MULTIVARIATE STATISTICAL ANALYSIS

Lecture 5
Factor Analysis

Associate Professor Lý Quốc Ngọc

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

fit@hcmus

Contents

5.1. Purpose of FA

Factor analysis is a statistical method used to describe the relationship between observed correlated variables and new unobserved variables called factors..

5.2. Problem statement

Let the observable random vector $X = (X_1, X_2, ..., X_p)'$ have
Mean vector $\mu = (\mu_1, \mu_2, ..., \mu_p)'$
and Covariance matrix $\sum$
The factor model postulates that X is linearly dependent upon a few unobservable random variables $F_1, F_2, ..., F_m$ called common factors.

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + ... + l_{1m}F_m + \varepsilon_1$$
$$X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + ... + l_{2m}F_m + \varepsilon_2$$
$$\vdots$$
$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + ... + l_{pm}F_m + \varepsilon_p$$

Or, in matrix notation:

$$X - \mu = L.F + \varepsilon$$

5.2. Problem statement

The coefficient l_{ij} is called the loading of the i[th] variable on the j[th] factor..

The p deviations $X_1 - \mu_1, X_2 - \mu_2, ..., X_p - \mu_p$ are expressed in terms of p+m random variables $F_1, F_2, ..., F_m$ $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$ which are unobservable..

.

5.2. Problem statement

Some assumptions about orthogonal factor models for random vectors F and ε :

$$E(F) = 0, \qquad Cov(F) = E[FF'] = I$$

$$E(\varepsilon) = 0, \qquad Cov(\varepsilon) = E[\varepsilon\varepsilon'] = \Psi = \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \Psi_p \end{bmatrix}$$

.

$$Cov(\varepsilon, F) = E[\varepsilon\, F'] = 0$$

5.3. Method

The orthogonal factor model implies a covariance structure for X

$$\sum = Cov(X) = E[(X-\mu)(X-\mu)']$$
$$= E[(LF+\varepsilon)(LF+\varepsilon)']$$
$$= E[LF(LF)'+\varepsilon(LF)'+LF\varepsilon'+\varepsilon\varepsilon']$$
$$= LE[FF']L'+E[\varepsilon F']+LE[F\varepsilon']+E[\varepsilon\varepsilon']$$
$$= LL'+\Psi$$

$$Cov(X,F) = E[(X-\mu).F'] = LE[FF']+E[\varepsilon F'] = L$$

5.3. Method

Covariance Structure for the orthogonal factor model

$$1)\ Cov(X) = LL' + \Psi$$

$$Var(X_i) = l_{i1}^2 + l_{i2}^2 \ldots + l_{i1}^2 + \Psi_i$$

$$Cov(X_i, X_k) = l_{i1}l_{k1} + l_{i2}l_{k2} + \ldots + l_{im}l_{km}$$

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

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

5.3. Method

We see that:

$$Var(X_i) = l_{i1}^2 + l_{i2}^2 \ldots + l_{i1}^2 + \Psi_i$$
$$= h_i^2 + \Psi_i$$

Therefore:

Part of the variance of X_i is contributed from the sum of squares of each factor loading, and a part is donated from the variance of the error random variable.

In other words, the variance and covariance come from the random variables of the random vector that can be reconstructed from the orthogonal factor model.

5.3. Method

Use **principal component analysis** to analyze the covariance matrix.

The spectral decomposition provides us with one factoring of the covariance matrix $\sum$:

$$\sum = \lambda_1 e_1 (e_1)' + \lambda_2 e_2 (e_2)' + \ldots + \lambda_p e_p (e_p)'$$

$$= \left[\sqrt{\lambda_1}\, e_1 \sqrt{\lambda_2}\, e_2 \cdots \sqrt{\lambda_p} e_p\right] \begin{bmatrix} \sqrt{\lambda_1}\, e_1' \\ \sqrt{\lambda_2}\, e_2' \\ \vdots \\ \sqrt{\lambda_p}\, e_p' \end{bmatrix}$$

5.3. Method

It is necessary to find factor analysis models with fewer variables to explain the covariance matrix..

When the last p-m eigenvalues are small, is to neglect their contribution to $\sum$. Neglecting this contribution, we obtain the approximation :

$$\Sigma \approx \left[\sqrt{\lambda_1}\, e_1 \sqrt{\lambda_2}\, e_2 \cdots \sqrt{\lambda_m} e_m\right] \begin{bmatrix} \sqrt{\lambda_1}\, e_1' \\ \sqrt{\lambda_2}\, e_2' \\ \vdots \\ \sqrt{\lambda_m}\, e_m' \end{bmatrix} = \underset{(p\times m)}{L} \cdot \underset{(m\times p)}{L'}$$

5.3. Method

Allowing for the specific factors ε, we find that the approximation becomes:

$$\sum \approx LL' + \Psi$$

$$\sum \approx \left[\sqrt{\lambda_1}\, e_1 \sqrt{\lambda_2}\, e_2 \cdots \sqrt{\lambda_m}\, e_m\right] \begin{bmatrix} \sqrt{\lambda_1}\, e_1' \\ \sqrt{\lambda_2}\, e_2' \\ \vdots \\ \sqrt{\lambda_m}\, e_m' \end{bmatrix} + \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \Psi_p \end{bmatrix}$$

$$\Psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2, i = 1,2,\ldots,p$$

5.3. Method

Sau khi tìm được ma trận tải, cần tìm F .
Giả định rằng chúng ta tìm được vector trung bình μ, ma trận hệ số tải L và phương sai nhân tố lỗi Ψ của mô hình nhân tố:

$$X - \mu = L.F + \varepsilon$$

Tổng bình phương sai số có trọng số:

$$error(f) = \sum_{i=1}^{p} \frac{\varepsilon_i^2}{\Psi_i} = \varepsilon' \Psi^{-1} \varepsilon = (x - \mu - Lf)' \Psi^{-1} (x - \mu - Lf)$$

$$Var(\varepsilon_i) = \Psi_i$$

Sai số $error(f)$ đạt cực tiểu tại f' sao cho:

$$\frac{\partial error(f)}{\partial f} = 0 \Rightarrow f' = (L' \Psi^{-1} l)^{-1} L' \Psi^{-1} (x - \mu)$$

5.3. Phương pháp

Sử dụng **phương pháp ước lượng triển vọng cực đại** để phân tích ma trận hiệp phương sai.

Bàn luận trên lớp
Các câu hỏi gợi ý:

5.4. Ý nghĩa hình học

Bàn luận trên lớp
Các câu hỏi gợi ý:

5.5. Kiểm chứng tính đúng đắn của mô hình

Bàn luận trên lớp
Các câu hỏi gợi ý:

5.6. Ví dụ

Bàn luận trên lớp
Các câu hỏi gợi ý: