Họ và Tên: PHẠM THỊ HOÀ
MSHV: 23C23007
Ngành: Lý Thuyết Xác Suất và Thống Kê Toán Học
Môn Học: Thống Kê Nhiều Chiều
GVHD: TS. Nguyễn Thị Mộng Ngọc

I. Phần lý thuyết và ứng dụng (5 điểm)

1. Lý thuyết chủ đề phân tích nhân tố

Nội dung chính:

1. Mô hình: Mô hình nhân tố trực giao
2. Phương pháp ước lượng (Methods of Estimation)
   - Phương pháp thành phần chính
   - Phương pháp ước lượng hợp lý cực đại
3. Xoay nhân tố (Factor Rotation)
4. Điểm của nhân tố (Factor Scores)
   - Phương pháp bình phương tối thiểu có trọng số
   - Phương pháp hồi quy
5. Kiểm định mô hình nhân tố trên mẫu lớn: Kiểm định tính hợp lý của mô hình.

1.1. Mô hình

Mục đích chính của phân tích nhân tố là giải thích mối quan hệ của các biến thông qua ma trận hiệp phương sai của chúng, dựa trên một vài đại lượng ngẫu nhiên cơ bản nhưng không quan sát được, được gọi là các nhân tố.

Mô hình nhân tố được phát triển bởi lập luận sau:

- Giả sử có thể phân chia các biến khảo sát thành các nhóm.

- Tất cả các biến trong một nhóm cụ thể đều có mối tương quan cao với nhau, nhưng lại có mối tương quan tương đối thấp với các biến trong một nhóm khác.

Khi đó, có thể tưởng tượng rằng mỗi nhóm các biến đại diện cho một cấu trúc cơ bản duy nhất, hoặc một nhân tố, chịu trách nhiệm cho các mối tương quan được quan sát.

Ví dụ, các mối tương quan từ nhóm điểm kiểm tra trong các môn cổ điển, tiếng Pháp, tiếng Anh, toán học và âm nhạc do Spearman thu thập đã gợi ý một nhân tố cơ bản là "trí thông minh". Một nhóm thứ hai của các biến, đại diện cho điểm số thể lực, nếu có, có thể tương ứng với một nhân tố khác. Đây là loại cấu trúc mà phân tích nhân tố tìm cách xác nhận.

Phân tích nhân tố có thể được coi là sự mở rộng của phân tích thành phần chính. Cả hai có thể được xem như là những nỗ lực để ước lượng ma trận hiệp phương sai. Tuy nhiên, sự ước lượng dựa trên mô hình phân tích nhân tố là phức tạp hơn.

Mô hình nhân tố trực giao (The Orthogonal Factor Model)

Vector ngẫu nhiên quan sát được X, với:

- Số thành phần: p ,
- Trung bình: μ,
- Ma trận hiệp phương sai: Σ.

Mô hình nhân tố giả định rằng: X phụ thuộc tuyến tính vào một số lượng ít các nhân tố không quan sát được:

- $F_1, F_2, \ldots, F_m$, được gọi là nhân tố chung (common factors),
- p sai số (errors) (hoặc còn gọi là hoặc là nhân tố cụ thể (specific factors)) $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_p$.

Mô hình phân tích nhân tố được mô tả:

$$
\begin{aligned}
X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \ldots + \ell_{1m}F_m + \boldsymbol{\epsilon}_1 \\
X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \ldots + \ell_{2m}F_m + \boldsymbol{\epsilon}_2 \\
&\vdots \\
X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \ldots + \ell_{pm}F_m + \boldsymbol{\epsilon}_p
\end{aligned}
$$

hoặc, dưới dạng ký hiệu ma trận,

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon}$$

(công thức 9-2)

Trong đó: ℓ_{ij} được gọi là hệ số tải (loading) của biến thứ i trên nhân tố thứ j, do đó ma trận L là ma trận của các hệ số tải nhân tố (matrix of factor loadings).

Lưu ý:

- Sai số thứ i $\boldsymbol{\epsilon}_i$ chỉ liên quan đến phản hồi X_i.

- p độ lệch : $X_1 - \mu_1, X_2 - \mu_2, \ldots, X_p - \mu_p$ được biểu diễn bằng $p + m$ theo các biến ngẫu nhiên $F_1, F_2, \ldots, F_m, \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_p$, với F_i và $\boldsymbol{\epsilon}_i$ không thể quan sát được.

Với rất nhiều đại lượng không quan sát được, việc xác minh trực tiếp mô hình nhân tố từ các quan sát trên $X_1, X_2, \ldots, X_p$ là không thể. Tuy nhiên, với một số giả định bổ sung về các vector ngẫu nhiên F và ε, mô hình nhân tố trực giao đưa ra một số mối quan hệ hiệp phương sai, mà có thể được kiểm tra.

Chúng ta giả định rằng:

$$E(\mathbf{F}) = \mathbf{0}_{m\times 1}, \quad \text{Cov}(\mathbf{F}) = E[\mathbf{F}\mathbf{F}'] = \mathbf{I}_{m\times m},$$

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}_{p\times 1}, \quad \text{Cov}(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}_{p\times p},$$

(công thức 9-3)

trong đó $\mathbf{0}_{m\times 1}$ và $\mathbf{0}_{p\times 1}$ là các vector không, $\mathbf{I}_{m\times m}$ là ma trận đơn vị kích thước $m \times m$, và $\boldsymbol{\Psi}$ là ma trận đường chéo với các phần tử $\psi_1, \psi_2, \ldots, \psi_p$ trên đường chéo chính.

Do F và $\boldsymbol{\epsilon}$ là độc lập, vì vậy

$$\text{Cov}(\boldsymbol{\epsilon}, \mathbf{F}) = E(\boldsymbol{\epsilon}\mathbf{F}') = \mathbf{0}_{p\times m},$$

Những giả định này và mối quan hệ trong (công thức 9-2) tạo thành mô hình nhân tố trực giao (orthogonal factor model).

Mô Hình Nhân Tố Trực Giao với m Nhân Tố Chung

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{L}\mathbf{F} + \boldsymbol{\epsilon}$$

(công thức 9-4)

trong đó:

- μ_i là trung bình của biến thứ i,
- ϵ_i là nhân tố cụ thể thứ i,
- F_j là nhân tố chung thứ j,

- ℓ_{ij} là hệ số tải của biến thứ i trên nhân tố thứ j.

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

$$F \text{ và } \epsilon \text{ là độc lập}$$

$$E(\mathbf{F}) = \mathbf{0}, \quad \text{Cov}(\mathbf{F}) = \mathbf{I}$$

$$E(\boldsymbol{\epsilon}) = \mathbf{0}, \quad \text{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi}, \text{ trong đó } \boldsymbol{\Psi} \text{ là một ma trận đường chéo}$$

Mô hình nhân tố trực giao hàm ý một cấu trúc hiệp phương sai cho X. Từ mô hình trong (công thức 9-4):

$$(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' = (\mathbf{LF} + \boldsymbol{\epsilon})((\mathbf{LF} + \boldsymbol{\epsilon})') = \mathbf{LF}(\mathbf{LF})' + \boldsymbol{\epsilon}(\mathbf{LF})' + \mathbf{LF}\boldsymbol{\epsilon}' + \boldsymbol{\epsilon}\boldsymbol{\epsilon}'$$

vì thế

$$\Sigma = \text{Cov}(\mathbf{X}) = \text{E}[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'] = \text{E}[\mathbf{LF}(\mathbf{LF})']+\text{E}[\boldsymbol{\epsilon}(\mathbf{LF})']+\text{E}[\mathbf{LF}\boldsymbol{\epsilon}']+\text{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'] = \mathbf{LL}'+\boldsymbol{\Psi}$$

theo (công thức 9-3).

Do F và $\boldsymbol{\epsilon}$ độc lập: $\text{Cov}(\boldsymbol{\epsilon}, \mathbf{F}) = \text{E}[\boldsymbol{\epsilon}, \mathbf{F}'] = 0$

Theo (công thức 9-4),

$(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}' = (\mathbf{LF} + \boldsymbol{\epsilon})\mathbf{F}' = \mathbf{LFF}' + \boldsymbol{\epsilon}\mathbf{F}'$,

vì vậy $\text{Cov}(\mathbf{X}, \mathbf{F}) = \text{E}[(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}'] = \text{E}[\mathbf{LF}(\mathbf{F}') + \text{E}[\boldsymbol{\epsilon}\mathbf{F}'] = \mathbf{L}$.

Cấu Trúc Hiệp Phương Sai cho Mô Hình Nhân Tố Trực Giao

1. $\text{Cov}(\mathbf{X}) = \mathbf{LL}' + \boldsymbol{\Psi}$

hoặc

$$\text{Var}(X_i) = \ell_{i1}^2 + \ldots + \ell_{im}^2 + \psi_i$$

$$\text{Cov}(X_i, X_k) = \ell_{i1}\ell_{k1} + \ldots + \ell_{im}\ell_{km}$$

(công thức 9-5)

2. $\text{Cov}(\mathbf{X}, \mathbf{F}) = \mathbf{L}$

hoặc

$$\text{Cov}(X_i, F_j) = \ell_{ij}$$

Mô hình $X - \mu = LF + \epsilon$ là tuyến tính theo các nhân tố chung. Nếu p các X thực sự liên quan đến các nhân tố con (underlying factors), nhưng mối quan hệ là phi tuyến, ví dụ như $X_1 - \mu_1 = \ell_{11}F_1^3 + \epsilon_1$, $X_2 - \mu_2 = \ell_{21}F_2^3 + \epsilon_2$, thì cấu trúc hiệp phương sai $LL' + \Psi$ được đưa ra bởi (công thức 9-5) có thể sẽ không phù hợp. Giả định rất quan trọng về tính tuyến tính vốn có trong việc xây dựng mô hình nhân tố truyền thống.

Phần của phương sai của biến thứ i được góp bởi m nhân tố chung được gọi là tính chung (communality) của biến thứ i. Phần của $\text{Var}(X_i) = \sigma_{ii}$ do nhân tố cụ thể thường được gọi là độc nhất (uniqueness). Ký hiệu phần chung của biến thứ i bởi h_i^2, chúng ta thấy từ (9-5) rằng

$$\sigma_{ii} = \ell_{i1}^2 + \ell_{i2}^2 + \ldots + \ell_{im}^2 + \psi_i$$

$$\text{Var}(X_i) = \text{communality} + \text{specific variance}$$

hoặc

$$h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \ldots + \ell_{im}^2$$

và

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

Tính chung (communality) thứ i là tổng các bình phương của các hệ số tải của biến thứ i trên m nhân tố chung.

1.2. Phương pháp

$x_1, x_2, \ldots, x_n$ trên p biến số tương quan chung.

Vấn đề đặt ra: mô hình nhân tố của (công thức 9-4), với một số lượng nhỏ các nhân tố, có đại diện chính xác cho dữ liệu không?

=> Chúng ta giải quyết vấn đề mô hình thống kê này bằng cách xác thực mối quan hệ hiệp phương sai trong (công thức 9-5).

Ma trận hiệp phương sai mẫu S là ước lượng của ma trận Σ.

Nếu các phần tử nằm ngoài đường chéo của S nhỏ hoặc các phần tử của ma trận tương quan R cơ bản bằng không, các biến số không liên quan và phân tích nhân tố sẽ không hữu ích. Trong những trường hợp này, các nhân tố cụ thể đóng vai trò chủ đạo, trong khi mục tiêu chính của phân tích nhân tố là xác định một vài nhân tố chung quan trọng.

Nếu Σ dường như lệch đáng kể so với một ma trận đường chéo, thì một mô hình nhân tố có thể được xem xét, và vấn đề ban đầu là một trong những ước lượng các tải nhân tố ℓ_{ij} và các phương sai cụ thể ψ_i.

Hai trong số các phương pháp ước lượng tham số phổ biến nhất là:

- Phân tích thành phần chính (và phương pháp nhân tố chính liên quan)

- Phương pháp ước lượng hợp lý cực đại.

Giải pháp từ cả hai phương pháp này có thể được xoay vị trí để đơn giản hóa việc giải thích các nhân tố, như đã mô tả trong Mục 9.4.

1.2.1. Phương pháp Thành phần Chính (và Nhân tố Chính)

Giả sử Σ có các cặp trị riêng - vectơ riêng (λ_i, e_i) với $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Khi đó

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

$$= \left[\sqrt{\lambda_1} e_1 \mid \sqrt{\lambda_2} e_2 \mid \ldots \mid \sqrt{\lambda_p} e_p\right]$$

$$= \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \sqrt{\lambda_2} e_2' \\ \vdots \\ \sqrt{\lambda_p} e_p' \end{bmatrix}$$

Cấu trúc này phù hợp với cấu trúc hiệp phương sai đã được định trước cho mô hình phân tích nhân tố có nhiều nhân tố bằng số biến số ($m = p$) và các phương sai cụ thể $\psi_i = 0$ cho mọi i. Ma trận tải nhân tố có cột thứ j được cho bởi $\sqrt{\lambda_j} e_j$. Cụ thể, chúng ta có thể viết

$$\Sigma_{(p\times p)} = L_{(p\times p)} L'_{(p\times p)} + \theta_{(p\times p)} = LL'$$

(công thức 9-11)

Ngoại trừ yếu tố tỉ lệ $\sqrt{\lambda_j}$, nhân tố loadings cho thành phần thứ j trong phân tích thành phần chính.

Mặc dù biểu diễn phân tích nhân tố của Σ trong (công thức 9-11) là chính xác, nó không đặc biệt hữu ích:

- Nó sử dụng nhiều nhân tố chung bằng số biến số và không cho phép bất kỳ sự biến thiên nào trong các nhân tố cụ thể ε trong (công thức 9-4).

- Khi các trị riêng cuối cùng từ $p - m$ là nhỏ, là bỏ qua sự đóng góp của $\lambda_{m+1} e_{m+1} e'_{m+1} + \ldots + \lambda_p e_p e'_p$ vào Σ. Bỏ qua sự đóng góp này, chúng ta thu được phép xấp xỉ

$$\Sigma = \left[\sqrt{\lambda_1} e_1 \mid \sqrt{\lambda_2} e_2 \mid \ldots \mid \sqrt{\lambda_m} e_m\right] = \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \sqrt{\lambda_2} e_2' \\ \vdots \\ \sqrt{\lambda_m} e_m' \end{bmatrix} = LL'$$

(công thức 9-12)

Biểu diễn gần đúng trong (công thức 9-12) giả sử rằng các nhân tố cụ thể ε trong (công thức 9-4) không quan trọng và cũng có thể được bỏ qua trong việc phân tách của Σ. Nếu các nhân tố cụ thể được bao gồm trong mô hình, các phương sai của chúng có thể được xem là các phần tử đường chéo của $\Sigma - LL'$, nơi LL' được định nghĩa như trong (công thức 9-12).

Khi cho phép các nhân tố cụ thể, chúng ta thấy rằng phép xấp xỉ trở thành

$$\Sigma \approx LL' + \Psi$$

$$= \left[\sqrt{\lambda_1}e_1 \mid \sqrt{\lambda_2}e_2 \mid \ldots \mid \sqrt{\lambda_m}e_m\right] + \begin{bmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \psi_p \end{bmatrix}$$

(công thức 9-13)

trong đó $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} \ell_{ij}^2$ cho $i = 1, 2, \ldots, p$.

Để áp dụng cách tiếp cận này vào một bộ dữ liệu $x_1, x_2, \ldots, x_n$, thông thường người ta trước tiên sẽ dùng định lý giới hạn trung tâm các quan sát bằng cách trừ đi trung bình mẫu $\bar{x}$. Các quan sát đã trung tâm có dạng

$$x_j - \bar{x} = \begin{bmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jp} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} x_{j1} - \bar{x}_1 \\ x_{j2} - \bar{x}_2 \\ \vdots \\ x_{jp} - \bar{x}_p \end{bmatrix}$$

với $j = 1, 2, \ldots, n$

Các quan sát này có ma trận hiệp phương sai mẫu S giống như các quan sát gốc.

Trong trường hợp các đơn vị của các biến số không đồng nhất, thường là mong muốn làm việc với các biến số đã chuẩn hóa

$$z_j = \begin{bmatrix} \frac{(x_{j1} - \bar{x}_1)}{\sqrt{S_{11}}} \\ \frac{(x_{j2} - \bar{x}_2)}{\sqrt{S_{22}}} \\ \vdots \\ \frac{(x_{jp} - \bar{x}_p)}{\sqrt{S_{pp}}} \end{bmatrix}$$

với $j = 1, 2, \ldots, n$

Ma trận hiệp phương sai mẫu của các biến số chuẩn hóa là ma trận tương quan mẫu R của các quan sát $x_1, x_2, \ldots, x_n$. Chuẩn hóa tránh được vấn đề của việc có một biến số với phương sai lớn ảnh hưởng không đúng mực đến việc xác định các tải nhân tố.

Biểu diễn trong (công thức 9-13), khi áp dụng cho ma trận hiệp phương sai mẫu S hoặc ma trận tương quan mẫu R, được biết đến là *giải pháp thành phần chính*.

Giải pháp Thành phần Chính của Mô hình Nhân tố (Principal Component Solution of the Factor Model)

Phân tích nhân tố thành phần chính của ma trận hiệp phương sai mẫu S được xác định theo các cặp trị riêng - vectơ riêng $(\hat{\lambda}_1, \hat{e}_1), (\hat{\lambda}_2, \hat{e}_2), \ldots, (\hat{\lambda}_p, \hat{e}_p)$, với $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p$. Giả sử $m < p$ là số lượng nhân tố chung. Sau đó ma trận của các tải nhân tố ước lượng $\{\hat{\ell}_{ij}\}$ được cho bởi

$$\hat{L} = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1 \mid \sqrt{\hat{\lambda}_2}\hat{e}_2 \mid \ldots \mid \sqrt{\hat{\lambda}_m}\hat{e}_m\right]$$

Các phương sai cụ thể ước lượng được cung cấp bởi các phần tử đường chéo của ma trận $S - \hat{L}\hat{L}'$, do đó

$$\hat{\Psi} = \begin{bmatrix} \hat{\psi}_1 & 0 & \ldots & 0 \\ 0 & \hat{\psi}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \hat{\psi}_p \end{bmatrix}$$

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

Các $\hat{h}_i^2$ được ước lượng như sau

$$\hat{h}_i^2 = \hat{\ell}_{i1}^2 + \hat{\ell}_{i2}^2 + \ldots + \hat{\ell}_{im}^2$$

Phân tích nhân tố thành phần chính của ma trận tương quan mẫu được bắt đầu bằng cách sử dụng R thay cho S.

Đối với giải pháp thành phần chính, các tải ước lượng cho một nhân tố nhất định không thay đổi khi số lượng nhân tố được tăng lên. Ví dụ, nếu $m = 1$,

$$\hat{L} = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1\right],$$

và nếu $m = 2$,

$$\hat{L} = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1 \mid \sqrt{\hat{\lambda}_2}\hat{e}_2\right],$$

trong đó $(\hat{\lambda}_1, \hat{e}_1)$ và $(\hat{\lambda}_2, \hat{e}_2)$ là hai cặp trị riêng - vectơ riêng đầu tiên cho S (hoặc R).

Theo định nghĩa của $\hat{\psi}_i$, các phần tử đường chéo của S bằng với các phần tử đường chéo của $\hat{L}\hat{L}' + \hat{\Psi}$. Tuy nhiên, các phần tử nằm ngoài đường chéo của S không thường xuyên được tái tạo bởi $\hat{L}\hat{L}' + \hat{\Psi}$. Vậy thì, làm thế nào chúng ta chọn số lượng nhân tố m?

Nếu số lượng nhân tố chung không được xác định bởi các cân nhắc a priori, như theo lý thuyết hoặc công trình nghiên cứu của các nhà khoa học khác, sự

chọn lựa của m có thể dựa trên các trị riêng ước lượng tương tự như với các thành phần chính. Hãy xem xét ma trận dư

$$S - (\hat{L}\hat{L}' + \hat{\Psi})$$

kết quả từ phép xấp xỉ của S bởi giải pháp thành phần chính. Nếu các phần tử đường chéo là không, và nếu các phần tử khác cũng nhỏ, chúng ta có thể chủ quan chọn mô hình nhân tố m là phù hợp.

Tổng bình phương các phần tử của $S-(\hat{L}\hat{L}'+\hat{\Psi})$ nhỏ hơn hoặc bằng $\hat{\lambda}_{m+1}^2 + \ldots + \hat{\lambda}_p^2$

Vì thế, một giá trị nhỏ cho tổng bình phương các phần tử sai số xấp xỉ là điều mong muốn. Lý tưởng nhất, các đóng góp của vài nhân tố đầu tiên đối với phương sai mẫu của các biến số nên lớn. Đóng góp vào phương sai mẫu s_{ii} từ nhân tố chung đầu tiên là $\hat{\ell}_{i1}^2$. Đóng góp vào *tổng phương sai mẫu*, $s_{11} + s_{22} + \ldots + s_{pp} = tr(S)$, từ nhân tố chung đầu tiên là

$$\hat{\ell}_{11}^2 + \hat{\ell}_{21}^2 + \ldots + \hat{\ell}_{p1}^2 = (\sqrt{\hat{\lambda}_1}\hat{e}_1)'(\sqrt{\hat{\lambda}_1}\hat{e}_1) = \hat{\lambda}_1$$

vì vectơ riêng $\hat{e}_1$ có độ dài đơn vị.

Tỷ lệ của tổng phương sai mẫu do nhân tố thứ j là:

$$\begin{cases} \dfrac{\hat{\lambda}_j}{s_{11} + s_{22} + \ldots + s_{pp}} & \text{đối với phân tích nhân tố của } S \\ \dfrac{\hat{\lambda}_j}{p} & \text{đối với phân tích nhân tố của } R \end{cases}$$

Điều này thường xuyên được sử dụng như một công cụ thử nghiệm để xác định số lượng nhân tố chung thích hợp. Số lượng nhân tố chung giữ lại trong mô hình tăng lên cho đến khi một "tỷ lệ phù hợp" của tổng phương sai mẫu được giải thích.

Một quy ước khác thường gặp trong các chương trình máy tính là đặt m bằng số lượng trị riêng của R lớn hơn một nếu ma trận tương quan được phân tách, hoặc bằng số lượng trị riêng dương của S nếu ma trận hiệp phương sai mẫu được phân tách. Những quy tắc ngón tay cái này không nên áp dụng một cách không phân biệt.

Ví dụ, $m = p$ nếu quy tắc cho S được tuân thủ, vì tất cả các trị riêng đều được kỳ vọng là dương với kích thước mẫu lớn. Phương pháp tốt nhất là giữ lại ít hơn thay vì nhiều nhân tố, giả định rằng chúng cung cấp một sự giải thích thỏa đáng cho dữ liệu và tạo ra một phép đo ứng với S hoặc R một cách chấp nhận được.

1.2.2. Phương Pháp Ước Lượng Hợp Lý Cực Đại (The Maximum Likelihood Method)

Nếu các nhân tố chung F và các nhân tố ϵ có thể được giả định có phân phối chuẩn, thì ước lượng hợp lý cực đại của các tải số nhân tố và phương

sai có thể được thu được. Khi F_j và ϵ_j đồng phân phối chuẩn, các quan sát $X_j - \mu = LF_j + \epsilon_j$ sau đó là chuẩn, và từ, hàm hợp lý cực đại là

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

$$= (2\pi)^{\frac{np}{2}} |\Sigma|^{\frac{(n-1)p}{2}} e^{-\frac{(n-1)}{2} tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})'\right)\right]} \times (2\pi)^{\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{n}{2}(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)} \quad (2)$$

mô hình này phụ thuộc vào L và Ψ thông qua $\Sigma = LL' + \Psi$. Mô hình này vẫn chưa được định nghĩa rõ ràng, vì có nhiều sự lựa chọn cho L được tạo ra bởi các phép biến đổi trực giao. Điều mong muốn là làm cho L được định nghĩa tốt bằng cách áp đặt điều kiện duy nhất thuận tiện về mặt tính toán

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

là một ma trận đường chéo.

Các ước lượng hợp lý tối đa $\hat{L}$ và $\hat{\Psi}$ phải được thu được thông qua tối ưu hóa số học.

Kết quả 9.1.

Giả sử $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên có phân phối $N(\mu, \Sigma)$, trong đó: $\Sigma = LL' + \Psi$ là ma trận hiệp phương sai cho mô hình m nhân tố chung của (công thức 9-4). Các ước lượng hợp lý cực đại $\hat{L}, \hat{\Psi}$, và $\hat{\mu} = \bar{x}$ với điều kiện $L'\Psi^{-1}L$ là một ma trận đường chéo.

Các ước lượng hợp lý cực đại của $\hat{h}_i^2$ là

$$\hat{h}_i^2 = \hat{\lambda}_{i1}^2 + \hat{\lambda}_{i2}^2 + \cdots + \hat{\lambda}_{im}^2 \quad \text{với} \quad i = 1, 2, \ldots, p \quad (3)$$

vậy

$$\left(\text{Tỉ lệ phương sai mẫu tổng cộng}\right) \text{do nhân tố thứ } j \text{ là} \quad (4)$$

$$= \frac{\hat{\lambda}_{1j}^2 + \hat{\lambda}_{2j}^2 + \cdots + \hat{\lambda}_{pj}^2}{s_{11} + s_{22} + \cdots + s_{pp}} \quad (5)$$

1.3. Xoay nhân tố (Factor Rotation)

Tất cả các tải trọng nhân tố thu được từ tải trọng ban đầu thông qua một phép biến đổi trực giao đều có khả năng tái tạo ma trận hiệp phương sai (hoặc tương quan) như nhau. Từ đại số ma trận, chúng ta biết rằng một phép biến đổi trực giao tương ứng với một phép xoay cứng (hoặc phản xạ) của các trục tọa độ. Vì lý do này, một phép biến đổi trực giao của tải trọng nhân tố, cũng như biến đổi trực giao ngụ ý của các nhân tố, được gọi là *phép xoay nhân tố*.

Nếu $\hat{L}$ là ma trận $p \times m$ của tải trọng nhân tố ước lượng thu được bằng bất kỳ phương pháp nào (thành phần chính, khả năng tối đa, và như thế), sau đó

$$\hat{L}^* = \hat{L}T, \tag{6}$$

Vì các tải trọng gốc có thể không dễ dàng giải thích được, nên thực hành thông thường là xoay chúng cho đến khi đạt được một cấu trúc "đơn giản" hơn. Lý luận này rất giống với việc mài nhọn tiêu điểm của một kính hiển vi để nhìn thấy chi tiết rõ ràng hơn.

Lý tưởng nhất, chúng ta muốn thấy một mô hình tải trọng sao cho mỗi biến số tải trọng cao trên một nhân tố duy nhất và có tải trọng nhỏ đến trung bình trên các nhân tố còn lại. Tuy nhiên, không phải lúc nào cũng có thể có được cấu trúc đơn giản này, mặc dù các tải trọng đã xoay cho dữ liệu mười môn phối hợp được thảo luận trong Ví dụ 9.11 cung cấp một mô hình gần như lý tưởng.

Chúng ta sẽ tập trung vào các phương pháp đồ họa và phân tích để xác định một phép xoay trực giao để có được cấu trúc đơn giản. Khi $m = 2$, hoặc các nhân tố chung được xem xét từng đôi một, sự chuyển đổi thành cấu trúc đơn giản thường có thể được xác định một cách đồ họa. Các nhân tố chung không tương quan được coi là đơn vị. trong đó $TT' = T'T = I$ (9-42)

là ma trận $p \times m$ của tải trọng "xoay". Hơn nữa, ma trận hiệp phương sai (hoặc tương quan) ước lượng vẫn không đổi, vì

$$\hat{L}\hat{L}' + \hat{\Psi} = \hat{L}T\hat{L}'T' + \hat{\Psi} = \hat{L}^*\hat{L}^{*'} + \hat{\Psi} \tag{7}$$

(9-43)

Phương trình (9-43) chỉ ra rằng ma trận dư, $S_n - \hat{L}\hat{L}' - \hat{\Psi} = S_n - \hat{L}^*\hat{L}^{*'} - \hat{\Psi}$, vẫn không đổi. Hơn nữa, các phương sai cụ thể $\hat{\psi}_i$, và do đó các độ chung $\hat{h}_i^2$, không thay đổi. Như vậy, từ góc độ toán học, không quan trọng là ta có ma trận $\hat{L}$ hay $\hat{L}^*$.

Việc vẽ các vectơ dọc theo các trục tọa độ vuông góc. Một biểu đồ của các cặp tải trọng nhân tố $(\hat{e}_{i1}, \hat{e}_{i2})$ cho ra p điểm, mỗi điểm tương ứng với một biến số. Các trục tọa độ sau đó có thể được xoay một cách trực quan qua một góc – gọi là ϕ – và các tải trọng xoay mới $\hat{e}'_{ij}$ được xác định từ các mối quan hệ

$$\hat{L}^* = \hat{L}\,T \quad (\mathrm{p} \times 2) \tag{8}$$

$$(\mathrm{p} \times 2)\,(2 \times 2) \quad (9-44) \tag{9}$$

trong đó

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

Mối quan hệ trong (9-44) hiếm khi được thực hiện trong một phân tích đồ họa hai chiều. Trong tình huống này, các cụm biến số thường rõ ràng khi quan

sát, và các cụm này cho phép xác định các nhân tố chung mà không cần kiểm tra độ lớn của các tải trọng đã xoay. Mặt khác, đối với $m > 2$, các hướng không dễ dàng hình dung được, và độ lớn của các tải trọng xoay cần được kiểm tra để tìm ra một giải thích ý nghĩa của dữ liệu gốc. Sự lựa chọn của một ma trận trực giao T thỏa mãn một tiêu chí phân tích của cấu trúc đơn giản sẽ được xem xét ngay sau.

Phép Xoay Xiên

Phép xoay trực giao phù hợp cho mô hình nhân tố trong đó các nhân tố chung được giả định là độc lập. Nhiều nhà nghiên cứu trong khoa học xã hội xem xét phép xoay xiên (không trực giao), cũng như phép xoay trực giao. Các phép xoay đầu tiên được gọi là *oblique rotations*.

Điều này thường được đề xuất sau khi người ta xem xét các tải trọng nhân tố ước lượng và nhận thấy chúng không theo mô hình được đề xuất. Tuy nhiên, phép xoay xiên thường xuyên là một công cụ hữu ích trong phân tích nhân tố.

Nếu chúng ta xem xét các nhân tố chung m là các trục tọa độ, điểm với các tọa độ $(\hat{e}_{i1}, \hat{e}_{i2}, \ldots, \hat{e}_{im})$ đại diện cho vị trí của biến số thứ i trong không gian nhân tố. Giả định rằng các biến số được nhóm vào các cụm không chồng lấn, một phép xoay trực giao để có được cấu trúc đơn giản tương ứng với một phép xoay cứng của các trục tọa độ sao cho, sau phép xoay, các trục đi qua càng gần các cụm càng tốt. Một phép xoay xiên để có được cấu trúc đơn giản tương ứng với một phép xoay không cứng của hệ thống tọa độ sao cho các trục xoay (không còn vuông góc) đi (gần như) qua các cụm. Một phép xoay xiên tìm cách biểu thị mỗi biến số theo ít nhân tố nhất có thể—ưu tiên là một nhân tố duy nhất. Phép xoay xiên đã được thảo luận trong nhiều nguồn (xem, ví dụ, [6] hoặc [10]) và sẽ không được tiếp tục trình bày trong cuốn sách này.

1.4. Điểm số của nhân tố

Trong phân tích nhân tố, sự quan tâm thường tập trung vào các tham số trong mô hình nhân tố. Tuy nhiên, các giá trị ước lượng của các nhân tố chung, được gọi là *điểm số nhân tố*, cũng có thể cần thiết. Những đại lượng này thường được sử dụng cho mục đích chẩn đoán, cũng như là đầu vào cho một phân tích tiếp theo.

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

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

Việc ước lượng trở nên phức tạp do thực tế là các đại lượng không quan sát được f_j và e_j vượt quá số lượng quan sát x_j. Để vượt qua khó khăn này, một số

phương pháp tiếp cận khá độc đáo nhưng có lý, đã được đề xuất để ước lượng giá trị nhân tố. Hai trong số các phương pháp tiếp cận này.

Cả hai phương pháp tiếp cận điểm số nhân tố đều có hai yếu tố chung:

1. Xử lý các ước lượng tải nhân tố $\hat{\lambda}_{ij}$ và phương sai cụ thể $\hat{\psi}_i$ như thể chúng là giá trị thực.

2. Liên quan đến các phép biến đổi tuyến tính của dữ liệu gốc, có thể đã được chuẩn hóa hoặc tiêu chuẩn hóa. Thông thường, các tải nhân tố *rotated* được ước lượng thay vì các tải nhân tố gốc được ước lượng, được sử dụng để tính toán điểm số nhân tố. Các công thức tính toán được đưa ra trong phần này, không thay đổi khi các tải nhân tố quay được thay thế cho các tải nhân tố không quay, do đó chúng tôi sẽ không phân biệt giữa chúng.

1.4.1. Phương pháp bình phương tối thiểu có trọng số (The Weighted Least Squares Method)

Giả sử trước hết rằng, vector trung bình μ, các tải số nhân tố L, và phương sai đặc thù Ψ được biết đến cho mô hình nhân tố

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

Xem xét các yếu tố cụ thể $\epsilon' = [\epsilon_1, \epsilon_2, \ldots, \epsilon_p]$ như là các sai số. Do $Var(\epsilon_i) = \psi_i, i = 1, 2, \ldots, p$ không nhất thiết phải bằng nhau, Bartlett đã đề xuất rằng phương pháp bình phương tối thiểu có trọng số có thể được sử dụng để ước lượng các giá trị nhân tố chung.

Tổng bình phương của các sai số, được trọng số bằng nghịch đảo của phương sai của chúng, là

$$\sum_{i=1}^{p} \frac{\epsilon_i^2}{\psi_i} = \epsilon' \Psi^{-1} \epsilon = (x - \mu - LF)' \Psi^{-1} (x - \mu - LF)$$

Bartlett đề xuất chọn các ước lượng $\hat{f}$ của f để tối thiểu hóa. Giải pháp là

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

Chúng ta lấy các ước lượng L, Ψ, và $\mu = \bar{x}$ như là các giá trị thực và thu được điểm số nhân tố cho trường hợp thứ j như sau

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

Khi L và Ψ được xác định bằng phương pháp cực đại, những ước lượng này phải thoả mãn điều kiện duy nhất, $L' \Psi^{-1} L = \Delta$ là một ma trận đường chéo. Ta có:

Điểm Số Nhân Tố Được Thu Thập Bằng Phương Pháp Bình Phương Tối Thiểu Có Trọng Số từ Ước Lượng Khả Năng Tối Đa

$$\hat{f}_j = (\hat{L}'\hat{\Psi}^{-1}\hat{L})^{-1}\hat{L}'\hat{\Psi}^{-1}(x_j - \hat{\mu})$$
$$= \hat{\Delta}^{-1}\hat{L}'\hat{\Psi}^{-1}(x_j - \bar{x}), \quad j = 1, 2, ..., n$$

hoặc, nếu ma trận tương quan được phân tích

$$\hat{f}_j = (\hat{L}_z'\hat{\Psi}_z^{-1}\hat{L}_z)^{-1}\hat{L}_z'\hat{\Psi}_z^{-1}z_j$$
$$= \hat{\Delta}_z^{-1}\hat{L}_z'\hat{\Psi}_z^{-1}z_j, \quad j = 1, 2, ..., n$$

trong đó $z_j = D^{-1/2}(x_j - \bar{x})$, như trong (8-25), và $\hat{\rho} = \hat{L}_z\hat{L}_z' + \hat{\Psi}_z$.

Điểm số nhân tố được tạo ra bởi (9-50) có vector trung bình mẫu $\mathbf{0}$ và ma trận hiệp phương sai mẫu $\mathbf{0}$. (Xem Bài tập 9.16.)

Nếu các tải số nhân tố quay $\mathbf{L}^* = \mathbf{LT}$ được sử dụng thay cho các tải số gốc trong (9-50), các điểm số nhân tố tiếp theo $\hat{f}_j^*$, liên quan đến $\hat{f}_j$ bởi $\hat{f}_j^* = \mathbf{T}\hat{f}_j$, $j = 1, 2, \ldots, n$.

Nhận xét. Nếu các tải số nhân tố được ước lượng bởi phương pháp thành phần chính, thông thường người ta sẽ tạo ra điểm số nhân tố sử dụng một quy trình bình phương tối thiểu không trọng số (thông thường). Giả định ngầm này tương đương với việc giả sử rằng các ψ_i gần như bằng nhau. Các điểm số nhân tố được tính như sau

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

hoặc

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

đối với dữ liệu được chuẩn hóa. Vì $\tilde{\mathbf{L}} = [\sqrt{\lambda_1}\mathbf{e}_1|\sqrt{\lambda_2}\mathbf{e}_2|\ldots|\sqrt{\lambda_m}\mathbf{e}_m]$, chúng ta có

$$\hat{f}_j = \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}}\mathbf{e}_1'(\mathbf{x}_j - \bar{\mathbf{x}}) \\ \frac{1}{\sqrt{\lambda_2}}\mathbf{e}_2'(\mathbf{x}_j - \bar{\mathbf{x}}) \\ \vdots \\ \frac{1}{\sqrt{\lambda_m}}\mathbf{e}_m'(\mathbf{x}_j - \bar{\mathbf{x}}) \end{bmatrix}$$

Đối với những điểm số nhân tố này, chúng ta có

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

và

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

chúng ta thấy rằng $\hat{f}_j$ không gì khác hơn là m thành phần chính đầu tiên (đã được điều chỉnh tỉ lệ), được đánh giá tại x_j.

1.4.2. Phương Pháp Hồi Quy

Bắt đầu lại với mô hình nhân tố gốc $X-\mu = LF+\epsilon$, chúng ta ban đầu xử lý ma trận tải số L và ma trận phương sai cụ thể Ψ như đã biết. Khi các nhân tố chung F và các nhân tố cụ thể (hoặc sai số) ϵ có phân phối chuẩn chung với trung bình và hiệp phương sai được cho bởi (9-3), tổ hợp tuyến tính $X-\mu = LF+\epsilon$ có phân phối $N_p(0, LL'+\Psi)$. (Xem Kết quả 4.3.) Hơn nữa, phân phối chung của $(X-\mu)$ và F là $N_{m+p}(0,\Sigma^*)$, nơi

$$\Sigma^* = \begin{bmatrix} LL'+\Psi & L \\ L' & I \end{bmatrix}$$

và 0 là vector $(m+p)\times 1$ của các số không. Sử dụng Kết quả 4.6, chúng ta thấy rằng phân phối có điều kiện của F khi biết X là phân phối chuẩn đa biến với

trung bình $= E(F|X) = L'\Sigma^{-1}(X-\mu) = L'(LL'+\Psi)^{-1}(X-\mu)$

và

hiệp phương sai $= Cov(F|X) = I - L'\Sigma^{-1}L = I - L'(LL'+\Psi)^{-1}L$

Các số lượng $L'(LL'+\Psi)^{-1}$ trong (9-53) là các hệ số trong một hồi quy (đa biến) của các nhân tố trên các biến. Ước lượng của các hệ số này sản xuất ra các điểm số nhân tố tương đồng với các ước lượng của giá trị trung bình có điều kiện trong phân tích hồi quy đa biến. (Xem Chương 7.) Do đó, với bất kỳ vector quan sát x_j, và lấy ước lượng khả năng tối đa $\hat{L}$ và $\hat{\Psi}$ làm giá trị thực, chúng ta thấy rằng vector điểm số nhân tố thứ j được cho bởi

$$\hat{f}_j = \hat{\Sigma}^{-1}(x_j-\hat{\mu}) = L'(LL'+\Psi)^{-1}(x_j-\bar{x}), \quad j=1,2,...,n$$

Điểm Số Nhân Tố Được Tính Bằng Hồi Quy

$$\hat{f}_j = \hat{\mathbf{L}}'\mathbf{S}^{-1}(\mathbf{x}_j-\bar{\mathbf{x}}), \quad j=1,2,\ldots,n$$

hoặc, nếu ma trận tương quan được phân tích,

$$\hat{f}_j = \hat{\mathbf{L}}_z'\mathbf{R}_z^{-1}\mathbf{z}_j, \quad j=1,2,\ldots,n$$

trong đó,

$$\mathbf{z}_j = \mathbf{D}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}}) \quad \text{và} \quad \hat{\mathbf{P}} = \hat{\mathbf{L}}_z\hat{\mathbf{L}}_z' + \hat{\mathbf{\Psi}}_z$$

1.5. Các kiểm định cần thiết

Kiểm định mô hình nhân tố trên mẫu lớn

Với các quan trắc $x_1, x_2, \ldots, x_N$, nhận từ mẫu ngẫu nhiên kích thước N lấy từ phân phối chuẩn nhiều chiều $N_p(\mu, \Sigma)$, ta có thể kiểm định tính hợp lý của mô hình nhân tố $\Sigma = \mathbf{L}\mathbf{L}' + \Psi$ bằng cách xét:

$$\mathbf{\Theta} = \{\theta | \theta \text{ là tập các ma trận đối xứng xác định dương cấp } p \times p,$$

$$\theta_0 = \{\Sigma | \Sigma = \mathbf{L}\mathbf{L}' + \Psi, \Psi = \text{diag}(\psi_1, \psi_2, \ldots, \psi_p)\}$$

và khảo sát bài toán kiểm định giả thuyết

$$H_0 : \Sigma \in \Theta_0 \quad (\text{hay } H_0 : \text{Mô hình nhân tố là phù hợp})$$

so với

$$H_A : \Sigma \notin \Theta_0 \quad (\text{hay } H_A : \Sigma \text{ Là ma trận xác định dương bất kỳ})$$

bằng phép kiểm định tỷ số hợp lý tổng quát. Ta có

$$\max_{\mu,\Sigma} L(\mu, \Sigma) = \frac{1}{(2\pi)^{np/2}|\mathbf{S}_n|^{n/2}} \exp\left(-\frac{np}{2}\right),$$

trong đó $\mathbf{S}_n$ là ma trận hiệp phương sai mẫu (không hiệu chỉnh), và

$$\max_{\mu,\Sigma\in\Theta_0} L(\mu, \Sigma) = \frac{1}{(2\pi)^{np/2}|\hat{\mathbf{\Sigma}}|^{n/2}} \exp\left(-\frac{1}{2}\text{tr}\left(\hat{\mathbf{\Sigma}}^{-1}\sum_{j=1}^{N}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'\right)\right)$$

$$= \frac{1}{(2\pi)^{np/2}}|\mathbf{L}\mathbf{L}' + \hat{\mathbf{\Psi}}|^{n/2} \exp\left(-\frac{1}{2}n\text{tr}\left((\mathbf{L}\mathbf{L}' + \hat{\mathbf{\Psi}})^{-1}\mathbf{S}_n\right)\right).$$

với $\hat{\mu} = \bar{x}, \hat{\Sigma} = \mathbf{L}\mathbf{L}' + \hat{\Psi}$, trong đó $\mathbf{L}$ và $\hat{\Psi}$ lần lượt là ước lượng hợp lý cực đại của $\mathbf{L}$ và Ψ. Hơn nữa, do $\text{tr}\left((\mathbf{L}\mathbf{L}' + \hat{\Psi})^{-1}\mathbf{S}_n\right) = p$, nên

$$\max_{\mu,\Sigma\in\Theta_0} L(\mu, \Sigma) = \frac{1}{(2\pi)^{np/2}|\mathbf{L}\mathbf{L}' + \hat{\Psi}|^{n/2}} \exp\left(-\frac{1}{2}np\right).$$

Ta được thông kê Wilks lambda,

$$\Lambda = \frac{\max_{\mu,\Sigma\in\Theta_0} L(\mu, \Sigma)}{\max_{\mu,\Sigma} L(\mu, \Sigma)} = \left(\frac{|\mathbf{L}\mathbf{L}' + \hat{\Psi}|}{|\mathbf{S}_n|}\right)^{-n/2}$$

Với mẫu lớn:

$$-2\ln\Lambda = -2\ln\left(\frac{|\mathbf{LL}' + \hat{\Psi}|}{|\mathbf{S}_n|}\right)^{-n/2} = n\ln\left(\frac{|\mathbf{LL}' + \hat{\Psi}|}{|\mathbf{S}_n|}\right)$$

có phân phối gần đúng với phân phối Chi-bình phương với $v - v_0$ bậc tự do, khi H_0 đúng. Vì $v = \frac{1}{2}p(p+1)$ và $v_0 = p(m+1) - \frac{1}{2}m(m-1)$

Hơn nữa, bằng cách thay giá trị n bằng thừa số nhân $(n-1-(2p+4m+5)/6)$,

$$(n - 1 - (2p + 4m + 5)/6)\ln\left(\frac{|\mathbf{LL}' + \mathbf{\Psi}|}{|\mathbf{S}_n|}\right)$$

có thể xấp xỉ tốt hơn bằng phân phối Chi-bình phương với $\frac{1}{2}(p-m)^2 - p - m$ bậc tự do. Với hiệu chỉnh Barlett khi n và $n - p$ lớn, ta bác bỏ H_0 ở ngưỡng kiểm định α khi

$$(n - 1 - (2p + 4m + 5)/6)\ln\left(\frac{|\mathbf{LL}' + \mathbf{\Psi}|}{|\mathbf{S}_n|}\right) > \chi^2_{\frac{1}{2}(p-m)^2-p-m,1-\alpha},$$

trong đó $\chi^2_{(p-m)^2-p-m,\gamma}$ là phần vị ở mức xác suất $\gamma = 1-\alpha$ của phân phối Chi-bình phương với $\frac{1}{2}(p-m)^2 - p - m$ bậc tự do. Do đó bậc tự do $\frac{1}{2}(p-m)^2 - p - m$ phải là số dương, số các nhân tố cần thỏa điều kiện $m \leq \frac{1}{2}(2p + 1 - \sqrt{8p+1})$.

II. Phần bài tập (5 điểm)

Bài 4.2.1 - Trang 205 Sách Johnson

$X_1, \ldots, X_{60}$ là mẫu ngẫu nhiên lấy từ phân phối chuẩn 4 chiều với μ và Σ.

a. Phân phối của: $\bar{X}$

Áp dụng mệnh đề 2.2.5 ta có:

$$\bar{X} \sim N_p(\mu, \frac{1}{n}\Sigma)$$

Vậy với $X_1, \ldots, X_{60}$ có: $n = 60$, $p = 4$

$$\Rightarrow \bar{X} \sim N_4(\mu, \frac{1}{60}\Sigma)$$

b. Phân phối của $(X_1 - \mu)'\Sigma^{-1}(X_1 - \mu)$

Áp dụng định lý 2.1.4 ta có: Nếu $X \sim N_p(\mu, \Sigma)$ thì

$$U = (X_1 - \mu)'\Sigma^{-1}(X_1 - \mu) \sim \chi^2(p)$$

Với $p = 4 \Rightarrow U \sim \chi^2(4)$

c. Phân phối của: $n(\bar{X} - \mu)'\Sigma^{-1}(\bar{X} - \mu)$

Áp dụng định lý 2.1.4 cho $\bar{X} \sim N_p(\mu, \frac{1}{n}\Sigma)$ hay cho $\sqrt{n}(\bar{X} - \mu) \sim N_p(0, \Sigma)$

Ta được $n(\bar{X} - \mu)'\Sigma^{-1}(\bar{X} - \mu) \sim \chi^2(p)$

Với p = 4, ta được: $n(\bar{X} - \mu)'\Sigma^{-1}(\bar{X} - \mu) \sim \chi^2(4)$

d. Phân phối xấp xỉ của: $n(\bar{X} - \mu)'\mathrm{S}^{-1}(\bar{X} - \mu)$

Từ câu c ta có: $n(\bar{X} - \mu)'\Sigma^{-1}(\bar{X} - \mu) \sim \chi^2(4)$. Do đó khi $\bar{X}$ có phân phối xấp xỉ chuẩn, phân phối mẫu $n(\bar{X} - \mu)'\Sigma^{-1}(\bar{X} - \mu)$ được xấp xỉ bằng phân phối Chi-bình phương với p = 4 bậc tự do. Việc thay Σ^{-1} bằng S^{-1} không ảnh hưởng nhiều đến xấp xỉ này khi n-p lớn. Vậy $n(\bar{X} - \mu)'\mathrm{S}^{-1}(\bar{X} - \mu) \sim \chi^2(4)$