

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN



PHÂN TÍCH THÔNG KÊ DỮ LIỆU NHIỀU BIẾN
LÝ THUYẾT CUỐI KỲ
BÁO CÁO
Đề tài
PHÂN TÍCH DỮ KIẾN (FACTOR ANALYSIS)

Giảng viên hướng dẫn:
Lý Quốc Ngọc

Nhóm thực hiện: **MSA**

Họ và tên	Mã số sinh viên
Tạ Tiến Thành Đạt	1712333
Nguyễn Quý Em	1712399
Châu Phương Gia	1712400
Nguyễn Thế Lợi	1712573

Ngày 19 tháng 08 năm 2020

MỤC LỤC

MỤC LỤC.....	2
PHÂN TÍCH DỮ KIẾN (FACTOR ANALYSIS)	3
1. Giới thiệu	3
2. Mô hình Nhân tố trực giao (The Orthogonal Factor Model).....	4
3. Các phương pháp ước lượng (Methods of Estimation)	10
3.1. Phương pháp Phân tích thành phần chính (The Principal Component Method).....	11
3.2. Phương pháp Triển vọng cực đại (The Maximum Likelihood Method) ..	17
3.3. Kiểm định số nhân tố cho mẫu lớn (A Large Sample Test for the Number of Common Factors)	19
4. Xoay nhân tố (Factor Rotation)	21
Xoay nghiêng	27
5. Điểm nhân tố (Factor Scores)	28
5.1. Phương pháp bình phương tối thiểu có trọng số (the weighted least squares method)	28
5.2 Phương pháp hồi quy	30
6. Quan điểm và chiến lược Phân tích dữ kiện (Perspectives and a Strategy for Factor Analysis)	31
7. Kết luận (Conclusion)	32
8. Cài đặt	32
8.1. Chuẩn bị	32
8.2. Các tác vụ cơ bản để phân tích dữ kiện	34
8.3. Tiến hành cài đặt	34
8.4. Tổng kết	44
9. Tài liệu tham khảo (Reference)	45

PHÂN TÍCH DỮ KIẾN (FACTOR ANALYSIS)

1. Giới thiệu

Phân tích dữ kiện hay *Phân tích nhân tố* (Factor Analysis) là một phương pháp thống kê dùng để mô tả (nếu có thể) mối quan hệ hiệp phương sai giữa các biến quan sát được theo các biến mới không quan sát được, gọi là *nhân tố* (factor).

Về cơ bản, *mô hình nhân tố* (factor model) được thúc đẩy bởi hai yếu tố chính:

- 1) Giả sử các biến có thể được nhóm theo sự tương quan (correlation) của chúng. Điều này có nghĩa là các biến trong cùng một nhóm sẽ có mối tương quan cao hơn so với các biến ở nhóm khác.
- 2) Mỗi nhóm sẽ có một yếu tố đại diện duy nhất, được gọi là nhân tố (factor).

Trong lịch sử, phân tích dữ kiện đã gây ra khá nhiều tranh cãi. Khái niệm hiện đại này được đưa ra vào đầu thế kỷ 20 bởi Karl Pearson, Charles Spearman và những người khác để nhằm xác định và đo lường trí thông minh. Chính vì sớm có sự liên kết với các cấu trúc như trí thông minh, nên phân tích dữ kiện đã được nuôi dưỡng và phát triển chủ yếu bởi các nhà khoa học có sự quan tâm đến Tâm lý học (Psychometrics). Những tranh cãi về tâm lý trong một số nghiên cứu sơ khai và việc thiếu cơ sở tính toán mạnh mẽ đã cản trở sự phát triển ban đầu để nó trở thành một phương pháp thống kê. Sự ra đời của máy tính tốc độ cao gần như đã giải quyết được tất cả các tranh luận và bác bỏ hầu hết các kỹ thuật cũ, tạo ra một bước ngoặt cả về lý thuyết lẫn tính toán cho Phân tích dữ kiện.

Phân tích dữ kiện được ứng dụng trong nhiều lĩnh vực như tâm lý, kinh tế, khoa học hành vi và đặc biệt là các ngành khoa học dữ liệu. Ví dụ: Các nhà tâm lý học có thể sử dụng Phân tích dữ kiện để đánh giá nhân cách con người, một yếu tố không thể quan sát hay thu thập được. Ví dụ khác là khi tìm sự tương quan giữa điểm số các môn học như Toán, Lý, Hoá, Tiếng Anh, Tiếng Pháp, Âm nhạc,...người ta có thể tìm ra và đánh giá được “nhân tố ẩn” là “sự thông minh”.

Phân tích dữ kiện có thể được coi là sự mở rộng của *Phân tích thành phần chính* (Principal Component Analysis). Cả hai phương pháp đều cố gắng xấp xỉ ma trận hiệp phương sai. Tuy nhiên, sự xấp xỉ dựa trên mô hình phân tích dữ kiện thì chi tiết hơn. Câu hỏi chính được đặt ra trong Phân tích dữ kiện là liệu rằng dữ liệu có phù hợp với cấu trúc giả định từ trước hay không.

2. Mô hình Nhân tố trực giao (The Orthogonal Factor Model)

Cho một vector ngẫu nhiên quan sát được \mathbf{X} , với p thành phần, có trung bình $\boldsymbol{\mu}$ và ma trận hiệp phương sai Σ . Mô hình nhân tố quy định rằng \mathbf{X} phụ thuộc tuyến tính vào một vài biến ngẫu nhiên không quan sát được F_1, F_2, \dots, F_m , được gọi là các *nhân tố chung* (common factors), và p thành phần bổ sung $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$, được gọi là *lỗi* (errors) hay cũng gọi là *nhân tố riêng* (specific factors). Mô hình phân tích dữ kiện (nhân tố):

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

Biểu diễn ở dạng ma trận

$$\underbrace{\mathbf{X} - \boldsymbol{\mu}}_{(p \times 1)} = \underbrace{\mathbf{L}}_{(p \times m)} \underbrace{\mathbf{F}}_{(m \times 1)} + \underbrace{\boldsymbol{\varepsilon}}_{(p \times 1)} \quad (2)$$

Hệ số l_{ij} được gọi là *hệ số tải* (loading) của biến thứ i trên nhân tố thứ j . Do đó, ma trận \mathbf{L} được gọi là *ma trận hệ số tải* (matrix of factor loading). Lưu ý rằng nhân tố riêng thứ i (ε_i) chỉ có sự đóng góp vào biến thứ i (X_i). Lúc này, p độ lệch giữa $X_1 - \mu_1, X_2 - \mu_2, \dots, X_p - \mu_p$ được thể hiện bởi $p + m$ biến ngẫu nhiên $F_1, F_2, \dots, F_m, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$ không quan sát được.

Với quá nhiều biến không quan sát được, việc xác định trực tiếp mô hình nhân tố từ các quan sát dựa trên $X_1, X_2, X_3, \dots, X_p$ là không có hi vọng. Tuy nhiên, với một số giả định sau đây về các vector ngẫu nhiên \mathbf{F} và $\boldsymbol{\varepsilon}$, mô hình (2) thể hiện mối quan hệ hiệp phương sai có thể được kiểm tra như dưới đây.

Giả sử:

$$\begin{aligned} E(\mathbf{F}) &= \underbrace{\mathbf{0}}_{(m \times 1)}, \quad Cov(\mathbf{F}) = E[\mathbf{F}\mathbf{F}'] = \underbrace{\mathbf{I}}_{(m \times m)} \\ E(\boldsymbol{\varepsilon}) &= \underbrace{\mathbf{0}}_{(p \times 1)}, \quad Cov(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \underbrace{\boldsymbol{\Psi}}_{(p \times p)} = \begin{bmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_p \end{bmatrix} \end{aligned} \quad (3)$$

Và ta giả định \mathbf{F} và $\boldsymbol{\varepsilon}$ là độc lập. Do đó:

$$Cov(\boldsymbol{\varepsilon}, \mathbf{F}) = E(\boldsymbol{\varepsilon}\mathbf{F}') = \underset{(p \times m)}{\mathbf{0}}$$

Những giả định trên và mối liên hệ ở công thức (2) cho chúng ta *mô hình nhân tố trực giao*

Mô hình Nhân tố trực giao với m nhân tố chung

$$\underbrace{\mathbf{X} - \boldsymbol{\mu}}_{(p \times 1)} = \underbrace{\mathbf{L}}_{(p \times m)} \underbrace{\mathbf{F}}_{(m \times 1)} + \underbrace{\boldsymbol{\varepsilon}}_{(p \times 1)}$$

- μ_i = trung bình của biến thứ i
- ε_i = nhân tố riêng thứ i
- F_i = nhân tố chung thứ i
- l_{ij} = hệ số tải của biến thứ i trên nhân tố thứ j (4)

Các vector ngẫu nhiên không quan sát được \mathbf{F} và $\boldsymbol{\varepsilon}$ thoả các điều kiện:

- \mathbf{F} và $\boldsymbol{\varepsilon}$ độc lập
- $E(\mathbf{F}) = \mathbf{0}, Cov(\mathbf{F}) = \mathbf{I}$
- $E(\boldsymbol{\varepsilon}) = \mathbf{0}, Cov(\boldsymbol{\varepsilon}) = \boldsymbol{\psi}$, với $\boldsymbol{\psi}$ là

Mô hình nhân tố trực giao “hàm chứa” cấu trúc hiệp phương sai cho \mathbf{X} . Từ mô hình (4), ta có:

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

Do đó:

$$\begin{aligned} \Sigma &= Cov(\mathbf{X}) = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' \\ &= \mathbf{LE}(\mathbf{FF}')\mathbf{L}' + E(\boldsymbol{\varepsilon}\mathbf{F}')\mathbf{L}' + \mathbf{LE}(\mathbf{F}\boldsymbol{\varepsilon}') + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\ &= \mathbf{LL}' + \boldsymbol{\psi} \end{aligned}$$

Ta có được kết quả trên là theo giả định (3). Bởi sự độc lập:

$$Cov(\boldsymbol{\varepsilon}, \mathbf{F}) = E(\boldsymbol{\varepsilon}, \mathbf{F}') = \mathbf{0}.$$

Hơn nữa, từ mô hình (4), $(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}' = (\mathbf{LF} + \boldsymbol{\varepsilon})\mathbf{F}' = \mathbf{LF}\mathbf{F}' + \boldsymbol{\varepsilon}\mathbf{F}'$. Do đó:

$$Cov(\mathbf{X}, \mathbf{F}) = E(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}' = \mathbf{LE}(\mathbf{FF}') + E(\boldsymbol{\varepsilon}\mathbf{F}') = \mathbf{L}.$$

Cấu trúc hiệp phương sai cho Mô hình nhân tố trực giao

1. $Cov(X) = LL' + \psi$

Hay:

$$Var(X_i) = l_{i1}^2 + \dots + l_{im}^2 + \psi_i \quad (5)$$

$$Cov(X_i, X_k) = l_{i1}l_{k1} + \dots + l_{im}l_{km}$$

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

Hay:

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

Mô hình $X - \mu = LF + \varepsilon$ là mô hình tuyến tính của các nhân tố. Nếu p biến ngẫu nhiên của X , trên thực tế, liên quan đến các nhân tố bên dưới, nhưng mối quan hệ là phi tuyến, chẳng hạn như $X_1 - \mu_1 = l_{11}F_1F_3 + \varepsilon_1$, $X_2 - \mu_2 = l_{21}F_2F_3 + \varepsilon_2$, ... thì cấu trúc hiệp phương sai $LL' + \psi$ ta có được ở (5) có thể không còn phù hợp. Giả định về sự tuyến tính là rất quan trọng trong việc xây dựng mô hình nhân tố truyền thống.

Một phần phương sai của biến ngẫu nhiên thứ i được đóng góp bởi m nhân tố chung được gọi là *phần chung (communality)* thứ i , một phần được đóng góp từ *phương sai riêng (specific variance)*. Ký hiệu phần chung thứ i là h_i^2 , từ (5), ta có:

$$\underbrace{\sigma_{ii}}_{Var(X_i)} = \underbrace{l_{i1}^2 + l_{i2}^2 + \dots + l_{im}^2}_{communality} + \underbrace{\psi_i}_{specific\ variance}$$

Hay:

$$h_i^2 = l_{i1}^2 + l_{i2}^2 + \dots + l_{im}^2 \quad (6)$$

Và:

$$\sigma_{ii} = h_i^2 + \psi_i, \quad i = 1, 2, \dots, p$$

Phần chung thứ i là tổng bình phương của các hệ số tải của biến thứ i .

Tóm lại, từ mô hình nhân tố trực giao, ta có thể xây dựng lại phương sai và hiệp phương sai cho các biến ngẫu nhiên.

Ví dụ 1 (Xác minh mối quan hệ $\Sigma = LL' + \psi$ cho hai nhân tố)

Xét ma trận hiệp phương sai

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

Ta sẽ đưa ma trận hiệp phương sai về dạng

$$\Sigma = LL' + \psi$$

Thật vậy,

$$\begin{bmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} 4 & 7 & -1 & 1 \\ 1 & 2 & 6 & 8 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

Có thể dễ dàng kiểm tra lại bằng đại số ma trận. Theo đó, ma trận hiệp phương sai có cấu trúc được tạo bởi mô hình 2 nhân tố trực giao. Khi đó:

$$L = \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \\ l_{31} & l_{32} \\ l_{41} & l_{42} \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix},$$

$$\psi = \begin{bmatrix} \psi_1 & 0 & 0 & 0 \\ 0 & \psi_2 & 0 & 0 \\ 0 & 0 & \psi_3 & 0 \\ 0 & 0 & 0 & \psi_4 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

Từ (2.6), ta có thể tính phần chung (communality) của X_1

$$h_1^2 = l_{11}^2 + l_{12}^2 = 4^2 + 1^2 = 17$$

Phương sai của X_1 có thể được tái lập:

$$\sigma_{11} = (l_{11}^2 + l_{12}^2) + \psi_1 = h_1^2 + \psi_1 = 17 + 2 = 19$$

Cũng chính là $\Sigma_{11} = \sigma_{11} = 19$.

Tương tự cho các biến còn lại.

Mô hình nhân tố giả định rằng p phương sai và $\frac{p(p-1)}{2}$ hiệp phương sai, tức là $\frac{p(p+1)}{2}$ tham số của \mathbf{X} , có thể được tái lập từ pm hệ số tải l_{ij} và p phương sai riêng ψ_i , tức là chỉ còn $pm + p$ tham số. Khi $m = p$, bất kỳ một ma trận hiệp phương sai Σ nào cũng có thể được tái lập một cách chính xác như \mathbf{LL}' , vì vậy ψ có thể là ma trận không. Tuy nhiên, khi m tương đối nhỏ hơn p thì phân tích dữ kiện mới thật sự hữu ích. Trong trường hợp này, phân tích dữ kiện cung cấp một các giải thích đơn giản hơn về hiệp phương sai của \mathbf{X} với ít tham số hơn thay vì $\frac{p(p+1)}{2}$ tham số của \mathbf{X} . Ví dụ, nếu \mathbf{X} chứa 15 biến và mô hình (9.4) với 3 nhân tố là phù hợp, thì $\frac{p(p+1)}{2} = \frac{15(15+1)}{2} = 120$ thành phần của Σ được mô tả chỉ bởi $mp + p = 3 \times 15 + 3 = 48$ tham số của mô hình nhân tố.

Nhưng thật không may cho các nhà phân tích dữ kiện vì không phải ma trận hiệp phương sai nào cũng có thể phân tích thành dạng $\mathbf{LL}' + \psi$ với số lượng nhân tố m nhỏ hơn nhiều so với p . Ví dụ dưới đây cho ta thấy một trong những vấn đề gặp phải khi cố gắng xác định các tham số l_{ij} và ψ_i từ các phương sai và hiệp phương sai của các biến quan sát.

Ví dụ 2 (Không tồn tại một lời giải thích hợp)

Cho $p = 3, m = 1$ và giả sử các biến ngẫu nhiên X_1, X_2, X_3 có ma trận hiệp phương sai xác định dương:

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

Sử dụng mô hình ở (4), 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 ở (5):

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

Cụ thể:

$$\Sigma = \begin{bmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{bmatrix} = \begin{bmatrix} l_{11} \\ l_{21} \\ l_{31} \end{bmatrix} [l_{11} \quad l_{21} \quad l_{31}] + \begin{bmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_3 \end{bmatrix}$$

Ta có các phương trình:

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

$$0.9 = l_{11}l_{21}$$

$$0.7 = l_{11}l_{31}$$

$$1 = l_{21}^2 + \psi_2$$

$$0.4 = l_{21}l_{31}$$

Xét cặp phương trình

$$0.7 = l_{11}l_{31}$$

$$0.4 = l_{21}l_{31}$$

$$\rightarrow l_{21} = \frac{4}{7}l_{11}$$

Thay kết quả trên vào phương trình $0.9 = l_{11}l_{21}$, ta được

$$0.9 = l_{11}l_{21} = l_{11}\frac{4}{7}l_{11} = \frac{4}{7}l_{11}^2 \Leftrightarrow l_{11}^2 = 1.575 \Leftrightarrow l_{11} = \pm 1.255.$$

Trong khi $Var(F_1) = 1$ (theo giả sử), $Var(X_1) = 1$ (theo ma trận đề cho)

$l_{11} = Cov(X_1, F_1) = Corr(X_1, F_1)$. Trong khi hệ số tương quan Correlation phải là một số nằm trong khoảng $[-1, 1] \rightarrow$ mâu thuẫn.

Hơn nữa, từ

$$1 = l_{11}^2 + \psi_1 \Leftrightarrow \psi_1 = 1 - l_{11}^2 = 1 - 1.575 = -0.575$$

Kết quả này là không thoả đáng vì giá trị phương sai $Var(\varepsilon_1) = \psi_1$ phải là một giá trị không âm.

Do đó, với ví dụ $m = 1$, ta hoàn toàn có thể tìm được lời giải từ phương trình $\Sigma = \mathbf{L}\mathbf{L}' + \boldsymbol{\psi}$. Tuy nhiên, lời giải này lại cho ra kết quả không phù hợp với mục đích thống kê, vì vậy nó không phải là lời giải phù hợp.

Khi $m > 1$, luôn luôn có một vài sự mơ hồ liên quan đến mô hình nhân tố. Để thấy rõ điều này, ta đặt \mathbf{T} là ma trận trực giao bất kì có kích thước $m \times m$. Khi đó: $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$. Ta có thể viết lại biểu diễn ở (2):

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{L}^*\mathbf{F}^* + \boldsymbol{\varepsilon} \quad (7)$$

Với:

$$\mathbf{L}^* = \mathbf{L}\mathbf{T}, \mathbf{F}^* = \mathbf{T}'\mathbf{F}$$

Khi đó:

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

Và

$$Cov(\mathbf{F}^*) = \mathbf{T}'Cov(\mathbf{F})\mathbf{T} = \mathbf{T}'\mathbf{T} = \underset{(m \times m)}{\mathbf{I}}$$

Các nhân tố F và $F^* = T'F$ có cùng các đặc tính thống kê. Mặc dù ma trận hệ số tải L^* , nói chung, khác với ma trận hệ số tải L , nhưng cả hai đều tạo ra cùng một ma trận hiệp phương sai Σ . Đó là:

$$\Sigma = LL' + \psi = LTT'L' + \psi = (L^*)(F^*) + \psi \quad (8)$$

Sự không rõ ràng này cung cấp cơ sở lý luận để dẫn đến khái niệm “xoay nhân tố” (factor rotation), khi các ma trận trực giao tương ứng với các phép quay (và ánh xạ) của hệ tọa độ cho X .

Các hệ số tải L chỉ được xác định bởi một ma trận trực giao T . Do đó, các hệ số tải

$$L^* = LT \text{ và } L \quad (9)$$

đều cho ra cùng một đại diện. Phần chung (communality), được tạo ra bởi các thành phần nằm trên đường chéo của $LL' = (L^*)(L^*)'$ không bị ảnh hưởng bởi việc chọn ma trận T .

Việc phân tích mô hình nhân tố được tiến hành bằng cách áp các điều kiện cho phép chúng ta ước lượng duy nhất một L và ψ . Ma trận hệ số tải sau đó được xoay (nhân lên bởi một ma trận trực giao), trong đó độ xoay được xác định dựa trên các tiêu chí “dễ giải thích” hơn. Khi thu được các hệ số tải và các phương sai riêng (specific variances), các nhân tố cũng được xác định, và các giá trị ước lượng (gọi là các *điểm nhân tố*, hay *factor scores*) cho chính các nhân tố cũng thường được xây dựng.

3. Các phương pháp ước lượng (Methods of Estimation)

Cho các quan sát x_1, x_2, \dots, x_n trên p biến tương quan, phân tích dữ kiện tìm kiếm câu trả lời cho câu hỏi “Liệu có mô hình nhân tố như ở (4), với một số lượng nhân tố nhỏ, có thể thể hiện đầy đủ dữ liệu ban đầu 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 (5).

Ký hiệu S là ma trận hiệp phương sai mẫu, là một ước lượng của phương sai tổng thể Σ mà ta không thể xác định được. Nếu các phần tử nằm ngoài đường chéo của ma trận S là nhỏ, hay nói cách khác, về cơ bản chúng bằng 0 trong ma trận tương quan mẫu R , các biến không có mối quan hệ với nhau, và việc phân tích dữ kiện lúc này sẽ không hữu ích. Trong những trường hợp này, các nhân tố riêng đóng vai trò chủ đạo, trong khi mục đích chính của phân tích dữ kiện là xác định một vài nhân tố chung quan trọng cho dữ liệu.

3.1. Phương pháp Phân tích thành phần chính (The Principal Component Method)

Nhắc lại về công thức Phân rã phổ

Phân rã phổ của một ma trận đối xứng A , kích thước $(k \times k)$ được cho bởi:

$$\underbrace{A}_{(k \times k)} = \lambda_1 \underbrace{e_1}_{(k \times 1)} \underbrace{e_1'}_{(1 \times k)} + \lambda_2 \underbrace{e_2}_{(k \times 1)} \underbrace{e_2'}_{(1 \times k)} + \dots + \lambda_k \underbrace{e_k}_{(k \times 1)} \underbrace{e_k'}_{(1 \times k)}$$

Với $\lambda_1, \lambda_2, \dots, \lambda_k$ là các trị riêng của A và e_1, e_2, \dots, e_k là các vector riêng chuẩn hoá. $e_i' e_i = 1$ với $i = 1, 2, 3, \dots, k$ và $e_i' e_j = 0$ với $i \neq j$.

Cho ma trận hiệp phương sai Σ có các cặp trị riêng - vector riêng là (λ_i, e_i) với $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$. Khi đó:

$$\Sigma = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \dots + \lambda_p e_p e_p'$$

$$\left[\sqrt{\lambda_1} e_1 : \sqrt{\lambda_2} e_2 : \dots : \sqrt{\lambda_p} e_p \right] \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \sqrt{\lambda_2} e_2' \\ \vdots \\ \sqrt{\lambda_p} e_p' \end{bmatrix} \quad (10)$$

Ta nhận thấy công thức trên có nét tương đồng với cấu trúc hiệp phương sai được cho bởi mô hình phân tích dữ kiện với số lượng nhân tố bằng số lượng biến, tức là $m = p$, và các phương sai riêng $\psi_i = 0$ với mọi i . Ma trận hệ số tải có cột thứ j được cho bởi các $\sqrt{\lambda_j} e_j$. Ta có thể viết:

$$\underbrace{\Sigma}_{(p \times p)} = \underbrace{L}_{(p \times p)} \underbrace{L'}_{(p \times p)} + \underbrace{0}_{(p \times p)} = LL' \quad (11)$$

Ngoài nhân tố tỉ lệ $\sqrt{\lambda_j}$, thì các hệ số tải nhân tố của nhân tố thứ j cũng là các hệ số cho thành phần chính (principal component) của tổng thể (population).

Mặc dù mô hình (11) là chính xác, nhưng thực tế nó không hữu ích vì nó sử dụng số lượng nhân tố bằng với số lượng các biến và các nhân tố riêng ϵ là không đổi (bằng 0). Trong khi chúng ta mong muốn có được một mô hình có thể giải thích cấu trúc hiệp phương sai với số nhân tố ít hơn số biến.

Có một cách tiếp cận đó là khi có m cặp trị riêng (và vector riêng) đóng góp vào ma trận hiệp phương sai, $p - m$ cặp giá trị riêng còn lại mang giá trị nhỏ, thì

$\lambda_{m+1}\mathbf{e}_{m+1}\mathbf{e}'_{m+1} + \dots + \lambda_p\mathbf{e}_p\mathbf{e}'_p$ đóng góp không đáng kể vào Σ trong (10). Vì vậy, ta có thể bỏ qua sự đóng góp không đáng kể này, và xấp xỉ:

$$\Sigma = [\sqrt{\lambda_1}\mathbf{e}_1 : \sqrt{\lambda_2}\mathbf{e}_2 : \dots : \sqrt{\lambda_m}\mathbf{e}_m] \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}'_1 \\ \sqrt{\lambda_2}\mathbf{e}'_2 \\ \vdots \\ \sqrt{\lambda_m}\mathbf{e}'_m \end{bmatrix} = \underset{(p \times m)}{\mathbf{L}} \underset{(m \times p)}{\mathbf{L}'} \quad (12)$$

Khi có thêm các nhân tố riêng (specific factors):

$$\Sigma = [\sqrt{\lambda_1}\mathbf{e}_1 : \sqrt{\lambda_2}\mathbf{e}_2 : \dots : \sqrt{\lambda_m}\mathbf{e}_m] \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}'_1 \\ \sqrt{\lambda_2}\mathbf{e}'_2 \\ \vdots \\ \sqrt{\lambda_m}\mathbf{e}'_m \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_p \end{bmatrix} \quad (13)$$

Với $\psi_i = \sigma_{ii} - \sum_{j=1}^m l_{ij}^2$; $i = 1, 2, \dots, p$.

Để áp dụng cách tiếp cận này cho tập dữ liệu $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, đầu tiên ta tịnh tiến các biến về tâm quan sát bằng cách trừ chúng cho trung bình mẫu $\bar{\mathbf{x}}$.

$$\mathbf{x}_j - \bar{\mathbf{x}} = \begin{bmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jp} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} x_{j1} - \bar{x}_1 \\ x_{j2} - \bar{x}_2 \\ \vdots \\ x_{jp} - \bar{x}_p \end{bmatrix}, j = 1, 2, \dots, n \quad (14)$$

có cùng ma trận hiệp phương sai mẫu \mathbf{S} với các quan sát ban đầu.

Trong nhiều trường hợp, để tránh vấn đề một số biến có phương sai lớn làm ảnh hưởng quá mức đến việc xác định các hệ số tải, nên ta sẽ chuẩn hoá các biến:

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

Nội dung trình bày ở (13), khi áp dụng cho ma trận phương sai mẫu \mathbf{S} hoặc ma trận tương quan \mathbf{R} , được xem như là lời giải của phân tích thành phần chính. Cái tên này xuất phát từ thực tế là các hệ số tải nhân tố là các hệ số tỉ lệ của những thành phần chính đầu tiên của mẫu.

**Lời giải Phân tích thành phần chính cho mô hình nhân tố
(Principal Component Solution of the Factor Model)**

Phân tích dữ kiện dựa trên thành phần chính của ma trận hiệp phương sai mẫu \mathbf{S} bởi các cặp trị riêng – vector riêng $(\hat{\lambda}_1, \hat{\mathbf{e}}_1), (\hat{\lambda}_2, \hat{\mathbf{e}}_2), \dots, (\hat{\lambda}_p, \hat{\mathbf{e}}_p)$, với $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p$. Cho $m < p$ là số nhân tố chung. Khi đó, ma trận xấp xỉ hệ số tải $\{\tilde{t}_{ij}\}$ là:

$$\tilde{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1} \hat{\mathbf{e}}_1 : \sqrt{\hat{\lambda}_2} \hat{\mathbf{e}}_2 : \dots : \sqrt{\hat{\lambda}_m} \hat{\mathbf{e}}_m \right]$$

Xấp xỉ các phương sai riêng được tạo bởi các phần tử đường chéo của ma trận $\mathbf{S} - \tilde{\mathbf{L}} \tilde{\mathbf{L}}'$, do đó:

$$\tilde{\boldsymbol{\psi}} = \begin{bmatrix} \tilde{\psi}_1 & 0 & \dots & 0 \\ 0 & \tilde{\psi}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \tilde{\psi}_p \end{bmatrix} \quad \text{với } \tilde{\psi}_i = s_{ii} - \sum_{j=1}^m \tilde{t}_{ij}^2 \quad (16)$$

Phần chung (communality) được xấp xỉ:

$$\tilde{h}_i^2 = \tilde{t}_{i1}^2 + \tilde{t}_{i2}^2 + \dots + \tilde{t}_{im}^2$$

Phân tích dữ kiện bằng thành phần chính của ma trận tương quan mẫu thu được bằng cách sử dụng \mathbf{R} thay cho \mathbf{S} . (17)

Theo định nghĩa của $\tilde{\psi}_i$, các phần tử trên đường chéo của \mathbf{S} bằng với các phần tử trên đường chéo của $\tilde{\mathbf{L}} \tilde{\mathbf{L}}' + \tilde{\boldsymbol{\psi}}$. Tuy nhiên, các phần tử không thuộc

đường chéo của hai ma trận này thì lại thường không bằng nhau. Vậy làm sao để xác định được giá trị m , tức là số lượng nhân tố?

Việc chọn m có thể dựa trên việc xấp xỉ các trị riêng theo cách tương tự như các thành phần chính. Xét *ma trận dư* (*residual matrix*)

$$\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\boldsymbol{\Psi}}) \quad (18)$$

là kết quả từ việc xấp xỉ \mathbf{S} bằng phương pháp giải theo phân tích thành phần chính ở trên. Các phần tử trên đường chéo của ma trận này đều bằng 0. Và nếu các phần tử ở các vị trí khác cũng nhỏ, thì chúng ta có thể dùng mô hình m nhân tố để xấp xỉ. Ta có:

$$\text{Tổng bình phương các phần tử của } (\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\boldsymbol{\Psi}})) \leq \hat{\lambda}_{m+1}^2 + \dots + \hat{\lambda}_p^2 \quad (19)$$

Do đó, một giá trị nhỏ của tổng các bình phương của các trị riêng được bỏ qua, và giá trị này có thể được xem là giá trị độ lỗi bình phương của phép xấp xỉ.

Một cách lý tưởng, sự đóng góp của một vài nhân tố đầu tiên đến các phương sai mẫu của các biến là lớn. Sự đóng góp vào phương sai mẫu s_{ii} từ nhân tố chung thứ nhất là \tilde{l}_{i1}^2 .

Sự đóng góp vào tổng phương sai mẫu $s_{11} + s_{22} + \dots + s_{pp} = \text{tr}(\mathbf{S})$ từ nhân tố chung thứ nhất sẽ là:

$$\tilde{l}_{11}^2 + \tilde{l}_{21}^2 + \dots + \tilde{l}_{p1}^2 = \left(\sqrt{\hat{\lambda}_1} \mathbf{\hat{e}}_1 \right)' \left(\sqrt{\hat{\lambda}_1} \mathbf{\hat{e}}_1 \right) = \hat{\lambda}_1$$

vì vector riêng $\mathbf{\hat{e}}_1$ có độ dài đơn vị. Gọi K là tỉ lệ của tổng phương sai mẫu đóng góp bởi nhân tố thứ j , thông thường:

$$K = \begin{cases} \frac{\hat{\lambda}_j}{s_{11} + s_{22} + \dots + s_{pp}} & \text{(phân tích cho } \mathbf{S}) \\ \frac{\hat{\lambda}_j}{p} & \text{(phân tích cho } \mathbf{R}) \end{cases} \quad (20)$$

Tiêu chí (20) thường được sử dụng như một công cụ heuristic (dựa theo kinh nghiệm – thử và sai) cho việc xác định xấp xỉ số lượng nhân tố chung. Số nhân tố chung được giữ lại trong mô hình sẽ tăng lên cho đến khi tìm được một tỉ lệ thông tin phù hợp (suitable proportion) giải thích tổng phương sai mẫu.

Một quy ước khác thường được sử dụng trong các chương trình máy tính là cho m bằng với số lượng trị riêng của \mathbf{R} lớn hơn 1 nếu phân tích theo ma trận tương quan mẫu, hoặc bằng với số lượng trị riêng dương của \mathbf{S} nếu phân tích theo ma trận hiệp phương sai mẫu.

Ví dụ 3 (Phân tích dữ kiện cho dữ liệu mức độ ưu tiên của người tiêu dùng)

Trong một nghiên cứu về mức độ ưu tiên của người tiêu dùng, một số lượng người tiêu dùng được chọn ngẫu nhiên và được khảo sát để đánh giá một vài thuộc tính của một sản phẩm mới. Kết quả đánh giá theo thang điểm 7 đã được chuẩn hoá tỉ lệ, ta xây dựng được ma trận tương quan như dưới đây:

	1	2	3	4	5
1	1.0	.02	.96	.42	.01
2	.02	1.0	.13	.71	.85
3	.96	.13	1.0	.50	.11
4	.42	.71	.50	1.0	.79
5	.01	.85	.11	.79	1.0

Với các biến 1, 2, 3, 4, 5 lần lượt là các tiêu chí Mùi vị, Giá cả phải chăng, Hương vị, Sự thích hợp để ăn vặt và Cung cấp nhiều năng lượng.

Đầu tiên ta sẽ quan sát tổng quát ma trận trên để đưa ra một số nhận định ban đầu. Dễ dàng nhận thấy rằng biến 1 và biến 3 có mối tương quan lớn (hệ số tương quan 0.96), biến 2 và biến 5 cũng vậy (hệ số tương quan 0.85) nên rất có thể đây là hai nhóm (1, 3) và (2, 5). Còn lại biến 4, biến 4 có hệ số tương quan với các biến 1 và 3 lần lượt là 0.42 và 0.50, còn với các biến 2 và 5 thì lần lượt là 0.71 và 0.79. Rõ ràng mối tương quan giữa biến 4 với nhóm (2, 5) là cao hơn so với nhóm (1, 3). Với những nhận định ban đầu này, ta hi vọng rằng mối quan hệ tuyến tính giữa các biến có thể được thể hiện bằng hai hoặc ba nhân tố chung.

Hai trị riêng $\hat{\lambda}_1 = 2.85$ và $\hat{\lambda}_2 = 1.81$ của \mathbf{R} là các trị riêng duy nhất lớn hơn 1. Hơn nữa, $m = 2$ nhân tố chung chiếm tỉ lệ tích lũy

$$\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p} = \frac{2.85 + 1.81}{5} = 0.93$$

của tổng (chuẩn hoá) phương sai mẫu. Các giá trị xấp xỉ của hệ số tải, phần chung, phương sai riêng được tính theo (15), (16), (17) có kết quả như bên dưới

Biến quan sát (variable)	Xấp xỉ hệ số tải (factor loadings) $\tilde{l}_{ij} = \sqrt{\lambda_i} \hat{e}_{ij}$		Phần chung (communalities) \tilde{h}_i^2	Phương sai riêng (specific variances) $\tilde{\psi}_i = 1 - \tilde{h}_i^2$
	F_1	F_2		
1. Mùi vị	0.56	0.82	0.98	0.02
2. Giá cả phải chăng	0.78	-0.53	0.88	0.12
3. Hương vị	0.65	0.75	0.98	0.02
4. Thích hợp ăn vật	0.94	-0.10	0.89	0.11
5. Cung cấp nhiều năng lượng	0.80	-0.54	0.93	0.07
Trị riêng	2.85	1.81		
Tỉ lệ tích lũy (chuẩn hoá) của tổng phương sai mẫu	0.571	0.932		

Ta sẽ kiểm tra xem các kết quả trên có được chấp nhận hay không bằng cách tái lập lại ma trận tương quan \mathbf{R}

Ta có:

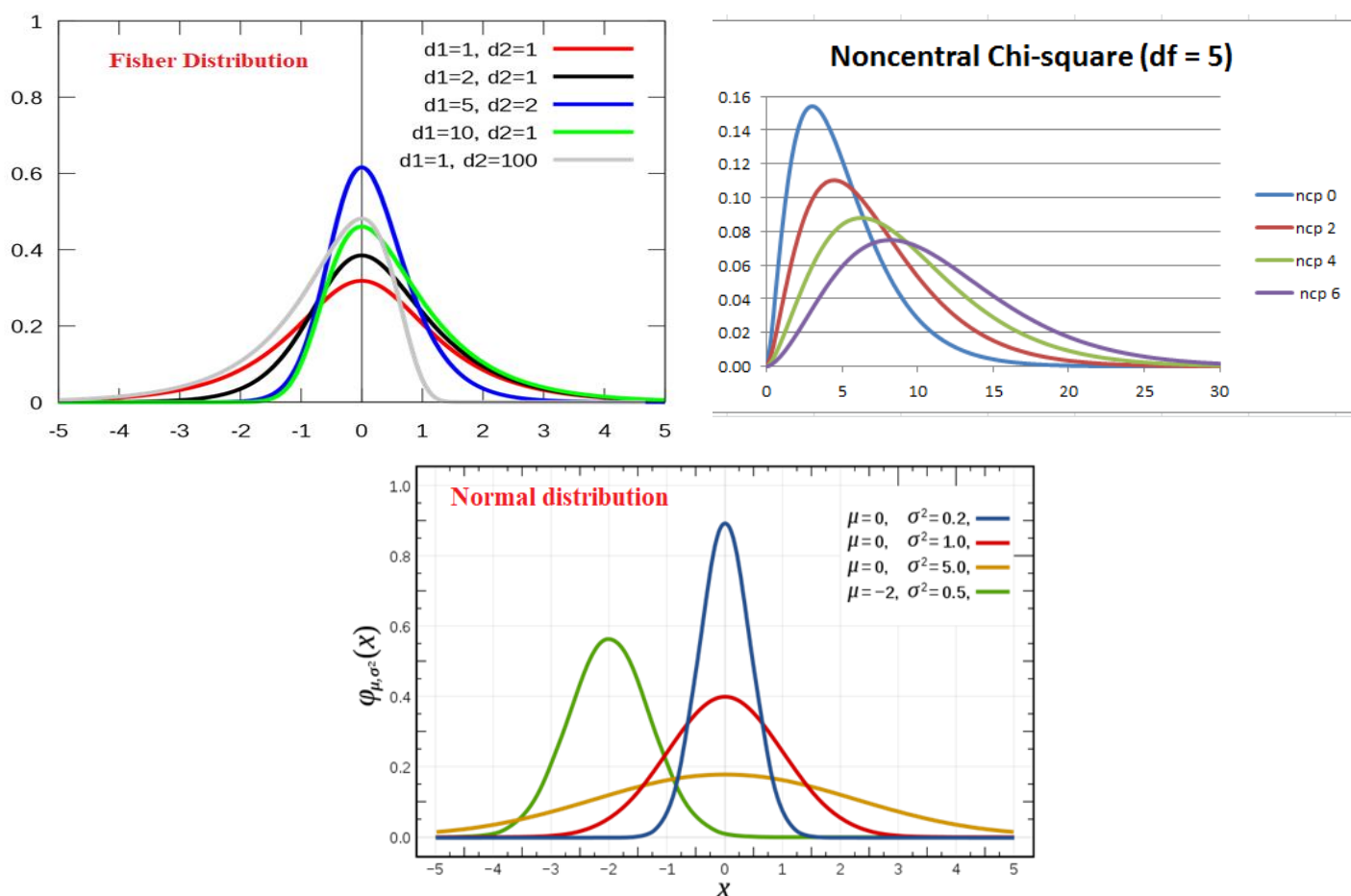
$$\begin{aligned}
\tilde{\mathbf{L}} \tilde{\mathbf{L}}' + \tilde{\boldsymbol{\Psi}} &= \begin{bmatrix} \tilde{l}_{11} & \tilde{l}_{12} \\ \tilde{l}_{21} & \tilde{l}_{22} \\ \tilde{l}_{31} & \tilde{l}_{32} \\ \tilde{l}_{41} & \tilde{l}_{42} \\ \tilde{l}_{51} & \tilde{l}_{52} \end{bmatrix} \begin{bmatrix} \tilde{l}_{11} & \tilde{l}_{21} & \tilde{l}_{31} & \tilde{l}_{41} & \tilde{l}_{51} \\ \tilde{l}_{12} & \tilde{l}_{22} & \tilde{l}_{32} & \tilde{l}_{42} & \tilde{l}_{52} \end{bmatrix} + \begin{bmatrix} \tilde{\psi}_1 & 0 & 0 & 0 & 0 \\ 0 & \tilde{\psi}_2 & 0 & 0 & 0 \\ 0 & 0 & \tilde{\psi}_3 & 0 & 0 \\ 0 & 0 & 0 & \tilde{\psi}_4 & 0 \\ 0 & 0 & 0 & 0 & \tilde{\psi}_5 \end{bmatrix} \\
&= \begin{bmatrix} .56 & .82 \\ .78 & -.53 \\ .65 & .75 \\ .94 & -.10 \\ .80 & -.54 \end{bmatrix} \begin{bmatrix} .56 & .78 & .65 & .94 & .80 \\ .82 & -.53 & .75 & -.10 & -.54 \end{bmatrix} + \begin{bmatrix} .02 & 0 & 0 & 0 & 0 \\ 0 & .12 & 0 & 0 & 0 \\ 0 & 0 & .02 & 0 & 0 \\ 0 & 0 & 0 & .11 & 0 \\ 0 & 0 & 0 & 0 & .07 \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{1.0} & .01 & .97 & .44 & .00 \\ .01 & \mathbf{1.0} & .11 & .79 & .91 \\ .97 & .11 & \mathbf{1.0} & .53 & .11 \\ .44 & .79 & .53 & \mathbf{1.0} & .81 \\ .00 & .91 & .11 & .81 & \mathbf{1.0} \end{bmatrix}, \text{ có giống ma trận } \begin{bmatrix} 1.0 & .02 & .96 & .42 & .01 \\ .02 & 1.0 & .13 & .71 & .85 \\ .96 & .13 & 1.0 & .50 & .11 \\ .42 & .71 & .50 & 1.0 & .79 \\ .01 & .85 & .11 & .79 & 1.0 \end{bmatrix}
\end{aligned}$$

không? Rất giống!!!

Do đó, trên cơ sở mô tả thuần túy thì chúng ta có thể đánh giá rằng mô hình hai nhân tố với các hệ số tải đã tính được ở trên là phù hợp với dữ liệu. Các giá trị phần chung – commutativities (0.98, 0.88, 0.98, 0.89, 0.93) chỉ ra rằng hai nhân tố này chiếm một tỉ lệ lớn trong phương sai mẫu của mỗi biến.

3.2. Phương pháp Triển vọng cực đại (The Maximum Likelihood Method)

Trong thực tế, dữ liệu mà ta có được không có đầy đủ các thống kê cần thiết để áp dụng cho phân phối chuẩn, như trung bình tổng thể μ chẳng hạn. Do đó, ta phải dùng các phân phối khác gần giống phân phối chuẩn. Ví dụ, đồ thị của phân phối Fisher và phân phối Chi Square dưới đây có hình dạng gần giống với đồ thị của phân phối chuẩn.



Nội hàm của phương pháp này là: Những gì ta thấy được trong thực nghiệm thì phải dễ xảy ra hơn là những gì ta không thấy.

Mặc dù những yếu tố “thiên nga đen” (những thứ có xác suất xảy ra thấp nhưng có tác động lớn nếu nó xảy ra) là thực sự có. Nhưng nội hàm trên là hợp lý vì các yếu

tổ “thiên nga đen” – đúng với cái tên của nó – là cực kì hiếm.

Dữ liệu mà chúng ta có được chính là những thứ dễ xảy ra. Do đó xác suất của nó phải cao hơn những sự kiện mà ta không quan sát được. Điều này có nghĩa là ta đã biết trước kết quả và cần tìm tham số mô hình cũng như nguyên nhân dẫn đến kết quả này.

Giả sử \mathbf{X} có phân phối xác suất P_{θ} phụ thuộc vào bộ $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$. Mục tiêu của chúng ta là cực đại hàm triển vọng

$$Likelihood(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) = p_{\theta}(x_1) \dots p_{\theta}(x_n)$$

Nếu các nhân tố chung \mathbf{F} và các nhân tố riêng $\boldsymbol{\varepsilon}$ được coi là có phân phối chuẩn, thì sẽ tồn tại phép ước lượng triển vọng cực đại cho các hệ số tải nhân tố và các phương sai riêng. F_j và ε_j có phân phối chuẩn nên $\mathbf{X}_j - \boldsymbol{\mu}_j = \mathbf{L}\mathbf{F}_j + \boldsymbol{\varepsilon}_j$ cũng có phân phối chuẩn. Hàm triển vọng

$$\begin{aligned} L(\boldsymbol{\mu}, \Sigma) &= (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} e^{-\left(\frac{1}{2}\right)tr\left[\Sigma^{-1}\left(\sum_{j=1}^n (x_j - \bar{x})(x_j - \bar{x})' + n(\bar{x} - \boldsymbol{\mu})(\bar{x} - \boldsymbol{\mu})'\right)\right]} \\ &= (2\pi)^{-\frac{(n-1)p}{2}} |\Sigma|^{-\frac{(n-1)}{2}} e^{-\left(\frac{1}{2}\right)tr\left[\Sigma^{-1}\left(\sum_{j=1}^n (x_j - \bar{x})(x_j - \bar{x})'\right)\right]} \\ &\times (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\left(\frac{n}{2}\right)(\bar{x} - \boldsymbol{\mu})'\Sigma^{-1}(\bar{x} - \boldsymbol{\mu})} \end{aligned} \quad (21)$$

Công thức trên phụ thuộc vào \mathbf{L} và $\boldsymbol{\psi}$ thông qua mối quan hệ $\Sigma = \mathbf{L}\mathbf{L}' + \boldsymbol{\psi}$. Mô hình này vẫn chưa được định nghĩa rõ, bởi vì có nhiều cách chọn \mathbf{L} có được từ phép biến đổi trực giao. Điều chúng ta mong muốn là làm cho \mathbf{L} được xác định rõ ràng bằng cách đặt điều kiện về tính duy nhất để thuận tiện cho việc tính toán:

$$\mathbf{L}'\boldsymbol{\psi}^{-1}\mathbf{L} = \Delta \text{ là ma trận chéo} \quad (22)$$

Ước lượng triển vọng cực đại $\hat{\mathbf{L}}$ và $\hat{\boldsymbol{\psi}}$ thu được bằng cách cực đại số học biểu thức (21). Các chương trình máy tính hiện nay cho phép chúng ta thực hiện phép ước lượng dễ dàng hơn.

Lược đồ tính toán đề xuất (Recommended Computational Scheme)

Với $m > 1$, điều kiện $\mathbf{L}'\boldsymbol{\psi}^{-1}\mathbf{L} = \Delta$, một ràng buộc hiệu quả $\frac{m(m-1)}{2}$ áp đặt lên các phần tử của \mathbf{L} và $\boldsymbol{\psi}$. Phương pháp giải:

1. Tính ước lượng khởi tạo của phương sai riêng $\psi_1, \psi_2, \dots, \psi_p$:

$$\widehat{\psi}_i = \left(1 - \frac{1}{2} \times \frac{m}{p}\right) \left(\frac{1}{s^{ii}}\right)$$

Với s^{ii} là phần tử thứ i trên đường chéo của ma trận S^{-1} .

2. Cho $\hat{\psi}$, tính m trị riêng đầu tiên $\hat{\lambda}_1 > \hat{\lambda}_2 > \dots > \hat{\lambda}_m > 1$ và các vector riêng $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_m$ của ma trận

$$\mathbf{S}^* = \hat{\psi}^{-\frac{1}{2}} \mathbf{S}_n \hat{\psi}^{-\frac{1}{2}}$$

Cho $\hat{\mathbf{E}} = [\hat{e}_1 : \hat{e}_2 : \dots : \hat{e}_m]$ là ma trận $p \times m$ của các vector riêng chuẩn hoá và $\hat{\mathbf{\Lambda}} = \text{diag}[\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_m]$ là ma trận đường chéo $m \times m$ của các trị riêng. Đồng thời ta có được kết quả $\hat{\mathbf{\Lambda}} = \mathbf{I} + \hat{\mathbf{\Delta}}$ và $\hat{\mathbf{E}} = \hat{\psi}^{-\frac{1}{2}} \hat{\mathbf{L}} \hat{\mathbf{\Delta}}^{-\frac{1}{2}}$ (chứng minh trong [1] phụ lục 9A). Ta có ước lượng:

$$\hat{\mathbf{L}} = \hat{\psi}^{\frac{1}{2}} \hat{\mathbf{E}} \hat{\mathbf{\Delta}}^{\frac{1}{2}} = \hat{\psi}^{\frac{1}{2}} \hat{\mathbf{E}} (\hat{\mathbf{\Lambda}} - \mathbf{I})^{\frac{1}{2}}$$

3. Từ \mathbf{L} có được ở bước trên, ta tính ψ bằng cách lấy ma trận đường chéo của \mathbf{S} trừ cho $\hat{\mathbf{L}} \hat{\mathbf{L}}'$. Lặp lại bước 2 và 3 cho đến khi $\hat{\mathbf{L}}$ và $\hat{\psi}$ hội tụ.

3.3. Kiểm định số nhân tố cho mẫu lớn (A Large Sample Test for the Number of Common Factors)

Ban đầu, chúng ta đã giả định tổng thể có phân phối chuẩn. Điều này dẫn đến việc ta phải kiểm định xem mô hình chúng ta đưa ra có phù hợp với mục đích thống kê hay không. Nghĩa là ta sẽ đi kiểm định giả thuyết H_0 cho mô hình m nhân tố chung $\Sigma = \mathbf{L}\mathbf{L}' + \psi$

$$H_0: \underset{(p \times p)}{\Sigma} = \underset{(p \times m)}{\mathbf{L}} \underset{(m \times p)}{\mathbf{L}}' + \underset{(p \times p)}{\psi} \quad (23)$$

Và đối thuyết: $H_1: \Sigma$ là ma trận dương xác định khác.

Nhắc lại

a) Cực đại hàm triển vọng:

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

với $|\hat{\Sigma}| = \left[\frac{n-1}{n} \right]^p |\mathbf{S}|$.

b) Kết quả thừa nhận:

Cho X_1, X_2, \dots, X_n là một ngẫu nhiên có được từ một tổng thể thuộc phân phối chuẩn với trung bình μ và hiệp phương sai Σ . Khi đó:

$$\hat{\mu} = \bar{\mathbf{X}} \text{ và } \hat{\Sigma} = \frac{1}{n} \sum_{j=1}^n (\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})' = \frac{(n-1)}{n} \mathbf{S}$$

là ước lượng triển vọng cực đại của μ và Σ

c) Kiểm định cho trung bình tổng thể có phân phối chuẩn

$$H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0; H_1: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

Bác bỏ H_0 , chấp nhận H_1 với mức ý nghĩa α , nếu:

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\mathbf{s}^2)^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) > t_{n-1}^2\left(\frac{\alpha}{2}\right)$$

Khi Σ không có gì đặc biệt khác, thì cực đại hàm triển vọng của nó tỉ lệ thuận với

$$|\mathbf{S}_n|^{-\frac{n}{2}} e^{-\frac{np}{2}} \quad (24)$$

Xét giả thuyết H_0 ở (23), cực đại hàm triển vọng tỉ lệ với

$$\begin{aligned} & |\hat{\Sigma}|^{-\frac{n}{2}} \exp \left(-\frac{1}{2} \text{tr} \left[\hat{\Sigma}^{-1} \left(\sum_{j=1}^n (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})' \right) \right] \right) \\ &= |\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}}|^{-\frac{n}{2}} \exp \left(-\frac{1}{2} n \text{tr} [(\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}})^{-1} \mathbf{S}_n] \right) \end{aligned} \quad (25)$$

Sử dụng Nhắc lại (c), (24) và (25), ta tìm được tỉ lệ triển vọng thống kê cho kiểm định H_0 là:

$$\begin{aligned} -2 \ln \Lambda &= -2 \ln \left[\frac{\text{maximized likelihood under } H_0}{\text{maximized likelihood}} \right] \\ &= -2 \ln \left(\frac{|\hat{\Sigma}|}{|\mathbf{S}_n|} \right)^{-\frac{n}{2}} + n [\text{tr}(\hat{\Sigma}^{-1} \mathbf{S}_n) - p] \end{aligned} \quad (26)$$

Với bậc tự do

$$\begin{aligned} v - v_0 &= \frac{1}{2} p(p+1) - \left[p(m+1) - \frac{1}{2} m(m-1) \right] \\ &= \frac{1}{2} [(p-m)^2 - p - m] \end{aligned} \quad (27)$$

Do $\text{tr}(\hat{\Sigma}^{-1} \mathbf{S}_n) - p = 0$ (xem thêm phần phụ lục 9A sách AMSA), nên ta có;

$$-2 \ln \Lambda = n \ln \left(\frac{|\hat{\Sigma}|}{|\mathbf{S}_n|} \right) \quad (28)$$

Sử dụng hiệu chỉnh Bartlett, ta bác bỏ H_0 với độ tin cậy α nếu:

$$\left(n - 1 - \frac{2p + 4m + 5}{6}\right) \ln \frac{|\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\psi}}|}{|\mathbf{S}_n|} > \chi^2_{\frac{(p-m)^2 - p - m}{2}}(\alpha) \quad (29)$$

khi n và $n - p$ lớn.

Do số bậc tự do $\frac{1}{2}[(p - m)^2 - p - m]$ phải là một số dương, nên kiểm định (29) phải thoả điều kiện

$$m < \frac{1}{2}(2p + 1 - \sqrt{8p + 1}) \quad (30)$$

Nhận xét: Khi thực hiện kiểm định (29), chúng ta kiểm định tính đầy đủ (phù hợp) của mô hình m nhân tố bằng cách so sánh các phương sai tổng quát hoá $|\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\psi}}|$ và $|\mathbf{S}_n|$. Nếu n lớn và m nhỏ so với p , tập giả thuyết H_0 sẽ thường bị bác bỏ, dẫn đến việc giữ lại nhiều nhân tố chung hơn. Tuy nhiên $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\psi}}$ có thể đủ “gần” \mathbf{S}_n khiến cho việc có thêm nhiều nhân tố hơn cũng không cung cấp thêm được thông tin gì hữu ích, mặc dù những nhân tố đó là “quan trọng”. Vì thế, cần có một số phán đoán trong việc lựa chọn giá trị m .

4. Xoay nhân tố (Factor Rotation)

Như đã biết, các hệ số tải nhân tố nhận được từ phép biến đổi trực giao ban đầu đều có thể tạo lại ma trận hiệp phương sai (hoặc ma trận tương quan). Trong đại số ma trận, ta biết rằng phép biến đổi trực giao tuân theo phép xoay trục tọa độ. Vì lí do này, phép biến đổi trực giao của hệ số tải nhân tố được gọi là *xoay nhân tố*.

Nếu $\hat{\mathbf{L}}$ là ma trận $p \times m$ ước lượng hệ số tải có được từ bất kì phương pháp nào (principal component, maximum likelihood, v.v..) thì

$$\hat{\mathbf{L}}^* = \hat{\mathbf{L}}\mathbf{T}, \quad \text{với } \mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$$

là ma trận hệ số tải đã được xoay. Ngoài ra, ma trận hiệp phương sai (hay hệ số ma trận) ước lượng được vẫn không đổi, do

$$\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\psi}} = \hat{\mathbf{L}}\mathbf{T}\mathbf{T}'\hat{\mathbf{L}} + \hat{\boldsymbol{\psi}} = \hat{\mathbf{L}}^*\hat{\mathbf{L}}^{*'} + \hat{\boldsymbol{\psi}}$$

Phương trình trên chỉ ra rằng ma trận còn sót lại $\mathbf{S}_n - \hat{\mathbf{L}}\hat{\mathbf{L}}' - \hat{\boldsymbol{\psi}} = \mathbf{S}_n - \hat{\mathbf{L}}^*\hat{\mathbf{L}}^{*'} - \hat{\boldsymbol{\psi}}$ vẫn không đổi. Hơn nữa, phương sai $\hat{\boldsymbol{\psi}}$ và công cộng $\widehat{\mathbf{h}}_i^2$ đều không đổi. Nên dưới góc nhìn toán học thì việc thu được $\hat{\mathbf{L}}$ hay $\hat{\mathbf{L}}^*$ đều không quan trọng.

Do các tải ban đầu có thể khó diễn đạt, ta thường xoay chúng để đạt được cấu trúc đơn giản hơn.

Chúng ta muốn thấy mẫu các tải sao cho từng biến có tải lớn trên một nhân tố nhất định và tải thấp đến trung bình trên các nhân tố còn lại. Tuy nhiên, không phải lúc nào cũng đạt được cấu trúc đơn giản như vậy.

Chúng ta sẽ tập trung dùng các phương pháp đồ thị và phân tích để xác định một phép xoay trực giao để đạt cấu trúc đơn giản hơn. Khi $m = 2$, hay các nhân tố chung được cho là 2 trong 1 lược, sự chuyển đổi về cấu trúc giản đơn thường được xác định bằng đồ thị. Những nhân tố chung không liên quan được xem là các vector đơn vị trên các trục vuông góc. Biểu đồ của cặp hệ số tải nhân tố $(\hat{l}_{i1}, \hat{l}_{i2})$ cho p điểm, mỗi điểm tương ứng với một biến. Sau đó các trục tọa độ có thể được xoay theo một góc, đặt tên là Φ , và các tải được xoay mới \hat{l}_{ij}^* được xác định từ các mối quan hệ (*).

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

Nếu xoay theo chiều kim đồng hồ thì:

$$\mathbf{T} = \begin{bmatrix} \cos \phi & \sin \phi \\ -\sin \phi & \cos \phi \end{bmatrix}$$

Nếu xoay ngược chiều kim đồng hồ thì:

$$\mathbf{T} = \begin{bmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{bmatrix}$$

Mối quan hệ trên hiếm khi được sử dụng cho phân tích đồ thị trong không gian hai chiều. Lúc này, các cụm biến thường có thể quan sát được bởi mắt thường, và những cụm giúp nhận diện các nhân tố chung mà không cần phải kiểm tra cường độ của những tải được xoay. Đối với $m > 2$, không dễ hình dung các định hướng được, và cường độ của các tải được xoay phải được kiểm tra để có thể rút ra được kết luận có ý nghĩa từ dữ liệu gốc. Việc chọn ma trận trực giao \mathbf{T} thoả mãn một cấu trúc đơn giản sẽ được cân nhắc.

Ví dụ 4: Lawley và Maxwell trình bày ma trận tương quan mẫu của điểm số trong $p = 6$ nhóm ngành của $n = 220$ nam học sinh. Ma trận tương quan là

$$R = \begin{bmatrix} 1.0 & .439 & .410 & .288 & .329 & .248 \\ & 1.0 & .351 & .354 & .320 & .329 \\ & & 1.0 & .164 & .190 & .181 \\ & & & 1.0 & .595 & .470 \\ & & & & 1.0 & .464 \\ & & & & & 1.0 \end{bmatrix}$$

và kết quả triển vọng cực đại với $m = 2$ nhân tố chung cho ta các ước lượng tại bảng dưới

Variable	Estimated factor loadings		Communalities \widehat{h}_i^2
	F1	F2	
1. Gaelic (a)	.553	.429	.490
2. English (b)	.568	.288	.406
3. History (c)	.392	.450	.356
4. Arithmetic (d)	.740	-.273	.623
5. Algebra (e)	.724	-.211	.569
6. Geometry (f)	.595	-.132	.372

Tất cả các biến đều có tải dương đối với nhân tố đầu tiên. Lawley và Maxwell cho rằng nhân tố này thể hiện sự phản ứng tổng quan của học sinh đối với hướng dẫn và có thể được ghi nhận là nhân tố *tri thức tổng quát*. Nửa số tải là số dương và nửa còn lại đều âm đối với nhân tố thứ 2. Nhân tố với kết quả tải này được gọi là *nhân tố lưỡng cực*. (Việc quy ước cực âm và dương là chủ quan, các dấu của tải đối với một nhân tố có thể bị đảo ngược mà không ảnh hưởng đến quá trình phân tích) Nhân tố này khó xác định, mà các cá nhân được điểm cao hơn trung bình trong bài kiểm tra miệng cũng đồng thời được điểm cao hơn trung bình trong nhân tố này. Các cá nhân có điểm cao hơn trung bình trong bài kiểm tra toán được điểm thấp hơn trung bình trong nhân tố. *Nhân tố* này có thể được xem là nhân tố “toán - không toán”.

Cặp hệ số tải nhân tố $(\hat{l}_{i1}, \hat{l}_{i2})$ được biểu diễn thành các điểm trong đồ thị 9.1. Các điểm được đặt tên theo số của các biến tương ứng. Đồng thời được biểu hiện là một phép xoay trục giao của trục tọa độ theo góc 20° . Góc này được chọn để trục tọa độ mới có thể đi qua $(\hat{l}_{41}, \hat{l}_{42})$. Khi đó, tất cả các điểm đều nằm trong

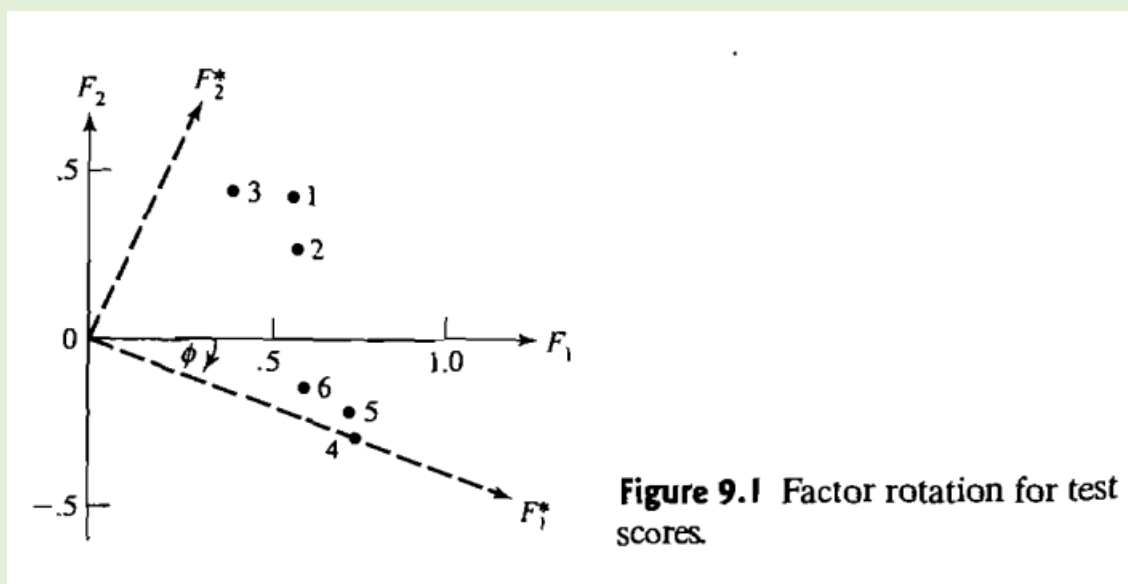
góc phần tư thứ nhất (mọi hệ số tải nhân tố đều dương), và các cụm biến dễ quan sát hơn.

Các biến liên quan đến bài kiểm tra toán có tải cao ở F_1^* và có tải không đáng kể tại F_2^* . Nhân tố đầu tiên có thể được đặt tên là *khả năng toán học*. Tương tự, 3 bài kiểm tra miệng có tải cao ở F_2^* và có tải tương đối tới thấp ở F_1^* . Nhân tố thứ 2 có thể được gọi là *khả năng ngôn ngữ*. Nhân tố *tri thức tổng quát* được xác định ban đầu được hòa vào trong nhân tố F_1^* và F_2^* .

Các hệ số tải nhân tố xoay có được từ (*) với góc 20° và các communality ước lượng tương ứng được thể hiện tại bảng dưới. Cường độ các hệ số tải nhân tố xoay củng cố cách dịch nghĩa nhân tố của hình dưới.

Ước lượng của communality không bị ảnh hưởng bởi phép xoay trục giao, vì $\hat{L}\hat{L}' = \hat{L}T T' \hat{L}' = \hat{L}^* \hat{L}^{*'}$, và các communalities là các thành phần chéo của các ma trận này.

Ta thấy rằng hình dưới có thể cho thấy một phép xoay nghiêng các tọa độ. Một trục tọa độ mới sẽ đi qua cụm {1, 2, 3} và trục còn lại sẽ đi qua {4, 5, 6}. Phép xoay nghiêng được đặt tên như vậy vì chúng tương ứng phép xoay *không cứng* của các trục tọa độ để cho ra các trục mới không vuông góc.



Variable	Estimated factor loadings		Communalities \widehat{h}_i^2
	F1	F2	
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

Tuy nhiên, có thể thấy rằng cách giải thích cho nhân tố nghiêng cho ví dụ này sẽ gần tương tự như cách giải thích đã đưa ra đối với xoay trực giao.

Kaiser đề xuất một phương pháp phân tích cấu trúc đơn giản có tên là *quy chuẩn varimax* (hoặc varimax thông thường). Xác định $\tilde{l}_{ij}^* = \frac{\tilde{l}_{ij}}{\hat{h}_i}$ là hệ số đã xoay tỉ lệ với căn của communalities. Khi đó quy trình varimax (thông thường) sẽ chọn ra phép biến đổi trực giao T mà cho ra

$$V = \frac{1}{p} \sum_{j=1}^m \left[\sum_{i=1}^p \tilde{l}_{ij}^{*4} - \frac{(\sum_{i=1}^p \tilde{l}_{ij}^{*2})^2}{p} \right]$$

lớn nhất có thể.

Phóng các hệ số đã xoay \tilde{l}_{ij}^* sẽ mang lại cho các biến có communalities nhỏ thêm ý nghĩa trong việc quyết định một cấu trúc đơn giản. Sau khi đã xác định phép biến đổi T, các tải \tilde{l}_{ij}^* sẽ được nhân lên \hat{h}_i lần để bảo toàn các communalities gốc.

Mặc dù phương trình trên trông phức tạp, nó có cách giải thích đơn giản. Trình bày bằng lời sẽ là

$$V \propto \sum_{j=1}^m (\text{phương sai bình phương của hệ số tải cho nhân tố thứ } j)$$

Tối đa hoá V đồng nghĩa với việc “phân tán” các bình phương nhiều nhất có thể của các tải trên từng nhân tố. Vì thế, chúng tôi hy vọng là có thể xác định được những nhóm hệ số lớn và vừa ở mọi cột của các tải đã xoay ở ma trận \tilde{L}^* .

Ta có những thuật toán máy tính để tối đa hoá V, và những phần mềm máy tính phổ biến nhất cho việc phân tích dữ kiện (các phần mềm xác suất SAS, SPSS, BMDP, và MINITAB) đều có thể xoay varimax. Phép xoay varimax các hệ số tải nhân tố có được từ các phương pháp (phân tích thành phần chính, triển vọng cực đại, ...) sẽ phần lớn không trùng nhau. Ngoài ra, các xu hướng của các tải đã xoay có thể thay đổi đáng kể nếu cho thêm nhân tố chung vào trong quy trình xoay. Nếu có một nhân tố đơn lớn, nó sẽ thường bị che khuất đi bởi bất kì phép xoay trực giao. Ngược lại, nó luôn có thể được cố định lại trong khi các nhân tố khác xoay.

Ví dụ 5 (Các tải được xoay cho dữ liệu consumer-preference) Các hệ số tải nhân tố gốc (thu được thông qua phương pháp phân tích thành phần chính), các communalities, và các hệ số tải nhân tố được xoay (varimax) được trình bày tại bảng dưới.

Variable	Estimated factor loadings		Rotated estimated factor loadings		Communalities \widehat{h}_i^2
	F1	F2	F1	F2	
1. Taste	.56	.82	0.2	.99	.98
2. Good buy for money	.78	-.52	.94	-.01	.88
3. Flavor	.65	.75	.13	.98	.98
4. Suitable for snack	.94	-.10	.84	.43	.89
5. Provides lots of energy	.80	-.54	.97	-.02	.93
Cumulative proportion of total (standardized) sample variance explained	.571	.932	.507	.932	

Ta thấy được là biến 2, 4, và 5 định nghĩa nhân tố 1 (tải cao trên nhân tố 1, nhỏ hay không đáng kể trên nhân tố 2), trong khi biến 1 và 3 định nghĩa nhân tố 2 (tải cao tại nhân tố 2, nhỏ hay không đáng kể tại nhân tố 1). Biến 4 gần tương tự đối với nhân tố 1, tuy nhiên nó cũng có một số ít điểm tương đồng đối với nhân tố 2. Ta có thể gọi nhân tố 1 là nhân tố *giá trị dinh dưỡng* và nhân tố 2 là nhân tố *vị*.

Các hệ số tải nhân tố cho các biến được hình dung dựa theo các trục tọa độ gốc và các trục tọa độ xoay nhân tố (varimax) ở hình dưới.

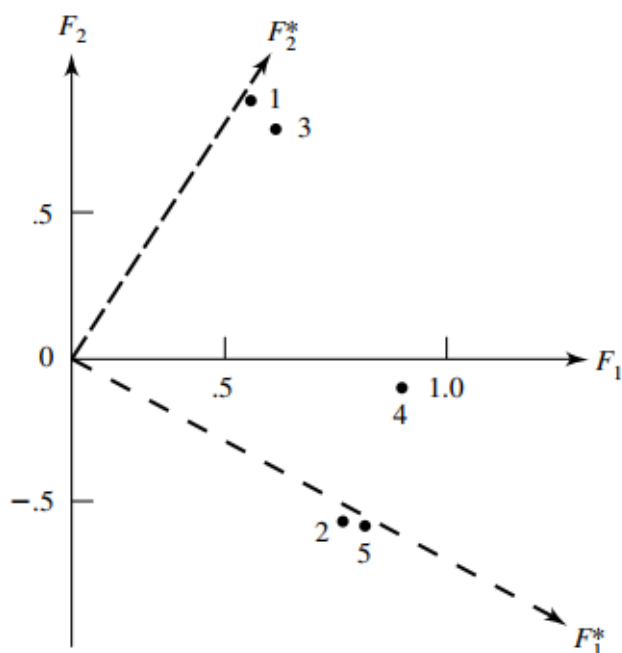


Figure 9.2 Factor rotation for hypothetical marketing data.

Việc xoay các hệ số tải nhân tố được khuyến khích sử dụng cụ thể cho các tải có được thông qua triển vọng cực đại, bởi các giá trị ban đầu đã được hạn chế để thỏa mãn điều kiện $\hat{L}'\hat{\psi}^{-1}\hat{L}$ là 1 ma trận chéo. Điều kiện này có ích cho việc lập trình, nhưng nó không giúp ích trong việc dịch và hiểu các nhân tố.

Xoay nghiêng

Các phép xoay trực giao thường phù hợp cho mẫu nhân tố có các nhân tố chung độc lập. Phép xoay nghiêng thường được dùng sau khi quan sát các hệ số tải nhân tố và không theo mẫu được công nhận.

Nếu ta cho m nhân tố chung là các trục tọa độ, điểm với m tọa độ $(\hat{l}_{i1}, \hat{l}_{i2}, \dots, \hat{l}_{im})$ thể hiện vị trí của biến thứ i trong *không gian nhân tố*. Cho rằng các biến được gộp vào các cụm không chồng lên nhau, một phép xoay trực giao về cấu trúc đơn giản hơn tương ứng với một phép xoay *cứng* của các trục tọa độ sao cho các trục sau khi xoay sẽ tiếp cận gần các cụm nhất có thể. Một phép xoay nghiêng về cấu trúc đơn giản hơn tương ứng với một phép di chuyển *không cứng* của hệ tọa độ sao cho các trục được xoay (không vuông góc nữa) đi qua các cụm. Phép xoay nghiêng thể hiện từng biến theo số nhân tố ít nhất có thể.

5. Điểm nhân tố (Factor Scores)

Trong phân tích dữ kiện, mỗi 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, được gọi là điểm nhân tố, cũng có thể được yêu cầu. Những đại lượng này thường được sử dụng cho mục đích chẩn đoán, cũng như đưa đến một phân tích tiếp theo.

Điểm nhân tố không phải là ước tính của các tham số chưa biết theo nghĩa thông thường. Thay vào đó, chúng là các ước tính giá trị cho các vector yếu tố ngẫu nhiên không được quan sát $F_j, j = 1, 2, \dots, n$. Đó là, điểm nhân tố

$$\hat{f}_j = \text{ước lượng giá trị } f_j \text{ đạt bởi } F_j \text{ (trường hợp thứ } j)$$

Việc ước lượng phức tạp là do các đại lượng không quan sát được f_j và ε_j nhiều hơn x_j quan sát được. Để khắc phục khó khăn này, chúng tôi mô tả hai trong số những cách tiếp cận dưới đây.

5.1. Phương pháp bình phương tối thiểu có trọng số (the weighted least squares method)

Trước tiên, giả sử rằng vector trung bình μ , hệ số tải L và phương sai cụ thể ψ được biết đến với mô hình nhân tố.

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

Hơn nữa, hãy xem xét các nhân tố riêng $\varepsilon = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p]$ là sai số. Do $Var(\varepsilon_i) = \psi_i, i = 1, 2, \dots, p$, không cần bằng nhau. Bartlett đã gợi ý rằng bình phương nhỏ nhất có trọng số được sử dụng để ước tính các giá trị nhân tố chung. Tổng bình phương của các lỗi, được tính theo nghịch đảo của phương sai, là

$$\sum_{i=1}^p \frac{\varepsilon_i^2}{\psi_i} = \varepsilon' \psi^{-1} \varepsilon = (x - \mu - Lf)' \psi^{-1} (x - \mu - Lf) \quad (31)$$

Bartlett đã phân tích chọn các ước lượng \hat{f} của f để cực tiểu (31) là

$$\hat{f} = (L' \psi^{-1} L)^{-1} L' \psi^{-1} (x - \mu) \quad (32)$$

Được thúc đẩy bởi (48), chúng ta lấy các ước tính \hat{L} , $\hat{\psi}$ và $\hat{\mu} = \bar{x}$ làm giá trị thực và lấy điểm nhân tố cho trường hợp thứ j :

$$\hat{f}_j = (\hat{L}'\hat{\psi}^{-1}\hat{L})^{-1}\hat{L}'\hat{\psi}^{-1}(x_j - \bar{x}) \quad (33)$$

Khi \hat{L} và $\hat{\psi}$ được xác định bằng phương pháp maximum likelihood, những ước tính này phải thỏa mãn điều kiện duy nhất, $\hat{L}'\hat{\psi}^{-1}\hat{L} = \hat{\Delta}$ một ma trận đường chéo.

Điểm nhân tố có được bởi phương pháp bình phương tối thiểu có trọng số từ ước lượng triển vọng cực đại

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

Đối với ma trận tương quan:

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

Trong đó $Z_j = D^{-1/2}(x - \bar{x})$, như trong (8-25), và $\hat{p} = \hat{L}_z \hat{L}'_z + \hat{\psi}_z$

(34)

Điểm nhân tố được tạo bởi (34) có vectơ trung bình mẫu $\mathbf{0}$ và mẫu không hiệp phương sai.

Nếu các tải $\hat{L}^* = \hat{L}T$ được sử dụng thay cho các tải ban đầu ở (34) các điểm nhân tố tiếp theo \hat{f}^*_j liên quan tới \hat{f}_j bởi $\hat{f}^*_j = T' \hat{f}_j$, $j=1,2,\dots,n$.

Nếu hệ số tải được ước tính bởi thành phần chính, phương pháp này thường tạo ra điểm nhân tố bằng cách sử dụng một thủ tục bình phương nhỏ nhất. Rõ ràng, điều này có nghĩa là giả định rằng ψ bằng hoặc gần bằng nhau. Điểm nhân tố sau đó là

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

Hoặc

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

Cho dữ liệu chuẩn hóa. Từ $\tilde{L} = [\sqrt{\hat{\lambda}_1} \hat{e}_1, \sqrt{\hat{\lambda}_2} \hat{e}_2, \dots, \sqrt{\hat{\lambda}_m} \hat{e}_m]$, ta có

$$\hat{f}_j = \begin{bmatrix} \frac{1}{\sqrt{\hat{\lambda}_1}} \hat{e}'_1(x_j - \bar{x}) \\ \frac{1}{\sqrt{\hat{\lambda}_2}} \hat{e}'_2(x_j - \bar{x}) \\ \vdots \\ \frac{1}{\sqrt{\hat{\lambda}_m}} \hat{e}'_m(x_j - \bar{x}) \end{bmatrix} \quad (35)$$

Đối với các điểm nhân tố này:

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^n \hat{f}_j &= \mathbf{0} & (\text{trung bình mẫu}) \\ \frac{1}{n-1} \sum_{j=1}^n \hat{f}_j \hat{f}_j' &= \mathbf{I} & (\text{hiệp phương sai mẫu}) \end{aligned}$$

5.2 Phương pháp hồi quy

Bắt đầu với mô hình ban đầu $\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon}$. Khi các nhân tố chung \mathbf{F} và các nhân tố riêng (hoặc sai số) $\boldsymbol{\varepsilon}$ được cùng phân phối chuẩn với trung bình và hiệp phương sai, tổ hợp tuyến tính $\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon}$ có phân phối $\mathbf{N}_p(\mathbf{0}, \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi})$. Hơn nữa, phân phối chung của $(\mathbf{X} - \boldsymbol{\mu})$ và \mathbf{F} là $\mathbf{N}_{m+p}(0, \Sigma^*)$, trong đó:

$$\Sigma^*_{(m+p) \times (m+p)} = \left[\begin{array}{c|c} \Sigma = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi} & \mathbf{L} \\ \hline \mathbf{L}' & \mathbf{I} \end{array} \right] \begin{matrix} (p \times p) & (p \times m) \\ (m \times p) & (m \times m) \end{matrix}$$

Và $\mathbf{0}$ là vector $(m+p) \times 1$. Ta rằng phân phối có điều kiện của $\mathbf{F}|\mathbf{x}$ là chuẩn đa biến với:

$$\text{Trung bình} = E(\mathbf{F}|\mathbf{x}) = \mathbf{L}'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{L}'(\mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}^{-1})^{-1}(\mathbf{x} - \boldsymbol{\mu}) \quad (36)$$

$$\text{Hiệp phương sai} = \text{Cov}(\mathbf{F}|\mathbf{x}) = \mathbf{I} - \mathbf{L}'\Sigma^{-1}\mathbf{L} = \mathbf{I} - \mathbf{L}'(\mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}^{-1})^{-1}\mathbf{L} \quad (37)$$

Điểm nhân tố thu được bằng hồi quy

$$\begin{aligned} \hat{f}_j &= \hat{L}' S^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}), & j=1,2,\dots,n \\ \text{Hoặc nếu ma trận tương quan được tính theo nhân tố} & \\ \hat{f}_j &= \hat{L}'_z R^{-1} z_j, & j=1,2,\dots,n \\ \text{Trong đó: } \mathbf{Z}_j &= \mathbf{D}^{-1/2} (\mathbf{x}_j - \bar{\mathbf{x}}) & \text{và} & \hat{\mathbf{p}} = \hat{\mathbf{L}}_z \hat{\mathbf{L}}'_z + \hat{\boldsymbol{\psi}}_z \end{aligned} \quad (38)$$

Một lần nữa, nếu tải xoay $\hat{\mathbf{L}}^* = \hat{\mathbf{L}}\mathbf{T}$ được sử dụng thay cho tải ban đầu trong (38), điểm nhân tố tiếp theo \hat{f}^*_j liên quan đến \hat{f}_j bởi:

$$\hat{f}^*_j = \mathbf{T}' \hat{f}_j, \quad j=1,2,\dots,n$$

Một thước đo bằng số về sự thống nhất giữa các điểm nhân tố được tạo ra từ hai phương pháp tính toán khác nhau được cung cấp bởi hệ số tương quan mẫu giữa các điểm trên cùng một hệ số. Nói cách khác kết quả giữa 2 phương pháp là tương đồng.

6. Quan điểm và chiến lược Phân tích dữ kiện (Perspectives and a Strategy for Factor Analysis)

Có rất nhiều quyết định cần phải được đưa ra trong các nghiên cứu về phân tích dữ kiện. Có lẽ rằng quyết định quan trọng nhất là việc lựa chọn số nhân tố chung m . Mặc dù phương pháp kiểm định mẫu lớn cho mô hình nhân tố có thể giúp chúng ta tìm ra được giá trị m , nhưng nó chỉ phù hợp với các dữ liệu xấp xỉ phân phối chuẩn. Hơn nữa, kết quả kiểm định hầu hết là bác bỏ mô hình có m nhỏ nếu số lượng biến và số lượng quan sát lớn. Tuy nhiên, đó là tình huống khi phân tích dữ kiện cho ra một xấp xỉ hữu ích. Thông thường, sự lựa chọn cuối cùng cho m dựa trên sự kết hợp của:

- (1) Tỷ lệ phương sai mẫu được giải thích.
- (2) Kiến thức về chủ đề, lĩnh vực mà dữ liệu biểu diễn (tâm lý, kinh tế, điểm số,...)
- (3) Tính hợp lý của kết quả.

Việc lựa chọn phương pháp giải và phép xoay thì ít quan trọng hơn. Trên thực tế, hầu hết các phân tích dữ kiện thích hợp nhất là những phân tích đã được xoay với nhiều hơn một phương pháp và tất cả các kết quả đều cho ra cùng một cấu trúc nhân tố. Hiện nay, việc phân tích vẫn được xem là “nghệ thuật” chứ không theo một chiến lược cứng nhắc duy nhất. Chúng ta sẽ cùng đi qua một số chiến lược có thể được lựa chọn:

1. Thực hiện phân tích dữ kiện theo phương pháp thành phần chính. Phương pháp này đặc biệt thích hợp cho lần đầu tiên duyệt qua dữ liệu.

(a) Tìm kiếm các quan sát bất thường bằng cách vẽ các điểm nhân tố. Đồng thời tính điểm chuẩn cho mỗi quan sát và bình phương khoảng cách.

(b) Thử phép xoay varimax.

2. Thực hiện phân tích dữ kiện theo phương pháp triển vọng cực đại và xoay varimax.

3. So sánh kết quả ở hai phương pháp trên.

(a) Các nhóm hệ số tải có tương tự nhau không?

(b) Vẽ các điểm nhân tố ở hai phương pháp trên.

4. Lặp lại 3 bước trên với một giá trị m khác. Xét xem các nhân tố vừa mở rộng thêm có đóng góp vào việc thể hiện dữ liệu hay không?

5. Đối với tập dữ liệu lớn thì chia tập dữ liệu làm hai và thực hiện phân tích trên mỗi tập con. So sánh hai kết quả này với nhau và với kết quả có được từ bộ dữ liệu hoàn chỉnh để kiểm tra tính ổn định của giải pháp.

7. Kết luận (Conclusion)

Phân tích dữ kiện có sức hấp dẫn to lớn đối với lĩnh vực hành vi và xã hội học. Trong những lĩnh vực này, người ta nhận thấy rằng những biểu hiện tự nhiên bên ngoài quan sát được của động vật hay con người là kết quả của các nhân tố ẩn không quan sát được. Phân tích dữ kiện cung cấp một giải pháp để giúp chúng ta giải thích các quan sát bề nổi dựa theo các nhân tố ẩn này.

Tuy nhiên, việc phân tích dữ kiện vẫn còn mang tính chủ quan của người thực hiện. Trong thực tế, phần lớn các kết quả phân tích không được “đẹp” như các ví dụ của chúng ta ở trên. Kết quả phân tích tốt hay không phụ thuộc vào dữ liệu và mức độ hiểu dữ liệu cũng như hiểu về các nhân tố đã tìm ra của người phân tích.

8. Cài đặt

Phần này trình bày cách cài đặt Phân tích dữ kiện bằng python áp dụng cho bộ dữ liệu thực tế về sự hài lòng của hành khách đi máy bay trên Kaggle.

8.1. Chuẩn bị

➤ Các gói thư viện cần thiết:

Tên thư viện	Chức năng chính	Lệnh cài đặt
--------------	-----------------	--------------

factor_analyzer	Đây là thư viện dùng để phân tích dữ kiện mà chúng ta sẽ dùng.	<code>!pip install factor_analyzer</code>
seaborn	Để vẽ heatmap. Giúp chúng ta dễ dàng đưa ra các đánh giá cần thiết để thực hiện phân tích dữ kiện chính xác hơn	<code>!pip install seaborn</code>
matplotlib	Chúng ta sẽ dùng thư viện này để vẽ đồ thị cho các trị riêng (eigen values)	<code>!pip install matplotlib</code>
numpy	Thao tác với dữ liệu mảng ndarray	<code>!pip install numpy</code>
pandas	Để đọc và giúp ta thao tác với dữ liệu ở dạng dataframe	<code>!pip install pandas</code>
pingouin	Để tính hệ số Cronbach's Alpha đo lường độ tin cậy của các nhân tố ta tìm được.	<code>!pip install pingouin</code>

➤ Chuẩn bị dữ liệu:



Chúng ta sẽ thực hiện phân tích dữ kiện trên bộ dữ liệu về Sự hài lòng của hành khách máy bay (Airline Passenger Satisfaction).

Link dữ liệu:

<https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction>

Đây là bộ dữ liệu khảo sát về sự hài lòng của các hành khách đi máy bay, bao gồm 103 904 quan sát. Bộ dữ liệu có 25 cột, trong đó có 14 cột là kết quả trả lời của hành khách, trên thang đánh giá từ 1 đến 5, đánh giá về các khía cạnh khác nhau của các chuyến bay như là: dịch vụ wifi, đồ ăn và thức uống, dịch vụ check-in,...

Bộ dữ liệu tải về từ Kaggle bao gồm 2 file là train.csv và test.csv. Chúng ta chỉ làm việc với file train.csv.

Name	Date modified	Type	Size
 test.csv	2/20/2020 4:51 PM	Microsoft Excel C...	2,967 KB
 train.csv	2/20/2020 4:51 PM	Microsoft Excel C...	11,908 KB

	id	Gender	Customer Age	Type of Tr/Class	Flight Dist	Inflight wi	Departure Ease	Of Gate loca	Food and	Online bo	Seat comf	Inflight er	On-board	Leg room	Baggage h	Checkin s	Inflight se	Cleanline	Departure	Arrival De	satisfaction		
0	70172	Male	Loyal Cus	13 Personal' Eco Plus	460	3	4	3	1	5	3	5	5	4	3	4	4	5	5	25	18	neutral or dissatisfied	
1	5047	Male	disloyal C	25 Business' Business	235	3	2	3	3	1	3	1	1	1	5	3	1	4	1	1	6	neutral or dissatisfied	
2	110028	Female	Loyal Cus	26 Business' Business	1142	2	2	2	2	5	5	5	5	4	3	4	4	4	5	0	0	satisfied	
3	24026	Female	Loyal Cus	25 Business' Business	562	2	5	5	5	2	2	2	2	2	5	3	1	4	2	11	9	neutral or dissatisfied	
4	119299	Male	Loyal Cus	61 Business' Business	214	3	3	3	3	4	5	5	3	3	4	4	3	3	3	0	0	satisfied	
5	111157	Female	Loyal Cus	26 Personal' Eco	1180	3	4	2	1	1	2	1	1	3	4	4	4	4	1	0	0	neutral or dissatisfied	
6	82113	Male	Loyal Cus	47 Personal' Eco	1276	2	4	2	3	2	2	2	2	3	3	4	3	5	2	9	23	neutral or dissatisfied	
7	96462	Female	Loyal Cus	52 Business' Business	2035	4	3	4	4	5	5	5	5	5	5	5	4	5	4	4	0	satisfied	
8	79485	Female	Loyal Cus	41 Business' Business	853	1	2	2	2	4	3	3	1	1	2	1	4	1	2	0	0	neutral or dissatisfied	
9	65725	Male	disloyal C	20 Business' Eco	1061	3	3	3	4	2	3	3	2	2	3	4	4	3	2	0	0	neutral or dissatisfied	
10	34991	Female	disloyal C	24 Business' Eco	1182	4	5	5	4	2	5	2	2	2	3	3	5	3	5	2	0	0	neutral or dissatisfied
11	51412	Female	Loyal Cus	12 Personal' Eco Plus	308	2	4	2	2	1	2	1	1	1	2	5	5	5	1	0	0	neutral or dissatisfied	
12	98628	Male	Loyal Cus	53 Business' Eco	834	1	4	4	4	1	1	1	1	1	1	3	4	4	1	28	8	neutral or dissatisfied	
13	83502	Male	Loyal Cus	33 Personal' Eco	946	4	2	4	3	4	4	4	4	4	4	5	2	2	2	4	0	0	satisfied
14	95789	Female	Loyal Cus	26 Personal' Eco	453	3	2	3	2	2	3	2	2	2	4	3	2	2	1	2	43	35	neutral or dissatisfied
15	100580	Male	disloyal C	13 Business' Eco	486	2	1	2	3	4	2	1	4	2	1	4	1	3	4	1	0	neutral or dissatisfied	
16	71142	Female	Loyal Cus	26 Business' Business	2123	3	3	3	3	4	4	4	4	5	3	4	5	4	4	49	51	satisfied	
17	127461	Male	Loyal Cus	41 Business' Business	2075	4	4	2	4	4	4	4	4	5	5	5	3	5	5	0	10	satisfied	
18	70354	Female	Loyal Cus	45 Business' Business	2486	4	4	4	4	3	4	4	5	5	5	5	3	5	4	7	5	satisfied	
19	66246	Male	Loyal Cus	38 Personal' Eco	460	2	3	3	2	5	3	5	5	1	2	4	3	2	5	17	18	neutral or dissatisfied	
20	39076	Male	Loyal Cus	9 Business' Eco	1174	2	4	2	4	2	2	1	2	1	5	3	4	3	2	0	4	neutral or dissatisfied	
21	22434	Female	Loyal Cus	17 Personal' Eco	208	3	1	3	3	5	3	5	5	2	5	3	3	4	5	0	0	neutral or dissatisfied	
22	43510	Female	Loyal Cus	43 Personal' Eco	752	3	5	3	5	5	4	5	5	3	3	3	5	3	4	52	29	neutral or dissatisfied	
23	114090	Female	Loyal Cus	58 Personal' Eco	2139	4	5	4	5	4	3	4	4	4	4	4	2	4	2	0	0	neutral or dissatisfied	
24	105420	Female	disloyal C	23 Business' Eco	452	5	0	5	1	1	5	1	1	4	5	5	5	3	5	1	54	44	satisfied
25	102956	Male	Loyal Cus	57 Personal' Eco	719	4	4	4	1	5	4	5	5	3	2	4	4	5	5	27	28	neutral or dissatisfied	
	10510	Female	disloyal C	23 Business' Eco	1564	3	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	satisfied

8.2. Các tác vụ cơ bản để phân tích dữ kiện

- Tác vụ 1: Xử lý dữ liệu để phù hợp cho việc phân tích.
- Tác vụ 2: Tính các kiểm định cần thiết để xem dữ liệu có phù hợp để phân tích dữ kiện hay không.
- Tác vụ 3: Phân tích sơ khởi (xác định số nhân tố bất kỳ) để tính các eigenvalues.
- Tác vụ 4: Dựa vào eigenvalues, chọn số lượng nhân tố thích hợp và tiến hành phân tích lại.
- Tác vụ 5: Dựa vào các hệ số tải có được từ phân tích ở bước 4, xác định các biến được giữ lại cho mỗi nhân tố.
- Tác vụ 6: Dựa vào ý nghĩa của các biến quan sát thuộc mỗi nhân tố để suy ra ý nghĩa của nhân tố.
- Tác vụ 7: Kiểm định độ tin cậy (độ tốt) của các nhân tố vừa tìm được.
- Tác vụ 8: Kết luận.

8.3. Tiến hành cài đặt

Phần này trình bày phương pháp phân tích dữ kiện sử dụng ngôn ngữ lập trình Python, soạn thảo và chạy bằng Jupyter Notebook.

Các bước chi tiết nêu ra dưới đây bao gồm việc import các thư viện cần thiết nêu ra ở mục 2.1, đọc và xử lý dữ liệu, triển khai cụ thể hơn các tác vụ đã nêu ra ở mục 2.2.

➤ Bước 1:

import các thư viện cần thiết đã nêu ở mục 2.1

```
In [1]: import numpy as np
import pandas as pd
from factor_analyzer import FactorAnalyzer
import matplotlib.pyplot as plt
import seaborn as sns
import pingouin as pg
```

➤ Bước 2:

Đọc dữ liệu bằng hàm `read_csv()` của pandas. Tất nhiên, chúng ta cần xem trong dữ liệu chứa nội dung gì để có thể tiếp tục công việc phân tích, nên ta sẽ hiển thị phần đầu của dữ liệu thông qua hàm `head()`.

```
In [2]: df = pd.read_csv('Data/train.csv')
df.head()
```

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	...	Inflight entertainment
0	0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	...	5	...	4
1	1	5047	Male	disloyal Customer	25	Business travel	Business	235	3	2	...	1	...	1
2	2	110028	Female	Loyal Customer	26	Business travel	Business	1142	2	2	...	5	...	4
3	3	24026	Female	Loyal Customer	25	Business travel	Business	562	2	5	...	2	...	2
4	4	119299	Male	Loyal Customer	61	Business travel	Business	214	3	3	...	3	...	3

5 rows × 25 columns

➤ Bước 3:

Nhận thấy hai cột đầu tiên (cột số thứ tự Unnamed và cột id) không có ý nghĩa về vấn đề cần quan tâm nên trước khi vẽ heatmap cho ma trận tương quan ở bước sau, ta sẽ loại bỏ 2 cột này ra khỏi dataframe.

Tham số `axis=1` có nghĩa các trường cần drop là cột (chiều ngang), thiết lập `inplace=True` để dataframe tự thay đổi và trả về chính nó (nếu để `inplace=False` thì nó sẽ trả về một object khác).

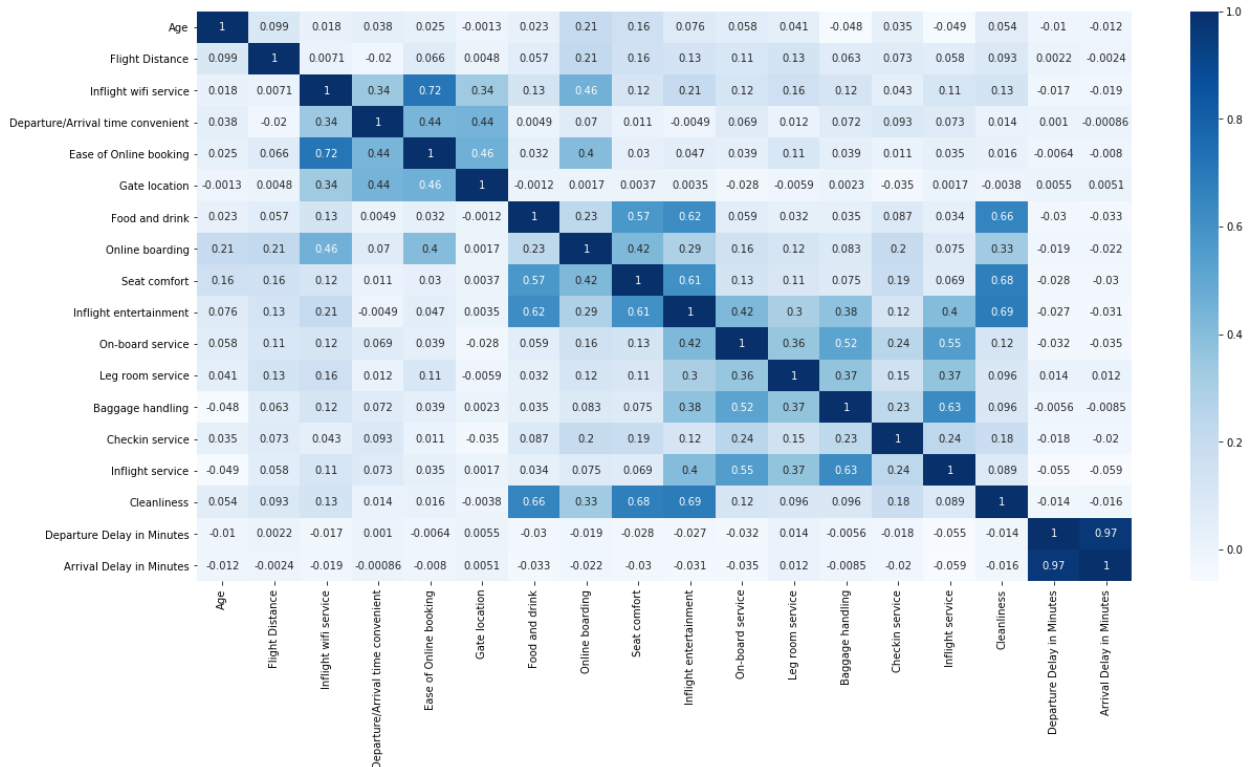
```
In [6]: df.drop(['Unnamed: 0', 'id'], axis=1, inplace=True)
```

➤ Bước 4:

Tính heatmap cho ma trận tương quan (correlation matrix)

```
In [8]: plt.figure(figsize=(20,10))
correlation_matrix = df.corr()
sns.heatmap(correlation_matrix, cmap='Blues')
```

Ta được kết quả:



➤ Bước 5:

Quan sát heatmap trên, ta nhận thấy hai biến “Departure Delay in Minutes” và “Arrival Delay in Minutes” là hai biến có mối tương quan lớn. Trên thực tế, nó có mối quan hệ kéo theo: Khởi hành chậm trễ => Đến nơi chậm trễ. Do đó, ta sẽ loại bỏ biến Đến nơi chậm trễ (Arrival Delay in Minutes) vì nó hơi “dư thừa”. Việc loại bỏ này giúp cho chương trình tính toán nhanh hơn một chút mà vẫn không làm mất mát thông tin.

```
In [5]: df.drop(['Arrival Delay in Minutes'], axis=1, inplace=True)
```

➤ Bước 6:

Bây giờ, chúng ta sẽ quan tâm 14 biến là 14 tiêu chí khảo sát sự hài lòng của hành khách. Chúng ta sẽ thực hiện phân tích dữ kiện dựa trên 14 biến này.

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

```
In [6]: #Subset of the data
        X = df[df.columns[6:20]]
```

➤ Bước 7:

Tiêu chuẩn để bộ dữ liệu phù hợp thực hiện phân tích dữ kiện là chỉ số KMO (Kaiser-Meyer-Olkin Measure of Sampling Adequacy) phải lớn hơn 0.5 (Garson, 2003). Thực hiện tính chỉ số KMO bằng hàm `calculate_kmo()`

```
In [7]: # Kaiser-Meyer-Olkin (KMO) Test
        # Dùng để xem xét sự thích hợp của phân tích nhân tố (0.5 <= kmo <=1)
        # KMO phải đạt 0.5 trở lên thì mới đủ điều kiện để thực hiện phân tích nhân tố
        from factor_analyzer.factor_analyzer import calculate_kmo
        kmo_all,kmo_model=calculate_kmo(X)
        kmo_model
```

0.781229425716438

Kết quả chỉ số KMO là xấp xỉ $0.781 > 0.5$. Do đó, chúng ta có thể thực hiện phân tích dữ kiện trên bộ dữ liệu này.

➤ Bước 8:

Một tiêu chuẩn nữa của việc phân tích dữ kiện là kiểm định Bartlett phải có mức ý nghĩa $\text{sig} < 0.05$. Vì khi đó ta sẽ bác bỏ giả thuyết H_0 : “Các biến quan sát không có tương quan với nhau trong tổng thể.” Mà điều cần có để thực hiện được phân tích dữ kiện là các biến phải có mối tương quan.

Tính kiểm định Barlett bằng hàm `calculate_barlett_sphericity()`

```
In [8]: from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value, sig=calculate_bartlett_sphericity(X)
print('Chi square value:', chi_square_value)
print('Sig. ', sig)
```

```
Chi square value: 601690.8930479608
Sig. 0.0
```

Kết quả cho thấy mức ý nghĩa $\text{sig} = 0.0 < 0.05 \Rightarrow$ Bác bỏ H_0 , có thể phân tích dữ kiện.

➤ Bước 9:

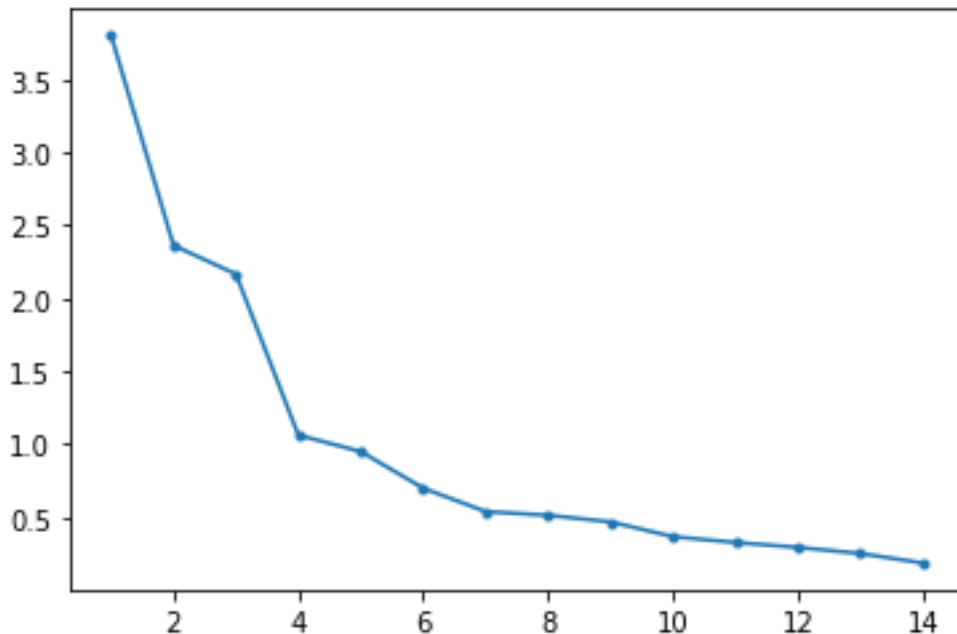
Thực hiện phân tích dữ kiện bằng hàm `fit()` trong class `FactorAnalyzer` với các tham số mặc định. Mục đích của bước này là để lưu kết quả có được từ hàm `fit()` vào đối tượng `fa` phục vụ cho bước sau.

```
In [9]: fa = FactorAnalyzer()
fa.fit(X)
```

➤ Bước 10: Ở bước 9 ta chỉ thực hiện tác vụ phân tích ban đầu (chưa xác định số nhân tố). Ở bước này, chúng ta sẽ gọi hàm `get_eigenvalues()` và vẽ đồ thị để từ đó lựa chọn số nhân tố cho phù hợp.

```
In [10]: #Get Eigen values and plot
ev, v = fa.get_eigenvalues()
print(ev)
plt.plot(range(1,X.shape[1]+1),ev, marker='.'))
```

```
[3.80011677 2.36198598 2.16589224 1.06327401 0.95093123 0.7003355
0.53995637 0.51465504 0.46947475 0.36866001 0.32840792 0.29509562
0.25317089 0.18804368]
```



➤ Bước 11:

Giá trị eigenvalue là thước đo mức độ phương sai của các biến mà một nhân tố giải thích. Thông thường, những nhân tố có *eigenvalue* ≥ 1 sẽ được giữ lại trong mô hình phân tích. Nếu dựa theo lý thuyết này thì ta sẽ chọn được 4 nhân tố và hệ số tải có được như bên dưới

```
[[ 9.26063710e-02  1.32302910e-01  6.05629505e-01  4.78034712e-01]
 [-6.28689994e-03  5.71663401e-02  5.89642950e-01  2.83064933e-04]
 [-3.61238330e-02  2.75031290e-02  7.66508969e-01  4.63395908e-01]
 [ 1.30970957e-02 -4.51418860e-02  6.80812592e-01 -1.00043760e-01]
 [ 7.70129743e-01  2.84511493e-03  3.29873832e-02  4.00335268e-02]
 [ 2.86895167e-01  1.18512193e-01  9.43010118e-02  7.56382275e-01]
 [ 7.54094309e-01  7.86463070e-02 -2.80995391e-02  2.13668389e-01]
 [ 7.66236835e-01  4.64665983e-01  4.12224708e-02  3.27084611e-02]
 [ 8.79310753e-02  7.00438053e-01  1.03848652e-02  5.18498254e-02]
 [ 5.75409430e-02  4.83207685e-01  4.05585608e-02  9.74845714e-02]
 [ 3.67376530e-02  7.63384135e-01  4.76166124e-02 -3.05831410e-02]
 [ 1.16816116e-01  2.85782472e-01 -2.55074815e-02  1.31914365e-01]
 [ 3.59624126e-02  7.99640607e-01  4.75951818e-02 -5.15726378e-02]
 [ 8.54325679e-01  8.24496911e-02 -4.58580297e-05  1.03142774e-01]]
```

Có gì đó không ổn khi mà các hệ số tải có giá trị quá nhỏ!

Để ý một chút, ta sẽ thấy đồ thị eigenvalues ở trên “lao dốc” khá mạnh sau nhân tố thứ 3. Vì vậy, số nhân tố ta chọn chính xác sẽ là 3 nhân tố.

Phân tích lại với số nhân tố là 3 và sử dụng phương pháp xoay varimax.

```
In [11]: fa = FactorAnalyzer(3, rotation='varimax')
          fa.fit(X)
          loads = fa.loadings_
          print(loadings)
```

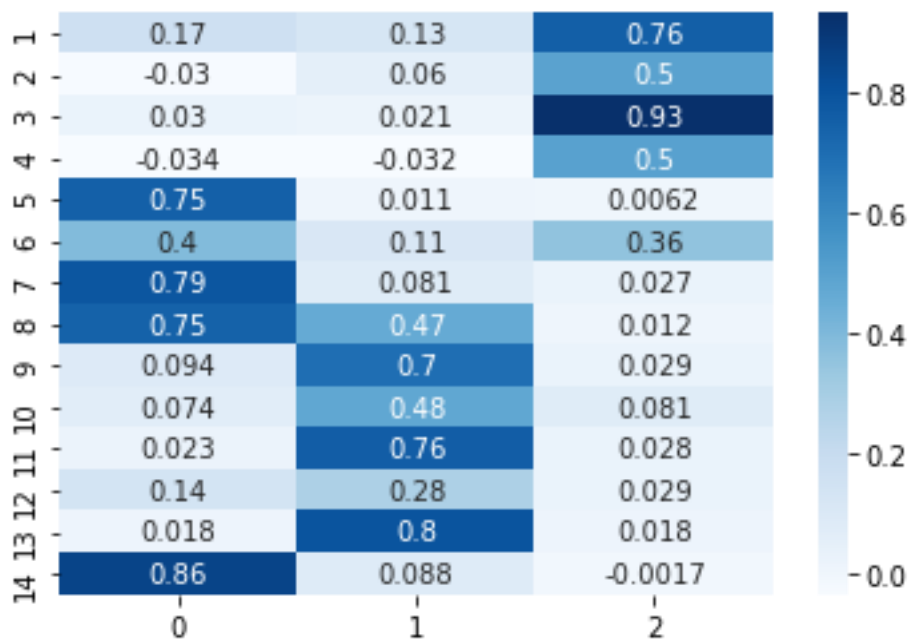
Ta được các hệ số tải:

```
[[ 0.16826952  0.12827119  0.75809134]
 [-0.02950837  0.05968117  0.50138365]
 [ 0.03023106  0.02091436  0.93277526]
 [-0.0338282  -0.03231121  0.50404385]
 [ 0.75263893  0.01094635  0.00616734]
 [ 0.39545345  0.1138114  0.35906543]
 [ 0.78999048  0.08146326  0.02725824]
 [ 0.7456934  0.46674984  0.01203424]
 [ 0.09388069  0.70115382  0.02900913]
 [ 0.07445487  0.48144209  0.08065029]
 [ 0.02346305  0.76474833  0.02769279]
 [ 0.14351222  0.28418169  0.02888186]
 [ 0.01813146  0.79977083  0.01825226]
 [ 0.85842046  0.08814824 -0.00170807]]
```

➤ Bước 12:

Trực quan các con số trên bằng heatmap sẽ giúp chúng ta dễ dàng quan sát hơn.

```
In [12]: ylabels=[]
          for i in range(1,15):
              ylabels.append(i)
          sns.heatmap(loadings, annot=True, cmap='Blues', yticklabels=ylabels)
```

➤ Bước 13:

- Theo [3], thì:

- + Hệ số tải ở mức ± 0.3 : Điều kiện tối thiểu để biến quan sát được giữ lại.
- + Hệ số tải ở mức ± 0.5 : Biến quan sát có ý nghĩa thống kê tốt.
- + Hệ số tải ở mức ± 0.7 : Biến quan sát có ý nghĩa thống kê rất tốt.

- Chúng ta sẽ giữ lại các biến có hệ số tải ≥ 0.5

+ Nhân tố 1: Giữ lại các biến 5, 7, 8, 14 có các hệ số tải lần lượt là 0.75, 0.79, 0.75, 0.86.

+ Nhân tố 2: Giữ lại các biến 9, 11, 13 có các hệ số tải lần lượt là 0.7, 0.76, 0.8.

+ Nhân tố 3: Giữ lại các biến 1, 2, 3, 4 có các hệ số tải lần lượt là 0.76, 0.5, 0.93, 0.5.

- Ta sẽ tạo 3 factors này.

+ Lấy tên các cột (các biến quan sát) từ dữ liệu X ta đã tạo.

```
In [13]: columns = []
for col in X.columns:
    columns.append(col)
print(columns)
```

```
['Inflight wifi service', 'Departure/Arrival time convenient', 'Ease of Online booking', 'Gate location', 'Food and drink', 'Online boardi
ng', 'Seat comfort', 'Inflight entertainment', 'On-board service', 'Leg room service', 'Baggage handling', 'Checkin service', 'Inflight se
rvice', 'Cleanliness']
```

+ Giữ lại các biến có hệ số tải từ 0.5 trở lên cho mỗi factors. Sử dụng hàm `np.where()` để lấy vị trí cần tìm. Lưu tên các biến quan sát dưới dạng

string.

```
In [14]: vars_of_factor1 = []
vars_of_factor2 = []
vars_of_factor3 = []

index1 = np.where(loads[:,0]>=0.5)[0]
for i in index1:
    vars_of_factor1.append(columns[i])
print('Variables of factor 1:', vars_of_factor1)

index2 = np.where(loads[:,1]>=0.5)[0]
for i in index2:
    vars_of_factor2.append(columns[i])
print('Variables of factor 2:', vars_of_factor2)

index3 = np.where(loads[:,2]>=0.5)[0]
for i in index3:
    vars_of_factor3.append(columns[i])
print('Variables of factor 3:', vars_of_factor3)
```

Kết quả:

```
Variables of factor 1: ['Food and drink', 'Seat comfort', 'Inflight entertainment', 'Cleanliness']
Variables of factor 2: ['On-board service', 'Baggage handling', 'Inflight service']
Variables of factor 3: ['Inflight wifi service', 'Departure/Arrival time convenient', 'Ease of Online booking', 'Gate location']
```

Dựa vào các biến thuộc mỗi nhân tố, ta suy ra ý nghĩa của nhân tố đó. Tức là nội dung đại diện của nhân tố thay cho các biến này.

Ta có bảng sau:

Nhân tố	Các biến quan sát	Ý nghĩa và gọi tên nhân tố
Nhân tố 1	Food and drink, Seat comfort, Inflight entertainment, Cleanliness.	- Các biến này có điểm chung là thể hiện sự thoải mái của chuyến bay. - Ta có thể gọi nó là nhân tố Comfort .
Nhân tố 2	On-board service, Baggage handling, Inflight service.	- Các biến này đều thể hiện về dịch vụ hàng không. - Ta đặt tên cho nhân tố này là: Service .
Nhân tố 3	Inflight wifi service, Departure/Arrival time convenient, Ease of Online booking, Gate location.	- Các biến này thể hiện sự thuận tiện. - Ta đặt tên cho nhân tố này là: Convenience .

➤ Bước 14:

Để đánh giá lại độ tin cậy của các nhân tố mà chúng ta vừa có được, tức là xem chúng có thực sự thể hiện tốt dữ liệu hay không, ta sẽ đi tính hệ số Cronbach's Alpha của 3 nhân tố này.

Mức giá trị hệ số Cronbach's Alpha [3]:

- Từ 0.8 đến gần bằng 1: thang đo lường rất tốt.
- Từ 0.7 đến gần bằng 0.8: thang đo lường tốt.
- Từ 0.6 trở lên: thang đo lường đủ điều kiện.

Để thực hiện bằng python, ta tiếp tục bằng cách tạo dataframe cho 3 nhân tố hiện có. Sau đó dùng hàm `cronbach_alpha()` trong thư viện `pingouin` để tính hệ số Cronbach's Alpha.

```
In [18]: factor1 = X[vars_of_factor1]
factor2 = X[vars_of_factor2]
factor3 = X[vars_of_factor3]
#Get cronbach alpha
factor1_alpha = pg.cronbach_alpha(factor1)
factor2_alpha = pg.cronbach_alpha(factor2)
factor3_alpha = pg.cronbach_alpha(factor3)

print('Factor 1 alpha:', factor1_alpha[0])
print('Factor 2 alpha:', factor2_alpha[0])
print('Factor 3 alpha:', factor3_alpha[0])
```

Kết quả:

```
Factor 1 alpha: 0.87628779166241
Factor 2 alpha: 0.7942916933090223
Factor 3 alpha: 0.7679754211110685
```

Ta thấy các hệ số Cronbach's Alpha của các nhân tố đều lớn hơn 0.7. Điều này có nghĩa là chúng thể hiện dữ liệu khá tốt và việc phân tích của chúng ta đã thành công.

8.4. Tổng kết

Từ 14 biến quan sát, chúng ta đã trải qua các bước xử lý, phân tích, ra quyết định, để rút lại còn 3 nhân tố chính có ảnh hưởng đến sự hài lòng của hành khách khi đi máy bay, đó là: **Comfort** (Sự thoải mái), **Service** (Dịch vụ), **Convenience** (Sự thuận tiện). Để số lượng khách hàng tăng lên, số vé bán ra nhiều hơn và tất nhiên là doanh thu cao hơn, nhà kinh doanh dịch vụ bay cần tập trung làm tăng chất lượng các mảng liên quan đến ba nhân tố trên. Ví dụ, để gia tăng sự thoải mái, người kinh doanh có thể tập trung đầu tư vào yếu tố “Nhân viên phục vụ vui vẻ, thân thiện” để gia tăng sự hài lòng của hành khách.

Qua việc phân tích trên, chúng ta thấy rằng phân tích dữ kiện có ý nghĩa rất lớn trong nhiều lĩnh vực như kinh doanh, tâm lý,... Nhà phân tích, một khi đã tìm ra các nhân tố ảnh hưởng đến vấn đề cần quan tâm thì cũng giống như nắm được “quy luật bí mật”. Điều này sẽ giúp chúng ta ra quyết định chính xác và đột phá hơn, đem lại giá trị cực kỳ cao.

9. Tài liệu tham khảo (Reference)

- [1] Richard Johnson, Dean Wichern, *Applied Multivariate Statistical Analysis*, Pearson.
- [2] Wolfgang Karl Hardle Léopold Simar, *Applied Multivariate Statistical Analysis*, Springer.
- [3] Edouard Duchesnay, Tommy Löfstedt, *Statistics and Machine Learning in Python – release 1.0 (2017)*
- [4] GS.Nguyễn Tiến Dũng, GS. Đỗ Đức Thái, *Nhập môn Xác suất & Thống kê hiện đại*, Sputnik.

-Hết-