat{V}^{-1/2}+\widehat{V}^{-1/2}\widehat{\psi}\widehat{V}^{-1/2}|}{|\widehat{V}^{-1/2}S_n\widehat{V}^{-1/2}|} \\ &= \frac{|\widehat{L}_z\widehat{L}'_z+\widehat{\psi}_z|}{|R|}\end{aligned} \tag{3.30}$$

Mặt khác:

$$\frac{|\widehat{L}_z\widehat{L}'_z+\widehat{\psi}_z|}{|R|} = \frac{\begin{vmatrix} 1.000 & & & & \\ 0.632 & 1.000 & & & \\ 0.513 & 0.572 & 1.000 & & \\ 0.115 & 0.322 & 0.182 & 1.000 & \\ 0.103 & 0.246 & 0.146 & 0.683 & 1.000 \end{vmatrix}}{\begin{vmatrix} 1.000 & & & & \\ 0.632 & 1.000 & & & \\ 0.510 & 0.574 & 1.000 & & \\ 0.115 & 0.322 & 0.182 & 1.000 & \\ 0.154 & 0.213 & 0.146 & 0.683 & 1.000 \end{vmatrix}} = \frac{0.17898}{0.17519} = 1.0216$$

Sử dụng chỉnh sửa của Bartlett, ta đánh giá thống kê thử nghiệm trong (3.28):

$$\begin{aligned}&\Big[n-1-\frac{(2p+4m+5)}{6}\Big]ln\frac{|\widehat{L}\widehat{L}'+\widehat{\psi}|}{|S_n|} \\ =&\Big[103-1-\frac{(10+8+5)}{6}\Big]ln(1.0216) \\ =&2.10\end{aligned}$$

Khi đó $\frac{1}{2}\big[(p-m)^2-p-m\big] = \frac{1}{2}\big[(5-2)^2-5-2\big] = 1$ và mức ý nghĩa $\alpha = 0.05$ suy ra:

$$\chi_1^2(0.05) = 3.84 > 2.10$$

Kết luận: Vậy với mức ý nghĩa 5% ta không thể bác bỏ H_0. Dữ liệu không mâu thuẫn với mô hình hai yếu tố, trên thực tế mức ý nghĩa quan sát được, hoặc giá trị $P, P[\chi_1^2 > 2.10] = 0.15$ nghĩa là H_0 sẽ không bị bác bỏ ở bất kỳ mức ý nghĩa hợp lý nào.

3.4 Các lý thuyết kiểm định bổ sung.

3.4.1 Kiểm định độ cầu của Bartlett.

Kiểm định độ cầu của Bartlett (Bartlett's test of sphericity): Dùng để xem xét các biến quan sát trong nhân tố có tương quan với nhau hay không bằng cách sử dụng ma trận tương quan quan sát với ma trận đơn vị.

Điều kiện cần để áp dụng phân tích nhân tố là các biến quan sát phản ánh những khía cạnh khác nhau của cùng một nhân tố phải có mối tương quan với nhau.

Chương trình tính ra xấp xỉ chi bình phương và giá trị sig(P). Nếu sig < 0.05 ta nói các biến có tương quan với nhau và bộ dữ liệu có ý nghĩa thống kê và có thể phân tích nhân tố.

Bartlett (1951) đề xuất thống kê này để xác định tính phù hợp của ma trận tương quan đối với phân tích nhân tố. Thống kê được phân phối xấp xỉ chi bình phương với $df = \frac{p(p-1)}{2}$ và được đưa ra bởi công thức:

Công thức

$$\chi^2 = -\ln|R|\left(n - 1 - \frac{2p+5}{6}\right)$$

Trong đó:

- $|R|$ là định thức của ma trận tương quan.
- N là cỡ mẫu.
- p là số biến.

Chú ý: Phép thử này đòi hỏi tính chuẩn mực đa biến. Nếu điều kiện này không được đáp ứng, tiêu chí Kaiser-Meyer-Olkin (KMO) vẫn có thể được sử dụng.

3.4.2 Hệ số KMO (Kaiser-Meyer-Olkin)

Chỉ số (Hệ số) KMO là một biện pháp thống kê để xác định mức độ phù hợp của dữ liệu để phân tích nhân tố . Phép thử đo lường mức độ đầy đủ của việc lấy mẫu đối với mỗi biến trong mô hình và mô hình hoàn chỉnh.

Công thức

$$KMO = \frac{\sum\sum_{j\neq k} r_{jk}^2}{\sum\sum_{j\neq k} r_{jk}^2 + \sum\sum_{j\neq k} p_{jk}^2}$$

Trong đó:

- r_{jk} là hệ số tương quan của biến thứ j và biến thứ k ($j \neq k$).
- p_{jk} là hệ số tương quan riêng phần của biến thứ j và biến thứ k ($j \neq k$) được kiểm soát bởi tất cả các biến quan sát khác.

Độ hiệu quả thang đo:

Kaiser đề xuất rằng:

- KMO $\geq$ 0.9: Rất tốt.
- 0.8 $\leq$ KMO < 0.9: Tốt.
- 0.7 $\leq$ KMO < 0.8: Được.
- 0.6 $\leq$ KMO < 0.7: Tạm được.
- 0.5 $\leq$ KMO < 0.6: Xấu.
- KMO < 0.5: Không chấp nhận được.

Theo trích dẫn của IBM, KMO lớn hơn 0.8 có thể được coi là tốt, tức là một dấu hiệu cho thấy phân tích thành phần hoặc nhân tố sẽ hữu ích cho các biến này. Điều này thường xảy ra khi hầu hết các tương quan bậc 0 là dương. Giá trị KMO nhỏ hơn 0.5 xảy ra khi hầu hết các tương quan bậc 0 là âm. Giá trị KMO nhỏ hơn 0.5 yêu cầu hành động khắc phục bằng cách xóa các biến vi phạm.

Thông thường khi phân tích nhân tố ta chỉ cần chỉ số KMO > 0.5.

Trị số của KMO phải đạt giá trị 0.5 trở lên (0.5$\leq$ KMO $\leq$ 1) là điều kiện đủ để phân tích nhân tố là phù hợp. Nếu trị số này nhỏ hơn 0.5, thì phân tích nhân tố có khả năng không thích hợp với tập dữ liệu nghiên cứu.

Cách tính hệ số tương quan riêng phần

Sử dụng hồi quy tuyến tính.

Một cách đơn giản để tính hệ số tương quan riêng phần cho một số dữ liệu là giải hai bài toán hồi quy tuyến tính liên quan, lấy phần dư và tính toán tương quan giữa các phần dư.

Gọi X và Y là các biến ngẫu nhiên nhận giá trị thực và gọi Z là biến ngẫu nhiên có giá trị vectơ n-chiều.Ta viết x_i, y_i và z_i để biểu thị thứ i của N i.i.d (độc lập và phân phối đồng nhất). Các quan sát từ một số phân phối xác suất chung trên các biến thực ngẫu nhiên X, Y và Z, với z_i đã được tăng thêm với 1 để cho phép số số hạng không đổi trong hồi quy. Việc giải quyết vấn đề hồi quy tuyến tính tương đương với việc tìm (n + 1) vectơ hệ số hồi quy thứ nguyên $\mathbf{w}_X^*$và $\mathbf{w}_Y^*$ như là:

$$\mathbf{w}_X^* = argmin_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (x_i - \langle w, x_i \rangle)^2 \right\}$$

$$\mathbf{w}_y^* = argmin_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (x_i - \langle w, x_i \rangle)^2 \right\}$$

Với N là số quan sát và $\langle w, v \rangle$ là tích vô hướng của w và v.

Phần dư sau đó là:

$$e_{X,i} = x_i - \langle w_X^*, z_i \rangle$$

$$e_{Y,i} = y_i - \langle w_Y^*, z_i \rangle$$

Và hệ số tương quan riêng phần của mẫu sau đó được đưa ra theo công thức thông thường cho tương quan mẫu, nhưng giữa các giá trị suy ra mới này:

$$\hat{\rho}_{XY.\mathbf{Z}} = \frac{N\sum_{i=1}^{N} e_{X,i}e_{Y,i} - \sum_{i=1}^{N} e_{X,i}\sum_{i=1}^{N} e_{Y,i}}{\sqrt{N\sum_{i=1}^{N} e_{X,i}^2 - \left(\sum_{i=1}^{N} e_{X,i}\right)^2}\sqrt{N\sum_{i=1}^{N} e_{Y,i}^2 - \left(\sum_{i=1}^{N} e_{y,i}\right)^2}}$$

$$= \frac{N\sum_{i=1}^{N} e_{X,i}e_{Y,i}}{\sqrt{N\sum_{i=1}^{N} e_{X,i}^2}\sqrt{N\sum_{i=1}^{N} e_{Y,i}^2}}$$

CHƯƠNG 4

QUAY NHÂN TỐ

4.1 Đặt vấn đề.

Xét ma trận tải L:

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

Tuy nhiên ta có thấy trong bảng số liệu các hệ số tải của các nhân tố biểu hiện khá cao.

Ví dụ: Ở Hương liệu 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. Lý tưởng nhất là mỗi biến sẽ xuất hiện như một yếu tố đóng góp đáng kể trong một cột.

$\Longrightarrow$ 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.

4.2 Phép quay 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 quay 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 quay nhân tố có 2 loại:

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

$\Longrightarrow$ Phép quay xiên thích hợp hơn phép quay 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 quay. 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:

$$L^* = LT \text{ trong đó } TT^{'} = T^{'}T = 1 \tag{4.1}$$

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

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

$$\hat{L}\hat{L}' + \hat{\Psi} = \hat{L}TT'\hat{L} + \hat{\Psi} = \hat{L}^*\hat{L}^{*T} + \hat{\Psi} \tag{4.2}$$

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

$$S_n - \hat{L}\hat{L}' - \hat{\Psi} = S_n - \hat{L}^*\hat{L}^{*T} - \hat{\Psi}$$

không đổi sau khi thực hiện quay nhân tố.

Hơn nữa các phương sai cụ thể $\hat{\Psi}_i$, và cộng đồng $\hat{h}_i$ không thay đổi. Do đó từ quan điểm toán học việc thu được $\hat{L}$ hay $\hat{L}'$ là không quan trọng. Vì tải trọng ban đầu có thể không dễ hiểu, nên thông thường, người ta sẽ quay chúng cho đến khi đạt được "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 quay 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 Φ 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:

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

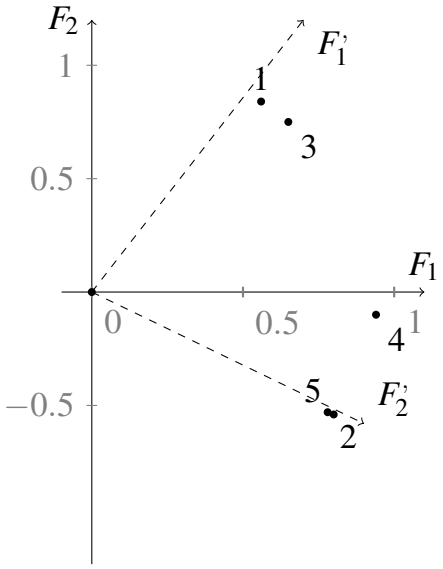
Đối với $m > 2$, các định hướng không dễ dàng hình dung được và độ lớn của các tải trọng quay phải được kiểm tra để tìm ra cách giải thích có ý nghĩa đối với dữ liệu gốc. Việc lựa chọn một ma trận trực giao T thỏa mãn phép đo, phân tích cấu trúc đơn giản sẽ được xem xét ngay sau đây.

Ví dụ 4.1

Chúng ta xét lại ví dụ (3.1) để hiểu thêm về phương pháp quay nhân tố:

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

Tìm cách quay 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.



⋆ Các cặp tải nhân tố được vẽ trên hệ tọa độ OF_1F_2 hệ tọa độ $OF_1'F_2'$là hệ tọa độ sau khi quay.

$$L' = \begin{bmatrix} \text{vị} & 0.02 & 0.99 \\ \text{giá} & 0.87 & 0.0065 \\ \text{hương} & 0.13 & 0.96 \\ \text{nhanh} & 0.81 & 0.4 \\ \text{n.lượng} & 0.97 & -0.016 \end{bmatrix}$$

⋆ Yếu tố 2 có tải trọng cao với vị và hương có thể là yếu tố về hương vị.

⋆ Yếu tố 1 có tải cao với giá , ăn nhanh, năng lượng có thể là yếu tố quyết định về giá thành và dinh dưỡng.

Ví dụ 4.2

Lawley và Maxwell trình bày ma trận tương quan mẫu về điểm thi trong $p = 6$ môn và 220 học sinh nam. Ma trận tương quan là:

$$\mathbf{R} = \begin{bmatrix} \textit{Gaelic} & \textit{English} & \textit{History} & \textit{Arithmetic} & \textit{Algebra} & \textit{Geometry} \\ 1.0 & .439 & .410 & .288 & .329 & .248 \\ & 1.0 & .351 & .354 & .320 & .329 \\ & & 1.0 & .164 & .190 & .181 \\ & & & 1.0 & .595 & .470 \\ & & & & 1.0 & .464 \\ & & & & & 1.0 \end{bmatrix}$$

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:

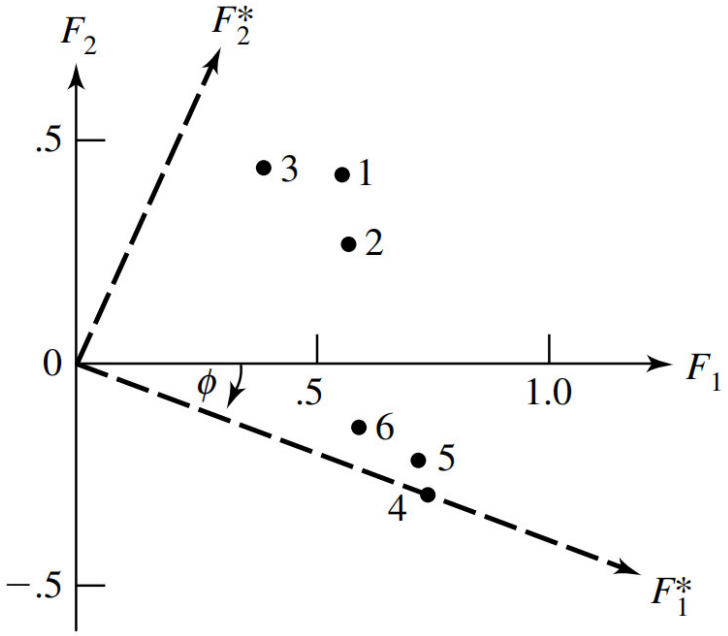
Variable	Estimated factor loadings F_1	F_2	Communalities $\hat{h}_i^2$
1. Gaelic	.553	.429	.490
2. English	.568	.288	.406
3. History	.392	.450	.356
4. Arithmetic	.740	−.273	.623
5. Algebra	.724	−.211	.569
6. Geometry	.595	−.132	.372

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

- Tất cả các biến đổi ở F_1 đều có tải dương lên nhân tố đầu tiên. Chúng ta có thể dự đoán yếu tố này dự đoán phản ứng tổng thể của học sinh đối với hướng dẫn và có thể được gọi là yếu tố trí thông minh chung.
- 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 quay 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 $(\hat{l_{41}}, \hat{l_{42}})$. Tất cả các điểm đều nằm trong góc phần tư thứ nhất và hai nhóm điểm riêng biệt được định nghĩa rõ ràng hơn.

Biểu diễn các tải nhân tố lên hệ trục tọa độ:



Các hệ số tải sau khi sử dụng phương pháp quay nhân tố trực giao:

Variable	Estimated rotated factor loadings F_1^*	F_2^*	Communalities $\hat{h}_i^{*2} = \hat{h}_i^2$
1. Gaelic	.369	.594	.490
2. English	.433	.467	.406
3. History	.211	.558	.356
4. Arithmetic	.789	.001	.623
5. Algebra	.752	.054	.568
6. Geometry	.604	.083	.372

Nhận xét:

- Các biến kiểm tra năng lực toán học có tải cao hơn trên F_1^* và có tải không đáng kể trên $F_2^* \Rightarrow$ Yếu tố đầu tiên có thể được gọi là yếu tố năng lực toán học.
- Tương tự thì 3 biến kiểm tra khả năng ngôn ngữ và xã hội có tải trong cao trên F_2^* và tải trọng nhỏ trên $F_1^* \Rightarrow$ Yếu tố thứ hai có thể được gọi là yếu tố về khả năng ngôn ngữ và xã hội.
- Yếu tố thông minh chung được xác định ban đầu nằm ẩn trong yếu tố F_1^* và F_2^*.

4.4 Các phương pháp quay nhân tố phân tích.

4.4.1 Giới thiệu.

Phương pháp quay 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 quay 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 quay trực giao.

Một số lượng lớn các phương pháp quay 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 quay 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 quay Quartimax.

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

Định nghĩa 4.3

(Phương pháp quay 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:

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

Chỉ số $j=00$ 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:

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

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 quay 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 quay quartimax, quy trình tính toán liên quan đến phép quay 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

Ma trận hệ số các ước tính khả năng xảy ra tối đa của các tải nhân tố ban đầu đối với dữ liệu giá cổ phiếu của các công ty. Giả định với mô hình m = 2 nhân tố.

$$\mathbf{L} = \begin{bmatrix} & F_1 & F_2 \\ \text{J P Morgan} & 0.115 & 0.755 \\ \text{Citibank} & 0.322 & 0.788 \\ \text{Wells Fargo} & 0.182 & 0.652 \\ \text{Royal Dutch Shell} & 1.000 & -0.000 \\ \text{ExxonMobil} & 0.683 & 0.32 \end{bmatrix}$$

Hệ số tải sau khi quay:

$$L' = \begin{bmatrix} & F_1^* & F_2^* \\ \text{J P Morgan} & 0.763 & 0.24 \\ \text{Citibank} & 0.821 & 0.227 \\ \text{Wells Fargo} & 0.669 & 0.104 \\ \text{Royal Dutch Shell} & 0.118 & 0.993 \\ \text{ExxonMobil} & 0.113 & 0.675 \end{bmatrix}$$

⋆ Các hệ số tải chỉ ra rằng cổ phiếu ngân hàng (JP Morgan, Citibank và Wells Fargo) chịu tải cao về yếu tố 1, trong khi cổ phiếu dầu (Royal Dutch Shell và ExxonMobil) chịu tải cao về yếu tố 2.

⋆ Yếu tố 1 đại diện cho các lực lượng kinh tế duy nhất khiến cổ phiếu ngân hàng di chuyển cùng nhau.

⋆ Yếu tố 2 đại diện cho các điều kiện kinh tế ảnh hưởng đến các kho dự trữ dầu.

Ví dụ 4.4

Tải luân phiên cho dữ liệu 10 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:

$$\mathbf{R} = \begin{bmatrix} 1.000 & .6386 & .4752 & .3227 & .5520 & .3262 & .3509 & .4008 & .1821 & -.0352 \\ .6386 & 1.0000 & .4953 & .5668 & .4706 & .3520 & .3998 & .5167 & .3102 & .1012 \\ .4752 & .4953 & 1.0000 & .4357 & .2539 & .2812 & .7926 & .4728 & .4682 & -.0120 \\ .3227 & .5668 & .4357 & 1.0000 & .3449 & .3503 & .3657 & .6040 & .2344 & .2380 \\ .5520 & .4706 & .2539 & .3449 & 1.0000 & .1546 & .2100 & .4213 & .2116 & .4125 \\ .3262 & .3520 & .2812 & .3503 & .1546 & 1.0000 & .2553 & .4163 & .1712 & .0002 \\ .3509 & .3998 & .7926 & .3657 & .2100 & .2553 & 1.0000 & .4036 & .4179 & .0109 \\ .4008 & .5167 & .4728 & .6040 & .4213 & .4163 & .4036 & 1.0000 & .3151 & .2395 \\ .1821 & .3102 & .4682 & .2344 & .2116 & .1712 & .4179 & .3151 & 1.0000 & .0983 \\ -.0352 & .1012 & -.0120 & .2380 & .4125 & .0002 & .0109 & .2395 & .0983 & 1.0000 \end{bmatrix}$$

Hệ số tải được tính theo hai phương pháp nhân tố chủ yếu và ước lượng hợp lý cực đại:

Table 9.4

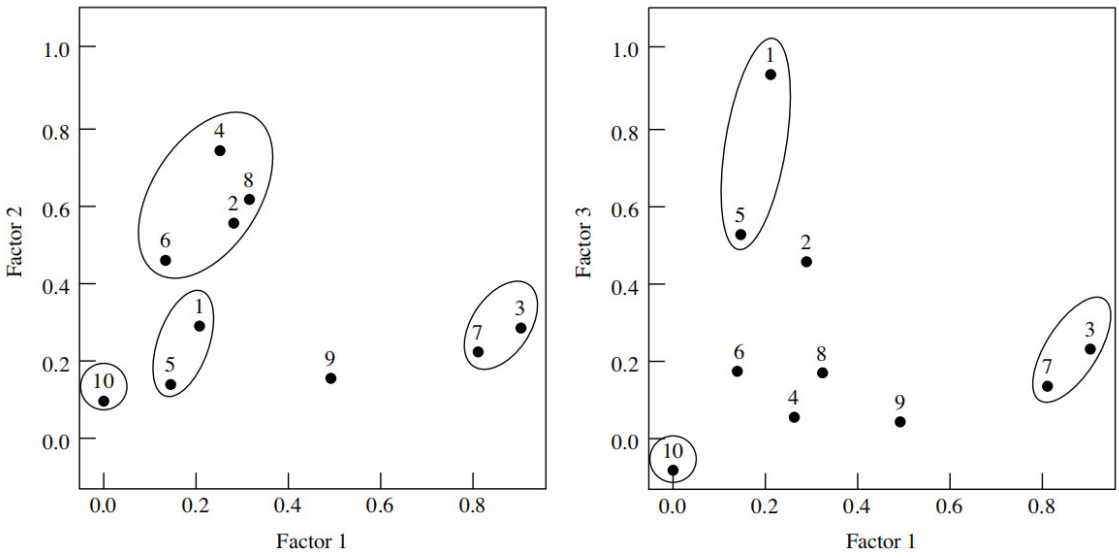
	Principal component					Maximum likelihood				
	Estimated factor loadings				Specific variances	Estimated factor loadings				Specific variances
Variable	F_1	F_2	F_3	F_4	$\tilde{\psi}_i = 1 - \tilde{h}_i^2$	F_1	F_2	F_3	F_4	$\hat{\psi}_i = 1 - \hat{h}_i^2$
1. 100-m run	.696	.022	−.468	−.416	.12	.993	−.069	−.021	.002	.01
2. Long jump	.793	.075	−.255	−.115	.29	.665	.252	.239	.220	.39
3. Shot put	.771	−.434	.197	−.112	.17	.530	.777	−.141	−.079	.09
4. High jump	.711	.181	.005	.367	.33	.363	.428	.421	.424	.33
5. 400-m run	.605	.549	−.045	−.397	.17	.571	.019	.620	−.305	.20
6. 100 m hurdles	.513	−.083	−.372	.561	.28	.343	.189	.090	.323	.73
7. Discus	.690	−.456	.289	−.078	.23	.402	.718	−.102	−.095	.30
8. Pole vault	.761	.162	.018	.304	.30	.440	.407	.390	.263	.42
9. Javelin	.518	−.252	.519	−.074	.39	.218	.461	.084	−.085	.73
10. 1500-m run	.220	.746	.493	.085	.15	−.016	.091	.609	−.145	.60
Cumulative proportion of total variance explained	.42	.56	.67	.76		.27	.45	.57	.62	

Hệ số tải sau khi thực hiện quay nhân tố:

	Principal component					Maximum likelihood				
	Estimated rotated factor loadings, $\tilde{\ell}^*_{ij}$				Specific variances	Estimated rotated factor loadings, $\hat{\ell}^*_{ij}$				Specific variances
Variable	F_1^*	F_2^*	F_3^*	F_4^*	$\tilde{\psi}_i = 1 - \tilde{h}_i^2$	F_1^*	F_2^*	F_3^*	F_4^*	$\hat{\psi}_i = 1 - \hat{h}_i^2$
100-m run	.182	.885	.205	−.139	.12	.204	.296	.928	−.005	.01
Long jump	.291	.664	.429	.055	.29	.280	.554	.451	.155	.39
Shot put	.819	.302	.252	−.097	.17	.883	.278	.228	−.045	.09
High jump	.267	.221	.683	.293	.33	.254	.739	.057	.242	.33
400-m run	.086	.747	.068	.507	.17	.142	.151	.519	.700	.20
110-m hurdles	.048	.108	.826	−.161	.28	.136	.465	.173	−.033	.73
Discus	.832	.185	.204	−.076	.23	.793	.220	.133	−.009	.30
Pole vault	.324	.278	.656	.293	.30	.314	.613	.169	.279	.42
Javelin	.754	.024	.054	.188	.39	.477	.160	.041	.139	.73
1500-m run	−.002	.019	.075	.921	.15	.001	.110	−.070	.619	.60
Cumulative proportion of total sample variance explained	.22	.43	.62	.76		.20	.37	.51	.62	

Nhận xét:

- Ta thấy rằng đẩy tạ, ném đĩa và phóng lao phụ thuộc rất nhiều vào một yếu tố, được gọi là "Sức mạnh bùng nổ của cánh tay".
- Tương tự như vậy, nhảy cao, chạy vượt rào 110 mét, nhảy sào và ở một mức độ nào đó nhảy xa phụ thuộc rất nhiều vào một yếu tố khác, gọi yếu tố này là "Sức mạnh bùng nổ của đôi chân".
- Chạy 100 mét, 400 mét, và ở một mức độ nào đó nhảy xa phụ thuộc rất nhiều vào yếu tố thứ ba, yếu tố này có thể được gọi là "Tốc độ chạy".
- Cuối cùng, chạy 1500 mét và chạy 400 mét tải nặng vào yếu tố thứ tư, gọi yếu tố này là "Sức bền khi chạy".
- Biểu đồ tải trọng khả năng tối đa quay vòng cho các cặp nhân tố (1,2) và (1,3):



- Các điểm thường được nhóm dọc theo các trục nhân tố. Đồ thị tải trọng thành phần chính quay rất giống nhau.

4.5 Phương pháp quay xiên.

"Oblimin trực tiếp" Là thủ tục quay 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. ζ là một tham số kiểm soát mức độ tương quan giữa các yếu tố.

- Các phép quay 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 quay 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 quay trực giao đến một cấu trúc đơn giản tương ứng với một phép quay cứng nhắc của các trục tọa độ sao cho các trục, sau khi quay, đi càng gần các cụm càng tốt.
- Một phép quay 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 quay 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.

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.

Trong phân tích nhân tố, mối quan tâm thường tập trung vào các tham số trong mô hình nhân tố. Tuy nhiên, giá trị ước tính của các yếu tố chung cũng được quan tâm và được gọi là điểm nhân tố. Những đại lượng này thường được sử dụng cho mục đích chẩn đoán, cũng như đầu vào cho một 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 các ước lượng giá trị cho các vectơ yếu tố ngẫu nhiên không được quan sát $F_j, j = 1, \ldots, n$.

Ta kí hiệu điểm nhân tố $\hat{f}_j$ = ước lượng giá trị f_j đạt được tại F_j Tiếp theo là hai phương pháp được sử dụng để tính toán điểm nhân tố. Điểm chung của hai phương pháp là:

1. Coi giá trị của các nhân tố tải $\hat{l}_{ij}$ và phương sai riêng $\hat{\Psi}$ là các giá trị thực.
2. Các phương pháp đều liên quan đến các phép biên đổi tuyến tính từ tập dữ liệu gốc, có thể là trung tâm hoá hoặc tiêu chuẩn hoá. Các nhân tố tải đã được quay sẽ được sử dụng để tính toán thay vì các nhân tố tải ước lượng ban đầu.

5.2 Phương pháp bình phương tối thiểu có trọng số.

5.2.1 Phương pháp bình phương tối thiểu.

Giả sử dữ liệu gồm các điểm (x_i, y_i) với $i = 1, 2, ..., n$. Chúng ta cần tìm một hàm số f thỏa mãn:

$$f(x_i) \approx y_i$$

Giả sử hàm f có thể thay đổi hình dạng, phụ thuộc vào một số tham số p_j với $j = 1, 2, ..., m$.

$$f(x) = f(p_j, x)$$

Ta cần tìm giá trị của các tham số p_j sao cho biểu thức sau đạt cực tiểu:

$$\chi^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

5.2.2 Phương pháp bình phương tối thiểu có trọng số để tính điểm nhân tố.

Ta xét một mô hình nhân tố đã được đề cập ở các chương trước:

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

Trong đó:

- μ là *vector* trung bình.
- L là nhân tố tải.
- ψ là phương sai riêng.
- $\varepsilon = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_p]$ là sai số.

Do $Var(\varepsilon_i) = \psi_i, i = 1,2,...,p$ không bằng nhau nên phương pháp bình phương tối thiểu có trọng số được sử dụng để ước lượng giá trị yếu tố chung.

Tổng bình phương các sai số và được tính trọng số bằng nghịch đảo các phương sai là

$$\sum_{i=1}^{p} \frac{\varepsilon_i^2}{\psi_i} = \varepsilon' \Psi^{-1} \varepsilon \tag{5.1}$$

Từ phương trình (5.1) ta được:

$$S = \sum_{i=1}^{p} \frac{\varepsilon_i^2}{\psi_i} = (x - \mu - Lf)' \psi^{-1} (x - \mu - Lf) \tag{5.2}$$

Lấy đạo hàm theo f từ (5.2):

$$\frac{\partial S}{\partial f} = -2(x - \mu - Lf)' \psi^{-1} L = 0$$

Vậy chọn ước lượng $\hat{f}$ của f để tối thiểu hoá phương trình ta thu được:

$$\hat{f} = (L' \psi^{-1} L)^{-1} L' \psi^{-1} (x - \mu) \tag{5.3}$$

Từ đó, ta chọn các ước lượng $\hat{L}, \hat{\psi}$ và $\hat{\mu} = \bar{x}$ là giá trị thực và thu được điểm nhân tố cho trường hợp thứ j:

$$\hat{f}_j = (\hat{L}' \hat{\psi}^{-1} \hat{L})^{-1} \hat{L}' \hat{\psi}^{-1} (x_j - \bar{x}) \tag{5.4}$$

Trong đó $\bar{x}$ là trung bình của mẫu.

5.2.3 Điểm nhân tố thu được bằng bình phương tối thiểu có trọng số từ ước lượng hợp lý cực đại.

Khi $\hat{L}$ và $\hat{\psi}$ được xác định bởi phương pháp hợp lý cực đại, các ước lượng phải thoả mãn điều kiện duy nhất, $\hat{L}' p\hat{s}i^{-1} \hat{L} = \hat{\Delta}$ là ma trận đường chéo. Từ đó ta có:

$$\begin{aligned} \hat{f}_j &= (\hat{L}' \hat{\psi}^{-1} \hat{L})^{-1} \hat{L}' \hat{\psi}^{-1} (x_j - \overline{\mu}) \\ &= \hat{\Delta}^{-1} \hat{L}' \hat{\psi}^{-1} (x_j - \bar{x}), j = 1,2,..,n \end{aligned} \tag{5.5}$$

Hoặc nếu được xác định bởi phương pháp thành phần chính:

$$\begin{aligned}\hat{f}_j &= (\hat{L}_z'\hat{\psi}_z^{-1}\hat{L}_z)^{-1}\hat{L}_z'\hat{\psi}_z^{-1}z_j \\ &= \hat{\Delta}_z^{-1}\hat{L}_z'\hat{\psi}_z^{-1}z_j, j = 1,2,..,n\end{aligned} \tag{5.6}$$

Trong đó $z_j = D^{-1/2}(x_j - \bar{x})$ và $\hat{\rho} = \hat{L}_z\hat{L}_z' + \hat{\psi}_z$. Điểm nhân tố ở hai công thức trên có hiệp phương sai mẫu bằng 0 và vectorr trung bình mẫu bằng 0.

Nhận xét

Nếu các nhân tố tải được ước tính bằng phương pháp thành phần chính, thông thường sẽ tạo điểm nhân tố bằng cách sử dụng bình phương nhỏ nhất không trọng số. Từ đó công thức tính điểm nhân tố là:

$$\hat{f}_j = (\tilde{L}'\tilde{L})^{-1}\tilde{L}'(x_j - \bar{x}) \tag{5.7}$$

Hoặc:

$$\hat{f}_j = (\tilde{L}_z'\tilde{L}_z)^{-1}\tilde{L}_z'z_j \tag{5.8}$$

Để tiêu chuẩn hoá dữ liệu.

Do $\tilde{L} = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1 \vdots \sqrt{\hat{\lambda}_2}\hat{e}_2 \vdots \ldots \vdots \sqrt{\hat{\lambda}_m}\hat{e}_m\right]$, ta có:

$$\hat{f}_j = \begin{bmatrix} \frac{1}{\sqrt{\hat{\lambda}_1}}\hat{e}_1'(x_j - \bar{x}) \\ \frac{1}{\sqrt{\hat{\lambda}_2}}\hat{e}_2'(x_j - \bar{x}) \\ \vdots \\ \frac{1}{\sqrt{\hat{\lambda}_m}}\hat{e}_m'(x_j - \bar{x}) \end{bmatrix} \tag{5.9}$$

Trong đó λ_i và e_i lần lượt là trị riêng và vector riêng của ma trận hiệp phương sau mẫu ban đầu. Ta có trung bình mẫu $\frac{1}{n}\sum_{j=1}^{n}\hat{f}_j = 0$ và hiệp phương sai mẫu $\frac{1}{n-1}\sum_{j=1}^{n}\hat{f}_j\hat{f}_j' = \mathbf{I}$.

5.3 Phương pháp hồi quy.

Ta xét một mô hình nhân tố đã được đề cập ở các chương trước:

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

Với ma trận tải L và ma trận phương sai riêng ψ đã biết.

Ta có nhân tố chung F và nhân tố riêng (hoặc sai số) ε đồng thời là phân phối chuẩn với trung bình và phương sai như sau:

$$E(F) = \underset{(m\times 1)}{0}, \qquad Cov(F) = E[FF'] = \underset{(m\times m)}{I}$$

$$E(\varepsilon) = \underset{(p\times 1)}{0}, \qquad Cov(\varepsilon) = E[\varepsilon\varepsilon'] = \underset{(p\times p)}{\psi} = \begin{bmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \dots & \dots & \ddots & \dots \\ 0 & 0 & \dots & \psi_p \end{bmatrix}$$

Định lý 5.1

Nếu X tuân theo phân phối $N_p(\mu,\Sigma)$, thì q tổ hợp tuyến tính:

$$\underset{(q\times p)}{A}\underset{(p\times 1)}{X} = \begin{bmatrix} a_{11}X_1 + \dots + a_{1p}X_p \\ a_{21}X_1 + \dots + a_{2p}X_p \\ \vdots \\ a_{q1}X_1 + \dots + a_{qp}X_p \end{bmatrix}$$ tuân theo phân phối chuẩn $N_q(A\mu, A\Sigma A')$.

Đồng thời, $\underset{(p\times 1)}{X} + \underset{(p\times 1)}{d}$ tuân theo phân phối $N_p(\mu + d,\Sigma)$ với d là vector của hằng số.

Khi đó tổ hợp tuyến tính $X-\mu = LF+\varepsilon$ có phân phối $N_p(0, LL'+\psi)$. Thêm nữa, phân phối đồng thời của $(X-\mu)$ và F là $N_{m+p}(0,\Sigma^*)$, trong đó:

$$\underset{(m+p)\times(m+p)}{\Sigma^*} = \begin{bmatrix} \underset{p\times p}{\Sigma = LL'+\psi} & \underset{p\times m}{L} \\ \underset{m\times p}{L'} & \underset{m\times m}{I} \end{bmatrix} \tag{5.10}$$

Và 0 là vector không kích thước $(m+p)\times 1$.

Định lý 5.2

Cho $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ tuân theo phân phối $N_p(\mu,\Sigma)$, trong đó $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ và $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ và $|\Sigma_{22}| > 0$. Khi đó phân phối chuẩn nhiều chiều X_1 điều kiện $X_2 = x_2$ với:

$$E(X_1|X_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)$$

$$Cov(X_1|X_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Theo định lý ta được phân phối chuẩn nhiều chiều của F|x với:

$$mean = E(F|x) = L'\Sigma^{-1}(x-\mu) = L'(LL'+\psi)^{-1}(x-\mu) \tag{5.11}$$

Và:

$$covariance = Cov(F|x) = I - L'\Sigma^{-1}L = I - L'(LL'+\psi)^{-1}L \tag{5.12}$$

Các đại lượng $L'(LL'+\psi)^{-1}$ là các hệ số trong một hồi quy (đa biến) của các nhân tố trên biến. Ước lượng của các hệ số này tạo ra điểm nhân tố tương tự với ước lượng của giá trị trung bình có điều kiện trong phân tích hồi quy đa biến.

Từ đó, với bất kỳ vector quan sát x_j và coi ước lượng hợp lý cực đại $\hat{L}$ và $\hat{\psi}$ là giá trị thực, ta được điểm nhân tố thứ j được cho bởi:

$$\hat{f}_j = \hat{L}'\Sigma^{-1}(x_j-\hat{x}) = \hat{L}'(\hat{L}\hat{L}'+\hat{\psi})^{-1}(x_j-\hat{x}), j = 1,2,...,n \tag{5.13}$$

Việc tính toán $\hat{f}_j$ có thể làm đơn giản hoá bằng việc sử dụng ma trận đơn vị:

$$\underset{m\times p}{\hat{L}'}\underset{p\times p}{(\hat{L}\hat{L}'+\hat{\psi})^{-1}} = \underset{m\times m}{(I+\hat{L}'\psi^{-1}\hat{L})^{-1}}\underset{m\times p}{\hat{L}'}\underset{p\times p}{\hat{\psi}^{-1}} \tag{5.14}$$

Từ phương trình (5.5), (5.6), (5.13) và (5.14) cho phép ta so sánh điểm nhân tố từ phương pháp hồi quy là $\hat{f}_j^R$ và phương pháp bình phương tối thiểu có trọng s là $\hat{f}_j^{LS}$. Từ phương trình trên ta được:

$$\hat{f}_j^{LS} = (\hat{L}'\hat{\psi}^{-1}\hat{L})^{-1}(I+\hat{L}'\hat{\psi}^{-1}\hat{L})\hat{f}_j^R = (I+(\hat{L}'\hat{\psi}^{-1}\hat{L})^{-1})\hat{f}_j^R \tag{5.15}$$

Nhận xét

- Với ước lượng hợp lý cực đại $(\hat{L}'\hat{\psi}^{-1}\hat{L})^{-1} = \hat{\Delta}^{-1}$ và phần tử của ma trận đường chéo gần bằng 0 thì 2 phương pháp kết quả gần như nhau.
- Trong nỗ lực giảm bớt ảnh hưởng (nếu có) của việc xác định số lượng các nhân tố, ta tính điểm nhân tố bằng việc sử dụng S (ma trận phương sai mẫu ban đầu) thay vì $\hat{\Sigma} = \hat{L}\hat{L}' + \hat{\psi}$. Từ đó ta có:

  $$\hat{f}_j = \hat{L}'S^{-1}(x_j - \bar{x}), j = 1,...,n \tag{5.16}$$

  Hoặc nếu được xác định bởi phương pháp thành phần chính:

  $$\hat{f}_j = \hat{L}_z'R^{-1}z_j, j = 1,...,n \tag{5.17}$$

  Trong đó $z_j = D^{-1/2}(x_j - \bar{x})$ và $\hat{\rho} = \hat{L}_z\hat{L}_z' + \hat{\psi}_z$.
- Hệ số tương quan giữa các điểm nhân tố tạo ra từ hai phương pháp được cung cấp bởi hệ số tương quan giữa các điểm trên cùng một nhân tố, không có phương pháp nào được khuyến nghị là vượt trội hẳn.

5.4 Ví dụ về tính toán điểm nhân tố.

Ta sẽ minh hoạ việc tính toán điểm nhân tố bằng hai phương pháp bình phương tối thiểu và hồi quy từ dữ liệu về giá cổ phiếu ở ví dụ ở chương nhân tố quay.

Nhắc lại hai công thức ứng với từng phương pháp:

Phương pháp bình phương tối thiểu có trọng số

$$\hat{f} = (\hat{L^*}'_z \hat{\psi}_z^{-1} \hat{L}_z^*)^{-1} \hat{L}_z^{*\prime} \hat{\psi}_z^{-1} z$$

Phương pháp hồi quy

$$\hat{f} = \hat{L^*}'_z R^{-1} z$$

Sử dụng phần mềm lập trình và phương pháp ước lượng hợp lý cực đại cho ta hệ số tải quay $\hat{L}_z^*$ và phương sai riêng $\hat{\psi}_z$:

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

Và ma trận phương sai mẫu ban đầu:

$$R = \begin{bmatrix} 1.000 & .632 & .511 & .115 & .155 \\ .632 & 1.000 & .574 & .322 & .213 \\ .511 & .574 & 1.000 & .183 & .146 \\ .115 & .322 & .183 & 1.000 & .683 \\ .155 & .213 & .146 & .183 & 1.000 \end{bmatrix}$$

Vectơ của quan sát chuẩn hóa $z' = [.50, -1.40, -.20, -.70, 1.40]$.

Từ đó cho chúng ta điểm số của các nhân tố 1 và 2 như sau:

- Sử dụng phương pháp bình phương tối thiểu:

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

- Sử dụng phương pháp hồi quy:

$$\hat{f} = \hat{L}_z^{*'} R^{-1} z = \begin{bmatrix} .331 & .526 & .221 & -.137 & .011 \\ -.040 & -.063 & -.026 & 1.023 & -.001 \end{bmatrix} \begin{bmatrix} .50 \\ -1.40 \\ -.20 \\ -.70 \\ 1.40 \end{bmatrix} = \begin{bmatrix} -.50 \\ -.64 \end{bmatrix}$$

Nhận xét

Trong bài toán này, kết quả từ hai phương pháp khá tương đồng nhau.

CHƯƠNG 6

ỨNG DỤNG THỰC TẾ.

6.1 Bộ dữ liệu 1.

6.1.1 Mô tả bài toán.

Mô tả dữ liệu

Bộ dữ liệu chứa một cuộc khảo sát về sự hài lòng của hành khách hàng không. Nó có 129487 quan sát và 15 cột, trong đó 15 cột thể hiện phản hồi của khách hàng, theo thang điểm từ 1 đến 5, đối với một cuộc khảo sát đánh giá các khía cạnh khác nhau của chuyến bay:

- Dịch vụ wifi trên máy bay.
- Đồ ăn và đồ uống.
- Lên máy bay trực tuyến.
- Chỗ ngồi thoải mái.
- ...

15 biến đó như sau:

1. Inflight wifi service.
2. Departure/Arrival time convenient.
3. Ease of Online booking.
4. Gate location.
5. Food and drink.
6. Online boarding.
7. Seat comfort.
8. Inflight entertainment.
9. On - board service.
10. Leg room service.
11. Baggage handling.
12. Checkin service.
13. Inflight service.
14. Cleanliness.
15. Checkin service.

6.1.2 Chạy chương trình.

Bước 1: Nhập dữ liệu.

	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes
0	3	4	3	1	5	3	5	5	4	3	4	4	5	5	25	18
1	3	2	3	3	1	3	1	1	1	5	3	1	4	1	1	6
2	2	2	2	2	5	5	5	5	4	3	4	4	4	5	0	0
3	2	5	5	5	2	2	2	2	2	5	3	1	4	2	11	9
4	3	3	3	3	4	5	5	3	3	4	4	3	3	3	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
129482	3	3	3	1	4	3	4	4	3	2	4	4	5	4	0	0
129483	4	4	4	4	4	4	4	4	4	5	5	5	5	4	0	0
129484	2	5	1	5	2	1	2	2	4	3	4	5	4	2	0	0
129485	3	3	3	3	4	4	4	4	3	2	5	4	5	4	0	0
129486	2	5	2	5	4	2	2	1	1	2	1	1	1	1	0	0

129487 rows × 16 columns

Chỉ giữ lại 15 biến sẽ sử dụng trong quá trình phân tích, tất cả các cột không liên quan ta sẽ lược bỏ. Đồng thời loại bỏ các dòng có dữ liệu bị thiếu.

Bước 2:

- Mô tả dữ liệu.

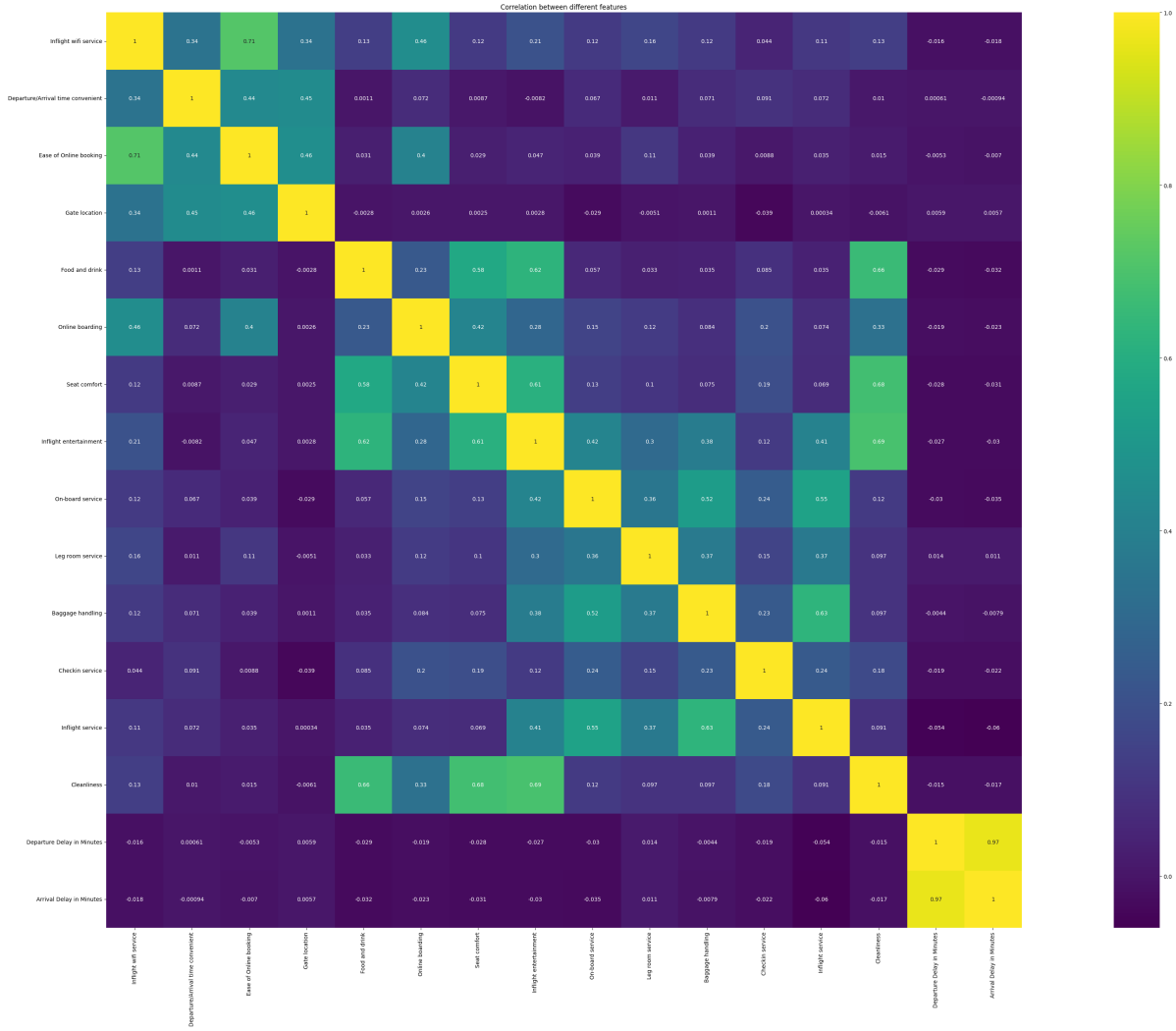
	count	mean	std	min	25%	50%	75%	max
Inflight wifi service	129487.0	2.728544	1.329235	0.0	2.0	3.0	4.0	5.0
Departure/Arrival time convenient	129487.0	3.057349	1.526787	0.0	2.0	3.0	4.0	5.0
Ease of Online booking	129487.0	2.756786	1.401662	0.0	2.0	3.0	4.0	5.0
Gate location	129487.0	2.976909	1.278506	0.0	2.0	3.0	4.0	5.0
Food and drink	129487.0	3.204685	1.329905	0.0	2.0	3.0	4.0	5.0
Online boarding	129487.0	3.252720	1.350651	0.0	2.0	3.0	4.0	5.0
Seat comfort	129487.0	3.441589	1.319168	0.0	2.0	4.0	5.0	5.0
Inflight entertainment	129487.0	3.358067	1.334149	0.0	2.0	4.0	4.0	5.0
On-board service	129487.0	3.383204	1.287032	0.0	2.0	4.0	4.0	5.0
Leg room service	129487.0	3.351078	1.316132	0.0	2.0	4.0	4.0	5.0
Baggage handling	129487.0	3.631886	1.180082	1.0	3.0	4.0	5.0	5.0
Checkin service	129487.0	3.306239	1.266146	0.0	3.0	3.0	4.0	5.0
Inflight service	129487.0	3.642373	1.176614	0.0	3.0	4.0	5.0	5.0
Cleanliness	129487.0	3.286222	1.313624	0.0	2.0	3.0	4.0	5.0
Departure Delay in Minutes	129487.0	14.643385	37.932867	0.0	0.0	0.0	12.0	1592.0
Arrival Delay in Minutes	129487.0	15.091129	38.465650	0.0	0.0	0.0	13.0	1584.0

- Tất cả các dữ liệu đều đã đủ và phù hợp. Nhưng có lưu ý biến "Arrival Delay in Minutes"có dạng float.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 129487 entries, 0 to 129486
Data columns (total 16 columns):
 #   Column                            Non-Null Count   Dtype
---  ------                            --------------   -----
 0   Inflight wifi service             129487 non-null  int64
 1   Departure/Arrival time convenient 129487 non-null  int64
 2   Ease of Online booking            129487 non-null  int64
 3   Gate location                     129487 non-null  int64
 4   Food and drink                    129487 non-null  int64
 5   Online boarding                   129487 non-null  int64
 6   Seat comfort                      129487 non-null  int64
 7   Inflight entertainment            129487 non-null  int64
 8   On-board service                  129487 non-null  int64
 9   Leg room service                  129487 non-null  int64
 10  Baggage handling                  129487 non-null  int64
 11  Checkin service                   129487 non-null  int64
 12  Inflight service                  129487 non-null  int64
 13  Cleanliness                       129487 non-null  int64
 14  Departure Delay in Minutes        129487 non-null  int64
 15  Arrival Delay in Minutes          129487 non-null  int64
dtypes: int64(16)
memory usage: 15.8 MB
```

Bước 3: Ma trận hệ số tương quan.

- Giờ ta sẽ xem xét biểu đồ tương quan của tất cả các biến để xem liệu có biến nào vô dụng hoặc quá tương quan với các biến khác hay không.



- Một số biến có mối tương quan khá cao, đặc biệt là những biến liên quan đến câu trả lời khảo sát. Mối tương quan giữa "Departure Delay in Minutes" và "Arrival Delay in Minutes" cực kỳ cao (0,98). Điều đó có ý nghĩa: Nếu máy bay khởi hành muộn hơn dự kiến thì nó cũng sẽ đến muộn hơn. Xem xét mối tương quan cực kỳ cao này, chúng ta quyết định chỉ xóa cột đó khỏi tập dữ liệu.

Bước 4: Kiểm tra một vài giả định.

- Kiểm định cầu Bartlett thu được kết quả như sau:

$$\begin{bmatrix} 752622.6835461138 & 0.0 \end{bmatrix}$$

Kiểm tra xem các biến có tương quan với nhau hay không bằng cách so sánh ma trận tương quan với ma trận đơn vị. Chỉ số sig < 0.05 chứng tỏ các biến có tương quan với nhau.

- Kiểm định KMO thu được kết quả là:

$$0.7810578543966449$$

Kiểm định KMO đo mức độ phù hợp của bộ dữ liệu để phân tích nhân tố, ước tính tỷ lệ phương sai giữa các biến. Chỉ số KMO > 0.6 được coi là phù hợp.

$\Longrightarrow$ Cả kiểm định cầu Bartlett và KMO đều chỉ ra rằng khung dữ liệu phù hợp để phân tích nhân tố.

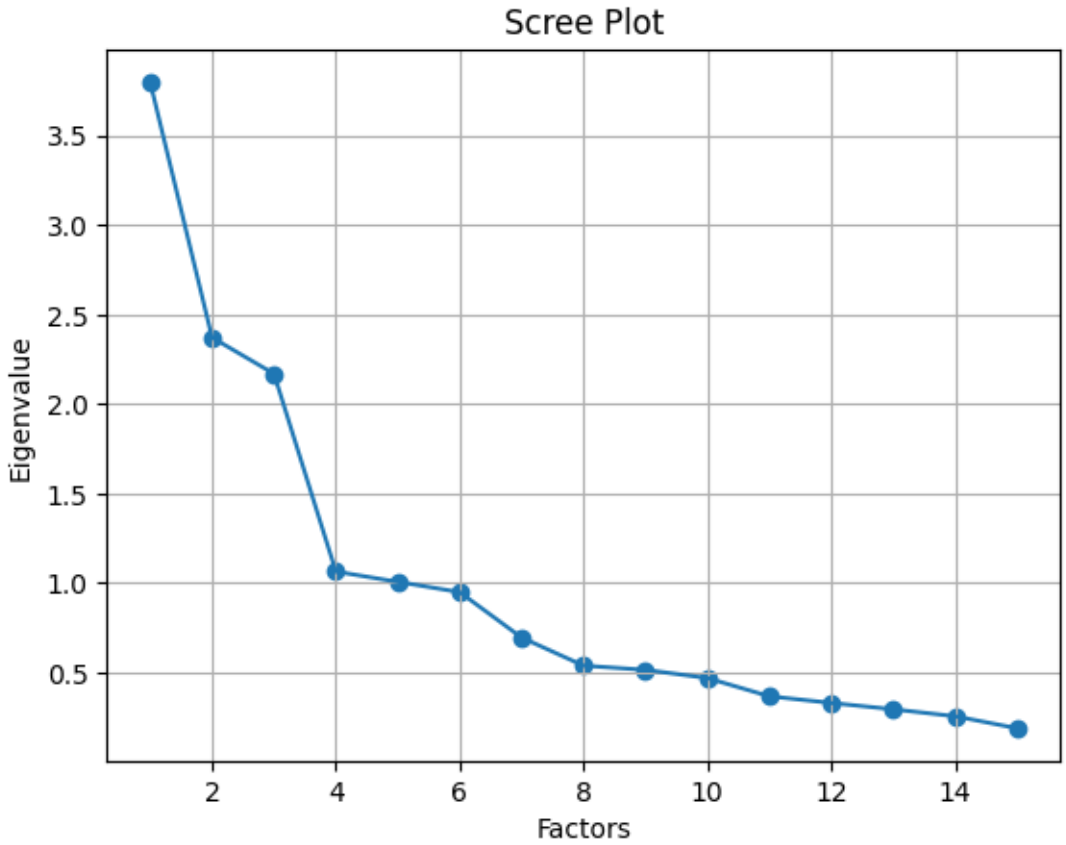
Bước 5: Xác định số lượng các nhân tố.

- Nếu không biết cần trích xuất bao nhiêu yếu tố trong phân tích, trước tiên ta có thể sử dụng phương pháp trích xuất thành phần chính, không xoay vòng, sử dụng số lượng yếu tố mặc định làm đánh giá sơ bộ.
- Để tìm ra chúng ta cần bao nhiêu yếu tố, chúng ta có thể xem xét giá trị riêng, là thước đo mức độ phương sai của các biến mà một yếu tố giải thích. Trong phương pháp trích xuất thành phần chính, phương sai của từng yếu tố sẽ bằng giá trị riêng của nó.
- Giá trị riêng biểu thị tổng lượng phương sai có thể được giải thích bằng một thành phần chính nhất định.

- Giá trị riêng > 1 có nghĩa là yếu tố giải thích nhiều phương sai hơn một biến duy nhất. Bởi vậy, ta sẽ lấy các giá trị riêng > 1.
- Ta thu được kết quả về giá trị riêng như sau:

$$\begin{bmatrix} 3.79955647 & 2.37147549 & 2.16897694 & 1.06313402 & 1.00576476 \\ 0.94813865 & 0.69417644 & 0.5372157 & 0.51374933 & 0.46777892 \\ 0.36604041 & 0.32883888 & 0.29328704 & 0.25443485 & 0.18743211 \end{bmatrix}$$

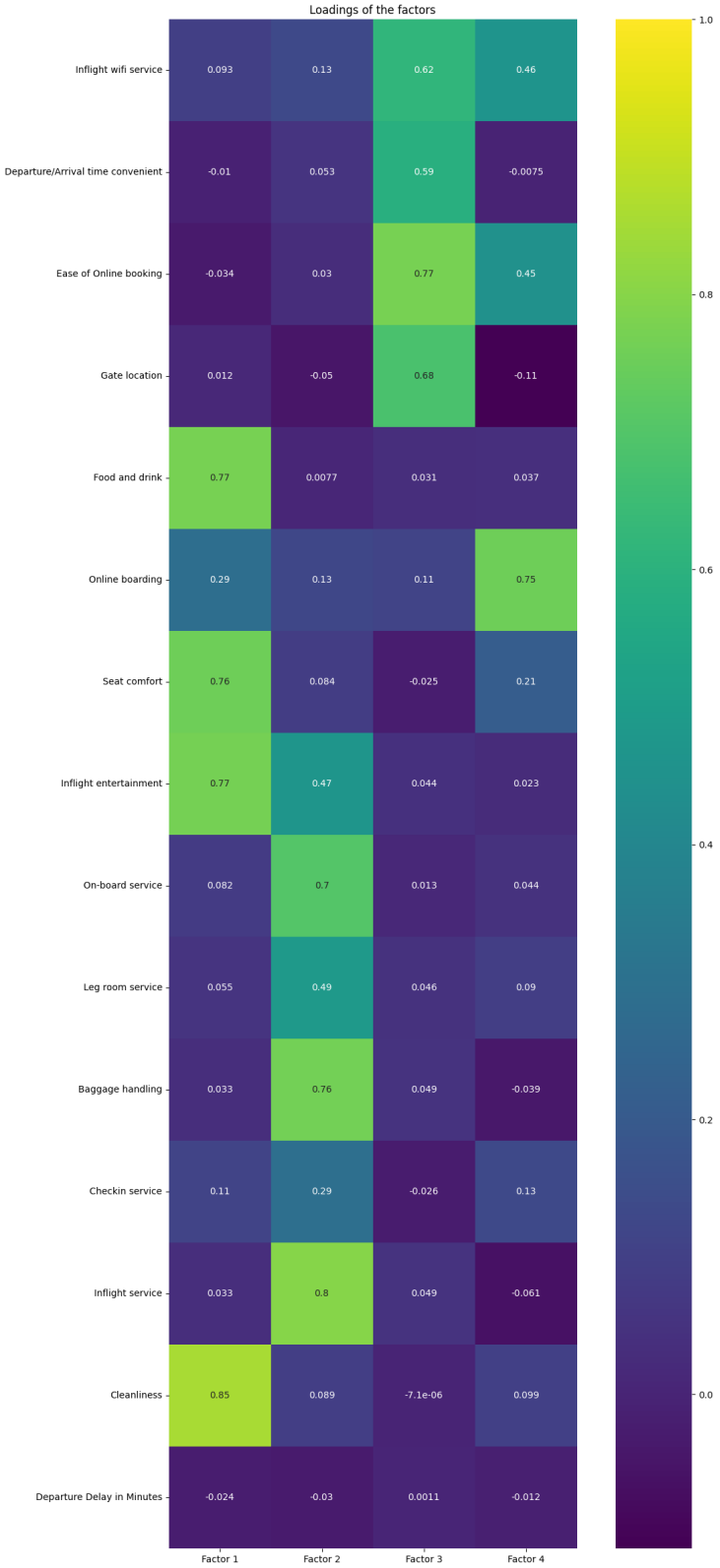
- Biểu đồ Scatter:



- Tiếp theo ta sẽ lấy ra bốn nhân tố > 1. Bây giờ ta đã xác định được số nhân tố, ta thực hiện phép quay Varimax để tìm ma trận tải trọng được biểu diễn qua các nhân tố này. Phép quay Varimax giúp cải thiện tính diễn giải và hiểu rõ hơn về mối quan hệ giữa các yếu tố.

- Ma trận tải trọng:

	Factor 1	*Factor 2*	*Factor 3*	*Factor 4*
$Inflightwifiservice$	0.093052	0.134820	0.615484	0.463961
$Departure/Arrivaltimeconvenient$	−0.010308	0.053033	0.589641	−0.007537
$EaseofOnlinebooking$	−0.034082	0.029731	0.773800	0.447062
$Gatelocation$	0.012412	−0.050015	0.682242	−0.112097
$Foodanddrink$	0.770890	0.007706	0.031326	0.036565
$Onlineboarding$	0.287200	0.126517	0.110047	0.754456
$Seatcomfort$	0.755716	0.083901	−0.025206	0.210944
$Inflightentertainment$	0.765443	0.469186	0.043598	0.022827
$On-boardservice$	0.082294	0.702358	0.013276	0.044409
$Legroomservice$	0.055267	0.485487	0.045599	0.089881
$Baggagehandling$	0.033114	0.763634	0.049389	−0.038551
$Checkinservice$	0.111999	0.289186	−0.025706	0.132541
$Inflightservice$	0.032951	0.799812	0.049316	−0.060815
$Cleanliness$	0.853401	0.088800	−0.000007	0.099339
$DepartureDelayinMinutes$	−0.023606	−0.029636	0.001088	−0.012298


Loadings of the factors
Inflight wifi service
Departure/Arrival time convenient
Ease of Online booking
Gate location
Food and drink
Online boarding
Seat comfort
Inflight entertainment
On-board service
Leg room service
Baggage handling
Checkin service
Inflight service
Cleanliness
Departure Delay in Minutes
0.093 0.13 0.62 0.46
-0.01 0.053 0.59 -0.0075
-0.034 0.03 0.77 0.45
0.012 -0.05 0.68 -0.11
0.77 0.0077 0.031 0.037
0.29 0.13 0.11 0.75
0.76 0.084 -0.025 0.21
0.77 0.47 0.044 0.023
0.082 0.7 0.013 0.044
0.055 0.49 0.046 0.09
0.033 0.76 0.049 -0.039
0.11 0.29 -0.026 0.13
0.033 0.8 0.049 -0.061
0.85 0.089 -7.1e-06 0.099
-0.024 -0.03 0.0011 -0.012
Factor 1
Factor 2
Factor 3
Factor 4
1.0
0.8
0.6
0.4
0.2
0.0

- Yếu tố "Departure Delay in Minutes"được thể hiện ít nhất thông qua 4 nhân tố này. Yếu tố "Inflight entertainment"thể hiện được nhiều nhất. Ta sẽ chọn ra các biến có tải trọng cao đối với mỗi nhân tố (> 0.5). Các biến này sẽ được giải thích thông qua mỗi nhân tố tương ứng. Ta chia ra được các biến quan sát phụ thuộc vào 4 nhân tố như sau:

  1. Comfort: Cleanliness, Food and Drink, Inflight Entertainment, Seat Comfort.
  2. Service: Inflight Services, Baggage Handling, Onboard Services, Leg Room.
  3. Convenience: Ease of Online Booking, Gate Location.
  4. Network: Online Boarding, Inflight Wifi Service.

Bước 6: Tính cộng đồng, ma trận dư.

- Tính cộng đồng (còn gọi là phương sai chung h^2) là tỷ lệ biến thiên của mỗi biến được giải thích bởi các nhân tố, là tổng phương sai được tính bởi nhân tố được chọn.
- Các giá trị gần bằng 1 cho thấy các yếu tố được trích xuất giải thích nhiều hơn về phương sai của một mục riêng lẻ.
- Các biến có tính cộng đồng thấp. Ví dụ như thấp hơn 0.4 không đóng góp nhiều vào việc đo lường các yếu tố cơ bản.
- Tỷ lệ phương sai của mỗi biến được tính theo các nhân tố được chọn:

$$\begin{bmatrix} 0.62091568 & 0.35065235 & 0.80067601 & 0.48067551 & 0.59664905 \\ 0.67980446 & 0.62327853 & 0.80846041 & 0.50222686 & 0.24891006 \\ 0.58815897 & 0.11440029 & 0.64691507 & 0.74604774 & 0.00158795 \end{bmatrix}$$

$$\begin{bmatrix} & Communalities \\ Inflightwifiservice & 0.620916 \\ Departure/Arrivaltimeconvenient & 0.350652 \\ EaseofOnlinebooking & 0.800676 \\ Gatelocation & 0.480676 \\ Foodanddrink & 0.596649 \\ Onlineboarding & 0.679804 \\ Seatcomfort & 0.623279 \\ Inflightentertainment & 0.808460 \\ On-boardservice & 0.502227 \\ Legroomservice & 0.248910 \\ Baggagehandling & 0.588159 \\ Checkinservice & 0.114400 \\ Inflightservice & 0.646915 \\ Cleanliness & 0.746048 \\ DepartureDelayinMinutes & 0.001588 \end{bmatrix}$$

- Ma trận dư PSI:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0.379084	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.0000	0.000000	0.000000	0.000000
1	0.000000	0.649348	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.0000	0.000000	0.000000	0.000000
2	0.000000	0.000000	0.199324	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.0000	0.000000	0.000000	0.000000
3	0.000000	0.000000	0.000000	0.519324	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.0000	0.000000	0.000000	0.000000
4	0.000000	0.000000	0.000000	0.000000	0.403351	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.0000	0.000000	0.000000	0.000000
5	0.000000	0.000000	0.000000	0.000000	0.000000	0.320196	0.000000	0.00000	0.000000	0.00000	0.000000	0.0000	0.000000	0.000000	0.000000
6	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.376721	0.00000	0.000000	0.00000	0.000000	0.0000	0.000000	0.000000	0.000000
7	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.19154	0.000000	0.00000	0.000000	0.0000	0.000000	0.000000	0.000000
8	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.497773	0.00000	0.000000	0.0000	0.000000	0.000000	0.000000
9	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.75109	0.000000	0.0000	0.000000	0.000000	0.000000
10	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.411841	0.0000	0.000000	0.000000	0.000000
11	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.8856	0.000000	0.000000	0.000000
12	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.0000	0.353085	0.000000	0.000000
13	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.0000	0.000000	0.253952	0.000000
14	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.0000	0.000000	0.000000	0.998412

- Ma trận tương quan phù hợp:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1.000000	0.365609	0.684518	0.362313	0.109018	0.461552	0.163988	0.171907	0.131125	0.140363	0.118546	0.095082	0.113034	0.137468	-0.011228
1	0.365609	1.000000	0.454823	0.400342	0.010658	0.062951	-0.019793	0.042527	0.043893	0.051387	0.069569	-0.001974	0.071614	-0.004840	-0.000594
2	0.684518	0.454823	1.000000	0.475894	0.014542	0.416416	0.051538	0.031802	0.048204	0.088017	0.042557	0.044143	0.033629	0.017959	-0.004732
3	0.362313	0.400342	0.475894	1.000000	0.026456	-0.012257	-0.035659	0.013220	-0.030028	-0.002561	0.000234	-0.045469	0.000869	-0.004989	0.003310
4	0.109018	0.010658	0.014542	0.026456	1.000000	0.253409	0.590144	0.595889	0.070892	0.051061	0.031550	0.092608	0.030886	0.662195	-0.018842
5	0.461552	0.062951	0.416416	-0.012257	0.253409	1.000000	0.384031	0.301215	0.147461	0.150124	0.082473	0.165921	0.070198	0.331278	-0.019688
6	0.163988	-0.019793	0.051538	-0.035659	0.590144	0.384031	1.000000	0.621539	0.130153	0.100309	0.079717	0.137509	0.077934	0.673335	-0.022948
7	0.171907	0.042527	0.031802	0.013220	0.595889	0.301215	0.621539	1.000000	0.394120	0.274127	0.384906	0.223315	0.401244	0.697162	-0.032207
8	0.131125	0.043893	0.048204	-0.030028	0.070892	0.147461	0.130153	0.394120	1.000000	0.350131	0.538013	0.217874	0.562419	0.137011	-0.023289
9	0.140363	0.051387	0.088017	-0.002561	0.051061	0.150124	0.100309	0.274127	0.350131	1.000000	0.371352	0.157327	0.386902	0.099205	-0.016748
10	0.118546	0.069569	0.042557	0.000234	0.031550	0.082473	0.079717	0.384906	0.538013	0.371352	1.000000	0.218162	0.616635	0.092240	-0.022885
11	0.095082	-0.001974	0.044143	-0.045469	0.092608	0.165921	0.137509	0.223315	0.217874	0.157327	0.218162	1.000000	0.225657	0.134426	-0.012872
12	0.113034	0.071614	0.033629	0.000869	0.030886	0.070198	0.077934	0.401244	0.562419	0.386902	0.616635	0.225657	1.000000	0.093102	-0.023679
13	0.137468	-0.004840	0.017959	-0.004989	0.662195	0.331278	0.673335	0.697162	0.137011	0.099205	0.092240	0.134426	0.093102	1.000000	-0.023999
14	-0.011228	-0.000594	-0.004732	0.003310	-0.018842	-0.019688	-0.022948	-0.032207	-0.023289	-0.016748	-0.022885	-0.012872	-0.023679	-0.023999	1.000000

- Ma trận phần dư:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0.000000	-0.020763	0.030371	-0.023765	0.023091	-0.004130	-0.042616	0.035980	-0.011097	0.020051	0.002002	-0.051235	-0.002735	-0.006305	-0.004818
1	-0.020763	0.000000	-0.017126	0.047068	-0.009601	0.009224	0.028500	-0.050717	0.023153	-0.040753	0.001077	0.093191	0.000552	0.014862	0.001204
2	0.030371	-0.017126	0.000000	-0.015740	0.016096	-0.011471	-0.022936	0.014867	-0.009164	0.021324	-0.003342	-0.035308	0.001727	-0.002809	-0.000598
3	-0.023765	0.047068	-0.015740	0.000000	-0.029281	0.014836	0.038152	-0.010469	0.000918	-0.002585	0.000863	0.006175	-0.000532	-0.001077	0.002633
4	0.023091	-0.009601	0.016096	-0.029281	0.000000	-0.019884	-0.014150	0.027477	-0.013416	-0.017846	0.003864	-0.007506	0.004538	-0.004169	-0.010509
5	-0.004130	0.009224	-0.011471	0.014836	-0.019884	0.000000	0.035138	-0.017207	0.006811	-0.026975	0.001090	0.038294	0.003774	-0.001946	0.000369
6	-0.042616	0.028500	-0.022936	0.038152	-0.014150	0.035138	0.000000	-0.009590	0.000502	0.003934	-0.005100	0.052329	-0.009022	0.006322	-0.004764
7	0.035980	-0.050717	0.014867	-0.010469	0.027477	-0.017207	-0.009590	0.000000	0.024742	0.026446	-0.005615	-0.103652	0.005317	-0.004671	0.005041
8	-0.011097	0.023153	-0.009164	0.000918	-0.013416	0.006811	0.000502	0.024742	0.000000	0.007746	-0.017613	0.026746	-0.010960	-0.014803	-0.007182
9	0.020051	-0.040753	0.021324	-0.002585	-0.017846	-0.026975	0.003934	0.026446	0.007746	0.000000	0.000248	-0.004612	-0.017069	-0.002428	0.031087
10	0.002002	0.001077	-0.003342	0.000863	0.003864	0.001090	-0.005100	-0.005615	-0.017613	0.000248	0.000000	0.016570	0.012858	0.004867	0.018460
11	-0.051235	0.093191	-0.035308	0.006175	-0.007506	0.038294	0.052329	-0.103652	0.026746	-0.004612	0.016570	0.000000	0.012080	0.042185	-0.005760
12	-0.002735	0.000552	0.001727	-0.000532	0.004538	0.003774	-0.009022	0.005317	-0.010960	-0.017069	0.012858	0.012080	0.000000	-0.002537	-0.030650
13	-0.006305	0.014862	-0.002809	-0.001077	-0.004169	-0.001946	0.006322	-0.004671	-0.014803	-0.002428	0.004867	0.042185	-0.002537	0.000000	0.009446
14	-0.004818	0.001204	-0.000598	0.002633	-0.010509	0.000369	-0.004764	0.005041	-0.007182	0.031087	0.018460	-0.005760	-0.030650	0.009446	0.000000

- Ta thấy các phần tử đều gần bằng 0. Chứng tỏ mô hình phân tích nhân tố này hoàn toàn phù hợp.

Bước 7:

- Tính tổng tải bình phương, tỷ lệ phương sai:

$$\begin{bmatrix} & \textit{Factor 1} & \textit{Factor 2} & \textit{Factor 3} & \textit{Factor 4} \\ \textit{0} & 2.59725 & 2.311847 & 1.814135 & 1.086127 \\ \textit{1} & 0.17315 & 0.154123 & 0.120942 & 0.072408 \\ \textit{2} & 0.17315 & 0.327273 & 0.448215 & 0.520624 \end{bmatrix}$$

- Nhận xét:
  - Dòng 0 thể hiện tổng tải trọng bình phương (SS Loadings) của 4 nhân tố.
  - Dòng 1 thể hiện tỷ lệ của phương sai.
  - Dòng 2 thể hiện phương sai tích lũy.
    $\Longrightarrow$ 4 nhân tố này thể hiện tổng 52,06% phương sai của bộ dữ liệu.
  - Chúng ta có thể sử dụng những yếu tố mới này làm biến số cho các phân tích khác hoặc để dự đoán.

MÃ NGUỒN.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Buoc 1: Nhap du lieu.
df= pd.read_excel("C:/Users/admin/Downloads/BACH KHOA MAU NANG/PHAN TICH SO
    LIEU/BAO CAO PHAN TICH SO LIEU - NHOM 2/dulieu1.xlsx")
df = df.dropna()
df

# Buoc 2:
# Mo ta du lieu.
df.describe().T
# Kiem tra du lieu con thieu.
df.info()

# Buoc 3: Ma tran he so tuong quan.
import seaborn as sns
correlation = df.corr()
plt.figure(figsize=(60, 30))
sns.heatmap(correlation, vmax=1, square=True, annot=True, cmap='viridis')
plt.title('Correlation between different features')
plt.show()
df.drop(['Arrival Delay in Minutes'], axis=1, inplace=True)

# Cai dat thu vien factor_analyzer.
!pip install factor_analyzer
from factor_analyzer import FactorAnalyzer
```

```
# Buoc 4: Kiem tra mot vai gia dinh.
# Kiem dinh cau Bartlett.
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
calculate_bartlett_sphericity(df)
# Kiem dinh KMO.
from factor_analyzer.factor_analyzer import calculate_kmo
kmo_all, kmo_model = calculate_kmo(df)
print(kmo_model)

# Buoc 5: Xac dinh so luong cac nhan to.
fa = FactorAnalyzer(rotation=None)
fa.fit(df)
ev, v = fa.get_eigenvalues() #eigenvalues
ev
# Ve bieu do gia tri rieng
plt.scatter(range(1,df.shape[1]+1),ev)
plt.plot(range(1,df.shape[1]+1),ev)
plt.title('Scree Plot')
plt.xlabel('Factors')
plt.ylabel('Eigenvalue')
plt.grid()
plt.show()
fa = FactorAnalyzer(4, rotation='varimax')
fa.fit(df)
ev, v = fa.get_eigenvalues()
ev
# Ma tran tai trong.
lmatrix = pd.DataFrame(fa.loadings_, index = list(df.columns), columns = ['Factor
    1', 'Factor 2', 'Factor 3', 'Factor 4'])
lmatrix #loading matrix
```

```
lmatrix
plt.figure(figsize=(15, 30))
sns.heatmap(lmatrix, vmax=1, square=True, annot=True, cmap='viridis')
plt.title('Loadings of the factors')
plt.show()

# Buoc 6: Tinh cong dong, ma tran du.
# Ty le phuong sai cua moi bien duoc tinh theo cac nhan to duoc chon
fa.get_communalities()
cmatrix = pd.DataFrame(fa.get_communalities(), index = list(df.columns), columns
    = ['Communalities'])
cmatrix
# Ma tran du PSI.
gc = fa.get_communalities()
PSI = np.zeros((15, 1)) # Khi   to   ma   trn   PSI
for i in range(0, len(gc)):
    PSI[i] = 1 - gc[i]
PPSI = np.zeros((15, 15))
for i in range(0, len(PSI)):
    PPSI[i, i] = PSI[i]
PPSI1= pd.DataFrame(PPSI)
PPSI1
# Ma tran tuong quan phu hop.
ld= fa.loadings_
transpose_matrix = np.transpose(ld)
Sigma = np.dot(ld, transpose_matrix) + PPSI
Sigma1 = pd.DataFrame(Sigma)
Sigma1
# Ma tran phan du.
cor1 = df.corr().values
resid = cor1 - Sigma
```

```
resid
resid_df = pd.DataFrame(resid)
resid_df


# Buoc 7:
lmatrix.sort_values('Factor 1', ascending=False)
lmatrix.sort_values('Factor 2', ascending=False)
lmatrix.sort_values('Factor 3', ascending=False)
lmatrix.sort_values('Factor 4', ascending=False)
fa.get_factor_variance()
vmatrix = {
    'factor_1': [2.59725008, 0.17315001, 0.17315001],
    'factor_2': [2.31184703, 0.15412314, 0.32727314],
    'factor_3': [1.81413456, 0.1209423, 0.44821544],
    'factor_4': [1.08612727, 0.07240848, 0.52062393]
}
df5 = pd.DataFrame(vmatrix)
print(df5)
```

6.2 Bộ dữ liệu 2.

6.2.1 Mô tả bài toán.

Mô tả dữ liệu

Bộ dữ liệu gồm 50 quan sát và 5 biến chứa thông tin về kết cấu bánh ngọt bao gồm dữ liệu về:

- Oil: Lượng dầu trong bánh (Thể hiện độ béo của bánh).
- Density: Độ đặc của bánh.
- Crispy: Độ giòn của bánh.
- Fracture: Góc gãy (Góc uốn cong tối đa của bánh).
- Hardness: Lực cần thiết để làm gãy bánh (Độ cứng).

6.2.2 Chạy chương trình.

Bước 1: Nhập dữ liệu thu được kết quả như sau:

```
'data.frame':   50 obs. of  5 variables:
 $ Oil     : num  16.5 17.7 16.2 16.7 16.3 19.1 18.4 17.5 15.7 16.4 ...
 $ Density : int  2955 2660 2870 2920 2975 2790 2750 2770 2955 2945 ...
 $ Crispy  : int  10 14 12 10 11 13 13 10 11 11 ...
 $ Fracture: int  23 9 17 31 26 16 17 26 23 24 ...
 $ Hardness: int  97 139 143 95 143 189 114 63 123 132 ...
```

Bước 2: Thực hiện phân tích nhân tố bằng hàm Factanal với số lượng nhân tố bằng 2, thu được các kết quả như sau:

```
> food_fa <- factanal(food, factors = 2)
> food_fa

Call:
factanal(x = food, factors = 2)

Uniquenesses:
     Oil  Density   Crispy Fracture Hardness
   0.334    0.156    0.042    0.256    0.407

Loadings:
          Factor1 Factor2
Oil       -0.816
Density    0.919
Crispy    -0.745   0.635
Fracture   0.645  -0.573
Hardness          0.764

                Factor1 Factor2
SS loadings       2.490   1.316
Proportion Var    0.498   0.263
Cumulative Var    0.498   0.761

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 0.27 on 1 degree of freedom.
The p-value is 0.603
>
```

- Uniquenesses: Độ duy nhất (tính nhiễu của các biến).
- SS Loadings: Các giá trị (2.490 và 1.316) là "sum of squares loadings"cho mỗi thành phần chính. Chúng đại diện cho mức độ phân tán của dữ liệu được giải thích bởi từng thành phần. Giá trị càng cao, thành phần đó càng giải thích được nhiều biến đổi trong dữ liệu. Trong trường hợp này, Factor 1 giải thích được nhiều biến đổi hơn Factor 2.
- Proportion Var: Các giá trị (0.498 và 0.263) là tỷ lệ phương sai được giải thích bởi mỗi thành phần chính. Chúng thể hiện phần trăm tổng phương sai của dữ liệu được giải thích bởi từng thành phần. Ví dụ: Factor 1 giải thích 49,8% tổng phương sai trong dữ liệu.

- Cumulative Var: Các giá trị (0.498 và 0.761) là phương sai tích lũy được giải thích bởi mỗi thành phần chính. Chúng thể hiện phần trăm tổng phương sai được giải thích bởi thành phần đó và tất cả các thành phần trước đó. Ví dụ: Factor1 và Factor2 cùng nhau giải thích 76,1% tổng phương sai trong dữ liệu.
- Sự phù hợp của mô hình: Chi-square statistic (0,27) và p-value (0,603). Các giá trị này đánh giá sự phù hợp của mô hình. Giá trị p cao (> 0,05) cho thấy mô hình 2 nhân tố là một mô hình phù hợp với dữ liệu.

Bước 3:

- Tính duy nhất (tính nhiễu) của các biến:

$$\begin{bmatrix} Oil & Density & Crispy & Fracture & Hardness \\ 0.3338599 & 0.1555255 & 0.0422238 & 0.2560235 & 0.4069459 \end{bmatrix}$$

- Tính cộng đồng của các biến:

$$\begin{bmatrix} Oil & Density & Crispy & Fracture & Hardness \\ 0.6661398 & 0.8444745 & 0.9577762 & 0.7439766 & 0.5930539 \end{bmatrix}$$

Bước 4:

- Ma trận PSI:

$$\begin{bmatrix} 0.3338599 & 0.0000000 & 0.0000000 & 0.0000000 & 0.0000000 \\ 0.0000000 & 0.1555255 & 0.0000000 & 0.0000000 & 0.0000000 \\ 0.0000000 & 0.0000000 & 0.0422238 & 0.0000000 & 0.0000000 \\ 0.0000000 & 0.0000000 & 0.0000000 & 0.2560235 & 0.0000000 \\ 0.0000000 & 0.0000000 & 0.0000000 & 0.0000000 & 0.4069459 \end{bmatrix}$$

- Ma trận tương quan quan sát (S):

	Oil	*Density*	*Crispy*	*Fracture*	*Hardness*
Oil	1.00000000	−0.7500240	0.5930863	−0.5337392	−0.09604521
Density	−0.75002399	1.0000000	−0.6709460	0.5721324	0.10793720
Crispy	0.59308631	−0.6709460	1.0000000	−0.8439650	0.41109340
Fracture	−0.53373917	0.5721324	−0.8439650	1.0000000	−0.37335844
Hardness	−0.09604521	0.1079372	0.4110934	−0.3733584	1.00000000

- Ma trận tương quan phù hợp (Sigma):

	Oil	*Density*	*Crispy*	*Fracture*	*Hardness*
Oil	0.99999967	−0.7500246	0.5956994	−0.5155194	−0.09526886
Density	−0.75002460	1.0000000	−0.6698646	0.5796718	0.10825742
Crispy	0.59569942	−0.6698646	1.0000000	−0.8439652	0.41108842
Fracture	−0.51551938	0.5796718	−0.8439652	1.0000001	−0.37339100
Hardness	−0.09526886	0.1082574	0.4110884	−0.3733910	0.99999979

- Ma trận phần dư: Các số gần bằng 0 cho thấy mô hình nhân tố là tốt.

	Oil	*Density*	*Crispy*	*Fracture*	*Hardness*
Oil	0.000000	0.000001	−0.002613	−0.018220	−0.000776
Density	0.000001	0.000000	−0.001081	−0.007539	−0.000320
Crispy	−0.002613	−0.001081	0.000000	0.000000	0.000005
Fracture	−0.018220	−0.007539	0.000000	0.000000	0.000033
Hardness	−0.000776	−0.000320	0.000005	0.000033	0.000000

Bước 5: Giải thích các nhân tố, điều chỉnh mô hình ba yếu tố.

- Mô hình không xoay:

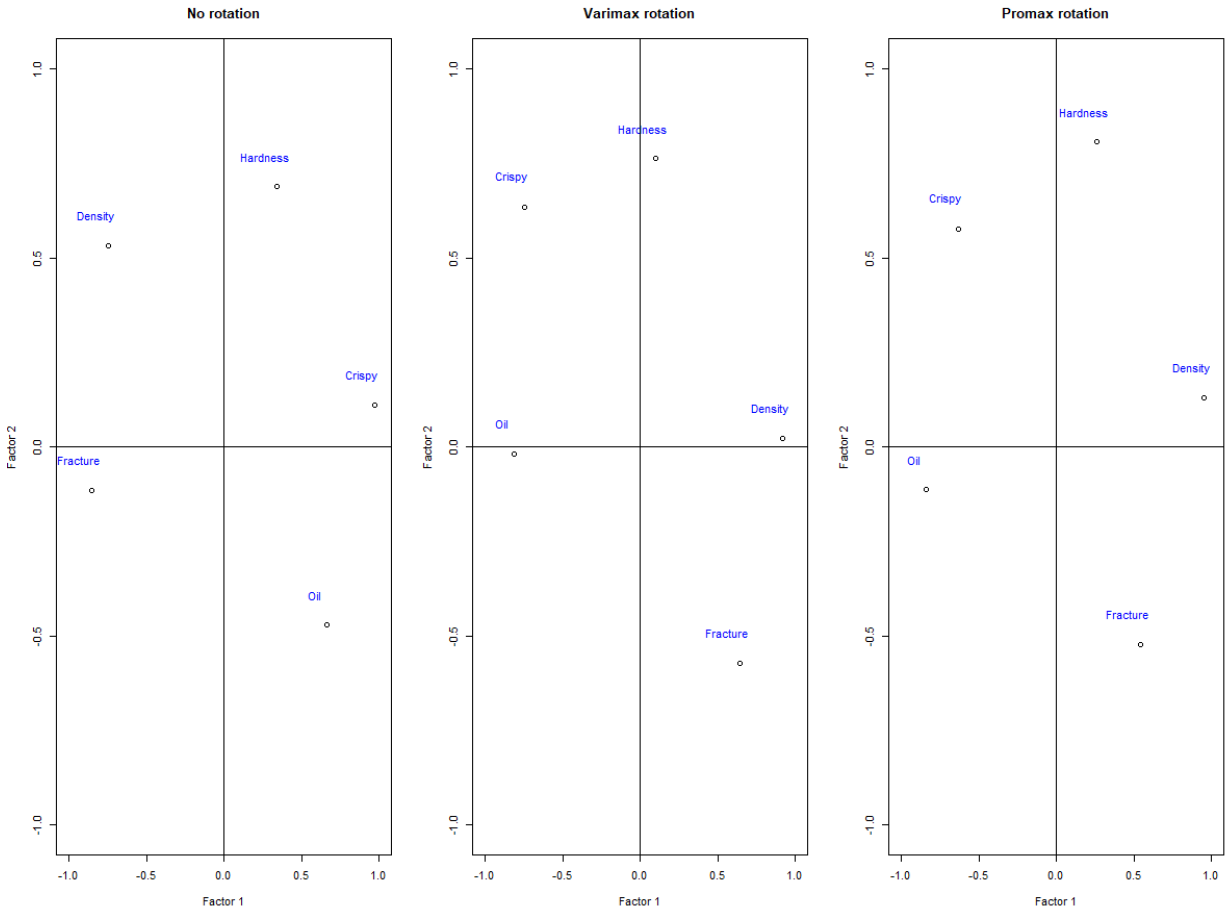
$$\begin{bmatrix} & \textit{Factor 1} & \textit{Factor 2} \\ \textit{Oil} & 0.6660743 & -0.4716830 \\ \textit{Density} & -0.7491979 & 0.5321437 \\ \textit{Crispy} & 0.9724288 & 0.1102655 \\ \textit{Fracture} & -0.8549300 & -0.1143299 \\ \textit{Hardness} & 0.3446544 & 0.6886707 \end{bmatrix}$$

- Mô hình có phép quay Varimax:

$$\begin{bmatrix} & \textit{Factor 1} & \textit{Factor 2} \\ \textit{Oil} & -0.81596104 & -0.01863759 \\ \textit{Density} & 0.91868262 & 0.02228699 \\ \textit{Crispy} & -0.74456598 & 0.63513597 \\ \textit{Fracture} & 0.64487779 & -0.57280821 \\ \textit{Hardness} & 0.09931344 & 0.76366928 \end{bmatrix}$$

- Mô hình có phép quay Promax:

$$\begin{bmatrix} & \textit{Factor 1} & \textit{Factor 2} \\ \textit{Oil} & -0.8433909 & -0.1134071 \\ \textit{Density} & 0.9498350 & 0.1290432 \\ \textit{Crispy} & -0.6347974 & 0.5767758 \\ \textit{Fracture} & 0.5451118 & -0.5232409 \\ \textit{Hardness} & 0.2600342 & 0.8080319 \end{bmatrix}$$



- Từ kết quả của phép quay có thể thấy Factor 1 chiếm tỷ lệ bánh ngọt loại bánh béo nhiều dầu (Oil), đặc (Density), có thể gọi là loại bánh béo mềm. Còn Factor 2 chiếm tỷ lệ bánh ngọt loại bánh cứng cần nhiều lực để bẻ gãy (Hardness), có thể gọi là loại bánh cứng.

MÃ NGUỒN.

```
# Buoc 1: Nhap du lieu.
setwd("C:\Users\admin\Downloads")
food <- read.csv("food-texture.csv",row.names = "X")
str(food)

# Buoc 2: Phan tich nhan to.
food_fa <- factanal(food, factors = 2)
food_fa$uniquenesses
food_fa$loadings

# Buoc 3:
#Tinh cong dong
h <- apply(food_fa$loadings, 1, function(row) sum(row^2))
h
# Tinh nhieu
1 - h

# Buoc 4:
Lambda <- food_fa$loadings
Psi <- diag(food_fa$uniquenesses)
S <- food_fa$correlation
Sigma <- Lambda %*% t(Lambda) + Psi
print(Lambda)
print(Psi)
print(S)
print(Sigma)
round(S - Sigma, 6)

# Buoc 5:
food_fa_none <- factanal(food, factors = 2, rotation = "none")
```

```
food_fa_varimax <- factanal(food, factors = 2, rotation = "varimax")
food_fa_promax <- factanal(food, factors = 2, rotation = "promax")
#no rotation
no_rotation <- data.frame(factor1 = food_fa_none$loadings[, 1],
                          factor2 = food_fa_none$loadings[, 2])
# varimax rotation
varimax_rotation <- data.frame(factor1 = food_fa_varimax$loadings[, 1],
                               factor2 = food_fa_varimax$loadings[, 2])
# promax rotation
promax_rotation <- data.frame(factor1 = food_fa_promax$loadings[, 1],
                              factor2 = food_fa_promax$loadings[, 2])
print(no_rotation)
print(varimax_rotation)
print(promax_rotation)

par(mfrow = c(1, 3))
plot(food_fa_none$loadings[, 1],
     food_fa_none$loadings[, 2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1, 1),
     xlim = c(-1, 1),
     main = "No rotation")
abline(h = 0, v = 0)

text(food_fa_none$loadings[, 1] - 0.08,
     food_fa_none$loadings[, 2] + 0.08,
     colnames(food),
     col = "blue")
abline(h = 0, v = 0)
```

```
plot(food_fa_varimax$loadings[, 1],
     food_fa_varimax$loadings[, 2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1, 1),
     xlim = c(-1, 1),
     main = "Varimax rotation")

text(food_fa_varimax$loadings[, 1] - 0.08,
     food_fa_varimax$loadings[, 2] + 0.08,
     colnames(food),
     col = "blue")
abline(h = 0, v = 0)

plot(food_fa_promax$loadings[, 1],
     food_fa_promax$loadings[, 2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1, 1),
     xlim = c(-1, 1),
     main = "Promax rotation")
abline(h = 0, v = 0)

text(food_fa_promax$loadings[, 1] - 0.08,
     food_fa_promax$loadings[, 2] + 0.08,
     colnames(food),
     col = "blue")
abline(h = 0, v = 0)
```

6.3 Bộ dữ liệu 3.

6.3.1 Mô tả bài toán.

Mô tả dữ liệu

Bộ dữ liệu gồm 100 quan sát về 11 yếu tố ảnh hưởng đến sự hài lòng của khách hàng. Ngoài 11 yếu tố, còn có 1 biến đo lường sự hài lòng của khách hàng.

1. ProQual: Product Quality (Chất lượng sản phẩm).
2. Ecom: E-Commerce (Thương mại điện tử).
3. TechSup: Technical Support (Hỗ trợ kỹ thuật cho khách hàng).
4. Advertising: Quảng cáo.
5. ProdLine: Product Line (Dòng sản phẩm).
6. SalesFImage: Sales Force Image (Hình ảnh đội ngũ bán hàng).
7. ComPricing: Competitive Pricing (Giá cả cạnh tranh).
8. WartyClaim: Warranty Claims (Bảo hành và giải quyết yêu cầu bồi thường).
9. OrdBilling: Order Billing (Đặt hàng và thanh toán).
10. DelSpeed: Delivery Speed (Tốc độ giao hàng).
11. CompRes: Complaint Resolution (Giải quyết khiếu nại).

6.3.2 Chạy chương trình.

Bước 1: Tiền xử lý dữ liệu: Dùng Excel làm sạch dữ liệu trước khi nạp vào SPSS, ta tiến hành bỏ đi các dữ liệu bị trống, dư thừa, không đúng định dạng.

Bước 2: Tính ma trận hệ số tương quan.

Correlation Matrix

		ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
Correlation	ProdQual	1.000	-.137	.096	.106	-.053	.477	-.152	-.401	.088	.104	.028
	Ecom	-.137	1.000	.001	.140	.430	-.053	.792	.229	.052	.156	.192
	TechSup	.096	.001	1.000	.097	-.063	.193	.017	-.271	.797	.080	.025
	CompRes	.106	.140	.097	1.000	.197	.561	.230	-.128	.140	.757	.865
	Advertising	-.053	.430	-.063	.197	1.000	-.012	.542	.134	.011	.184	.276
	ProdLine	.477	-.053	.193	.561	-.012	1.000	-.061	-.495	.273	.424	.602
	SalesFImage	-.152	.792	.017	.230	.542	-.061	1.000	.265	.107	.195	.272
	ComPricing	-.401	.229	-.271	-.128	.134	-.495	.265	1.000	-.245	-.115	-.073
	WartyClaim	.088	.052	.797	.140	.011	.273	.107	-.245	1.000	.197	.109
	OrdBilling	.104	.156	.080	.757	.184	.424	.195	-.115	.197	1.000	.751
	DelSpeed	.028	.192	.025	.865	.276	.602	.272	-.073	.109	.751	1.000

Bước 3:

- Kiểm định cầu Bartlett thu được kết quả:
  - Chi - square = 619.273.
  - df = 55.
  - Sig < 0.001.

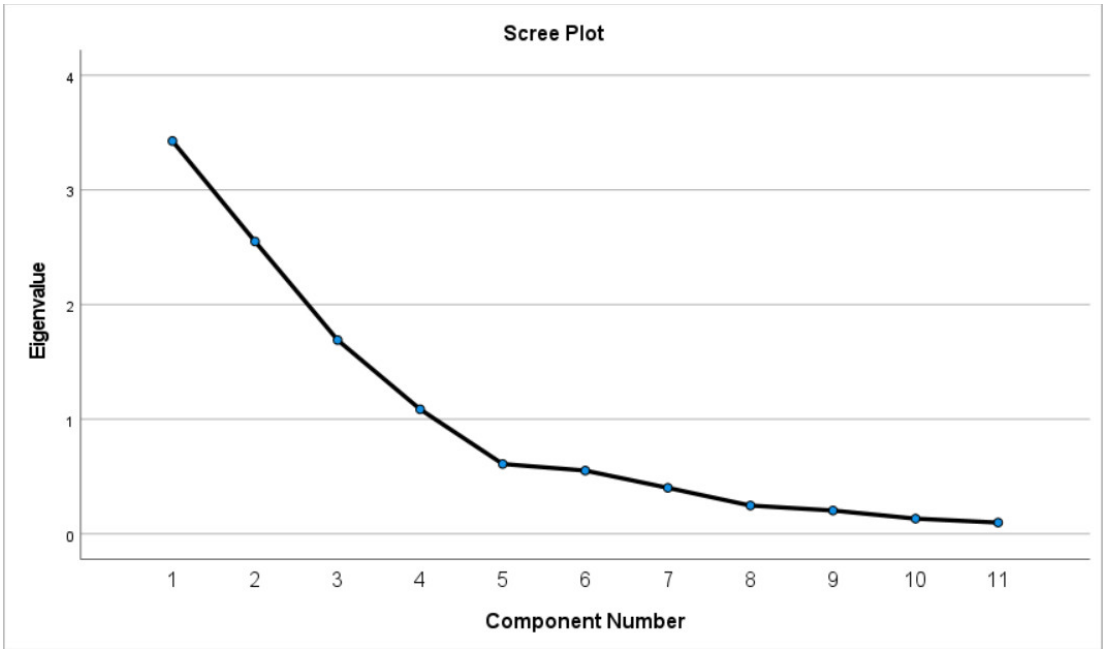
  Sig < 0.05, cho thấy các biến có tương quan với nhau, bộ dữ liệu có ý nghĩa thống kê và có thể phân tích nhân tố.

- Hệ số KMO: KMO = 0.653 > 0.5, dữ liệu phù hợp để phân tích nhân tố.

Bước 4: Tỷ lệ tích luỹ phương sai. Lựa chọn phương pháp phân tích thành phần chính.

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	3.427	31.154	31.154	3.427	31.154	31.154	3.189
2	2.551	23.190	54.344	2.551	23.190	54.344	2.449
3	1.691	15.373	69.717	1.691	15.373	69.717	2.024
4	1.087	9.878	79.595	1.087	9.878	79.595	2.199
5	.609	5.540	85.135				
6	.552	5.017	90.152				
7	.402	3.650	93.802				
8	.247	2.245	96.047				
9	.204	1.850	97.898				
10	.133	1.208	99.105				
11	.098	.895	100.000				

- Ma trận tương quan có bốn giá trị riêng lớn hơn 1, quan hệ giữa 11 biến trên có thể được giải thích qua 4 nhân tố. Tỷ lệ tích luỹ phương sai là 79.6% hay khi sử dụng 4 nhân tố thì 79.6% phương sai của các biến được giải thích.
- Biểu đồ Scree Plot cho biết mức độ thể hiện của các giá trị riêng:



Bước 5: Ma trận hệ số tải.

	1	*2*	*3*	*4*
DelSpeed	0.876	0.117	−0.302	−0.206
CompRes	0.871		−0.274	−0.215
OrdBilling	0.809		−0.220	−0.247
ProdLine	0.716	−0.455	−0.151	−0.212
SalesFimage	0.377	0.752	0.314	0.232
Ecom	0.307	0.713	0.306	0.284
ComPricing	−0.281	0.660		−0.348
Advertising	0.340	0.581	0.115	0.331
TechSup	0.292	−0.369	0.794	−0.202
WartyClaim	0.394	−0.306	0.778	−0.193
ProdQual	0.248	−0.501		0.670

Ta thấy hệ số tải của yếu tố ProQual cao với cả nhân tố 2 và 4, yếu tố ProLine có vẻ quan trọng với cả hai nhân tố 1 và 2. Điều này không cung cấp một cách giải thích dữ liệu đơn giản và rõ ràng nên cần sử dụng một phép quay.

Bước 6: Quay nhân tố.

- Lựa chọn phép quay xiên Promax với Kappa = 4 và ma trận xoay:

Component	*1*	*2*	*3*	*4*
1	1.000	0.207	0.159	0.286
2	0.207	1.000	−0.008	−0.237
3	0.159	−0.008	1.000	0.239
4	0.286	−0.237	0.239	1.000

- Ma trận tải sau khi xoay:

	1	*2*	*3*	*4*
DelSpeed	0.949	0.270		0.185
CompRes	0.938	0.203	0.133	0.235
OrdBilling	0.872	0.194	0.157	0.180
SalesFimage	0.221	0.922		−0.244
Ecom	0.147	0.879		−0.215
Advertising	0.221	0.744		
TechSup			0.943	0.210
WartyClaim	0.189		0.942	0.215
ProdQual	0.113	−0.124		0.848
ComPricing	−0.168	0.305	−0.335	−0.773
ProdLine	0.667		0.267	0.748

  - Với nhân tố thứ nhất, có các hệ số tải cao là 0.949 ứng với yếu tố DelSpeed, 0.938 ứng với yếu tố CompRes, 0.872 ứng với yếu tố OrdBilling, 0.667 ứng với yếu tố ProLine.
  - Nhận xét tương tự như trên với các nhân tố 2, 3, 4, chúng ta có thể nhóm các yếu tố lại như sau:

    1. Nhân tố 1: CompRes, DelSpeed, OrdBilling, ProLine.
    2. Nhân tố 2: SalesFImage, Ecom, Advertising.
    3. Nhân tố 3: TechSup, WartyClaim.
    4. Nhân tố 4: ProQual, ComPricing, ProLine.

- Tính cộng đồng (Phương sai chung): Tính cộng đồng của các biến đều lớn hơn 50%. Phương sai của biến Advertising được thể hiện ít nhất qua 4 nhân tố, chỉ 58.5%. Phương sai của biến TechSup và DelSpeed được giải thích rõ nhất qua 4 yếu tố.

	Initial	*Extraction*
ProdQual	1.000	0.768
Ecom	1.000	0.777
TechSup	1.000	0.893
CompRes	1.000	0.881
Advertising	1.000	0.576
ProdLine	1.000	0.787
SalesFimage	1.000	0.859
ComPricing	1.000	0.641
WartyClaim	1.000	0.892
OrdBilling	1.000	0.766
DelSpeed	1.000	0.914

Bước 7: Gọi tên và giải thích các nhân tố.

Nhân tố	Biến	Tên nhân tố	Mô tả
1	CompRes, DelSpeed, OrdBilling, ProLine.	Purchase (Mua hàng).	Các yếu tố đều liên quan đến việc mua hàng từ đặt hàng, thanh toán, giao hàng đến giải quyết các khiếu nại.
2	SalesFImage, Ecom, Advertising.	Marketing (Tiếp thị).	Các yếu tố liên quan đến tiếp thị như hình ảnh đội ngũ bán hàng, chi phí chi cho quảng cáo, hình thức thương mại điện tử.
3	TechSup, WartyClaim.	Post - Purchase Service (Dịch vụ sau khi mua hàng).	Các yếu tố liên quan đến dịch vụ sau khi mua hàng như hỗ trợ công nghệ, bảo hành...
4	ProQual, ComPricing, ProLine.	Product Position (Định vị sản phẩm).	Nhân tố định vị sản phẩm với các yếu tố về chất lượng sản phẩm, giá cả, dòng sản phẩm.

TỔNG KẾT

Có thể thấy, phân tích nhân tố là một công cụ mạnh mẽ để tìm cách hiểu các cấu trúc và mối quan hệ cơ bản trong dữ liệu, giúp giảm độ phức tạp của một số lượng lớn các biến quan sát bằng cách xác định một tập hợp nhỏ hơn các yếu tố cơ bản nắm bắt thông tin thiết yếu có trong dữ liệu, giúp xác định và khái niệm hóa các biến tiềm ẩn hoặc không thể quan sát có ảnh hưởng đến các biến quan sát. Phân tích nhân tố có thể được thực hiện bằng các phần mềm thống kê như SPSS, R, SAS hoặc Python, cung cấp các phương pháp trích xuất và xoay nhân tố khác nhau để phù hợp với các loại dữ liệu và mục tiêu nghiên cứu khác nhau.

Trên đây là toàn bộ nội dung bài báo cáo về chủ đề **"Phân tích nhân tố - Factor Analysis"** của Nhóm 2 chúng em. Báo cáo đã đưa ra được mô hình nhân tố trực giao, trình bày được các phương pháp ước lượng (Phương pháp phân tích thành phần chính, Phương pháp ước lượng hợp lý cực đại), tìm hiểu về phép quay nhân tố, điểm nhân tố và ứng dụng vào ba ví dụ thực tế.

Một lần nữa, nhóm chúng em xin gửi lời cảm ơn đến thầy **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!

Tài liệu

[1] Berry, M. J. A., and G. Linoff. *Data Mining Techniques: For Marketing, Sales and Customer Relationship Management* (2nd ed.) (paperback). New York: John Wiley, 2004.

[2] Berthold, M., and D. J. Hand. *Intelligent Data Analysis* (2nd ed.). Berlin, Germany: Springer-Verlag, 2003.

[3] Cormack, R. M. "A Review of Classification (with discussion)." *Journal of the Royal Statistical Society (A)*, 134, no. 3 (1971), 321–367.

[4] Dean W.Wichern, Richard A.Johnson (2015) *Applied Multivariate Statistical Analysis*, 6th Edition, Pearson Education Inc.

[5] Harmon, H. H. *Modern Factor Analysis* (3rd ed). Chicago: The University of Chicago Press, 1976.

[6] 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.

[7] ThS. Lê Xuân Lý, Bài giảng "Phân tích số liệu", Đại học Bách khoa Hà Nội, 2023.