Tutorial 5

CHEN Xiao

Department of Mathematics

April 3, 2020

Principal Component Analysis

Results 5.1 Let Σ be the covariance matrix associated with the random vector $X' = [X_1, X_2, \ldots, X_p]$. Let Σ have the eigenvalue-eigenvector pair $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then the ith principal component is given by

$$Y_i = \mathbf{e}_i' \boldsymbol{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, i = 1, 2, \ldots, p$$

With these choices,

$$\begin{aligned} \text{Var}(Y_i) &= \mathbf{e}_i' \boldsymbol{\Sigma} \mathbf{e}_i = \lambda_i, i = 1, 2, \ldots, p \\ \text{Cov}(Y_i, Y_k) &= \mathbf{e}_i' \boldsymbol{\Sigma} \mathbf{e}_k = 0, i \neq k \end{aligned}$$

If some λ_i are equal, the choices of corresponding coefficients vectors, $\mathbf{e}_i$, and hence Y_i are not unique.

Results 5.3 If $Y_1 = \mathbf{e}_1'\boldsymbol{X}$, $Y_2 = \mathbf{e}_2'\boldsymbol{X}, \ldots, Y_p = \mathbf{e}_p'\boldsymbol{X}$ are the principal components obtained from the covariance matrix $\boldsymbol{\Sigma}$, then

$$\rho_{Y_i,X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, i, k = 1, 2, \ldots, p$$

are the correlation coefficients between the components Y_i and the variables X_k. Here $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ are the eigenvalue-eigenvector pair for Σ

Summarizing Sample Variation by Principle Components

Suppose the data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ represent n independent drawings from some p-dimensional population withe mean vector μ and covariance matrix Σ. These data yield the sample mean vector $\overline{\mathbf{x}}$, the sample covariance matrix $\mathbf{S}$, and the sample correlation matrix $\mathbf{R}$.
If $\mathbf{S} = s_{ik}$ be $p \times p$ sample covariance matrix with eigenvalue-eigenvector pairs $\left(\hat{\lambda}_1, \hat{\mathbf{e}}_1\right), \left(\hat{\lambda}_2, \hat{\mathbf{e}}_2\right), \ldots, \left(\hat{\lambda}_p, \hat{\mathbf{e}}_p\right)$, the i th sample principal component is given by $\hat{y}_i = \hat{\mathbf{e}}_i' \mathbf{x} = \hat{e}_{i1} x_1 + \hat{e}_{i2} x_2 + \cdots + \hat{e}_{ip} x_p, i = 1, 2, \ldots, p$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$ and $\mathbf{x}$ is any observation on the variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p$. Also

$$\text{Sample variance } (\hat{y}_k = \hat{\lambda}_k, k = 1, 2, \ldots, p$$

$$\text{Sample covariance } (\hat{y}_i, \hat{y}_k) = 0, i \neq k$$

$$\text{Total sample variance } = \sum_{i=1}^{n} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p$$

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ik} \sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}, i, k = 1, 2, \ldots, p$$

The Orthogonal Factor Model

$$
\begin{aligned}
X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1 \\
X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_1 \\
&\vdots \\
X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \varepsilon_p
\end{aligned}
$$

or in matrix notation

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

The coefficient ℓ_{ij} is called the loading of the ith variable on the j th factor, so the matrix L is the matrix of factor loadings.

The unobservable random vectors F and ε satisfy the following condition F and ε are independent $\mathrm{E}(\mathbf{F}) = 0, \mathrm{Cov}(\mathbf{F}) = \mathbf{I}$ $\mathrm{E}(\varepsilon) = 0, \mathrm{Cov}(\varepsilon) = \Psi$, where Ψ is diagonal matrix.

Covariance structure for the Orthogonal Factor Model

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

2. $\mathrm{Cov}(\boldsymbol{X}, \mathbf{F}) = \mathbf{L}$ or

$\mathrm{Cov}\left(X_i, F_j\right) = \ell_{ij}$

The Principal Component Solution of the Factor Model

The principal component analysis of the sample covariance matrix S is specified in terms of its eigenvalue-eigenvector pairs $\left(\hat{\lambda}_1, \hat{\mathbf{e}}_1\right), \left(\hat{\lambda}_2, \hat{\mathbf{e}}_2\right), \ldots, \left(\hat{\lambda}_p, \hat{\mathbf{e}}_p\right)$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. Let $m < p$ be the number of common factors. Then the matrix of estimate factor loading $\left\{\tilde{\ell}_{ij}\right\}$ is give by

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

The estimate specific variances are provided by the diagonal elements of the matrix $\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}'$, so

$$\boldsymbol{\Psi} = \begin{bmatrix} \tilde{\psi}_1 & 0 & \cdots & 0 \\ 0 & \tilde{\psi}_2 & \cdots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \cdots & \tilde{\psi}_p \end{bmatrix} \quad \text{with} \quad \tilde{\psi}_i = s_{ii} - \sum_{j=1}^{m} \tilde{\ell}_{ij}^2$$

Example 1

(a) Show that the covariance matrix

$$\boldsymbol{\rho} = \begin{bmatrix} 1.00 & 0.63 & 0.45 \\ 0.63 & 1.00 & 0.35 \\ 0.45 & 0.35 & 1.00 \end{bmatrix}$$

for $p = 3$ standardized random variables Z_1, Z_2 and Z_3 can be generated by the $m = 1$ factor model $\mathbf{Z} = \mathbf{LF} + \varepsilon$ with $\mathrm{Var}(\mathbf{F}) = 1\, \mathrm{Cov}(\varepsilon, \mathbf{F}) = 0$, and $\mathrm{Var}(\varepsilon) = \boldsymbol{\Psi}$

(b)The eigenvalues and eigenvectors of the correlation matrix ρ above are

$$\begin{aligned} \lambda_1 = 1.96, \quad \mathbf{e}_1' &= [0.625, 0.593, 0.507] \\ \lambda_2 = 0.68, \mathbf{e}_2' &= [-0.219, -0.491, 0.843] \\ \lambda_3 = 0.36, \mathbf{e}_3' &= [0.794, -0.638, -0.177] \end{aligned}$$

Assume an $m = 1$ factor model, and its loading matrix $\mathbf{L}$ and matrix of specific variance Ψ are calculated by the principal component solution method. Calculate $\mathrm{Corr}\,(Z_i, F_1)$ for $i = 1, 2, 3$.

Solutions

(a)

$$Z_1 = 0.9F + \varepsilon_1$$
$$Z_2 = 0.7F + \varepsilon_2$$
$$Z_3 = 0.5F + \varepsilon_3$$

Let

$$L = \begin{bmatrix} 0.9 \\ 0.7 \\ 0.5 \end{bmatrix}, \Psi = \begin{bmatrix} 0.19 & 0 & 0 \\ 0 & 0.51 & 0 \\ 0 & 0 & 0.75 \end{bmatrix}$$

Assume $\mathrm{Var}(F) = 1\,\mathrm{Cov}(\boldsymbol{\varepsilon}, \mathbf{F}) = 0, \mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$

$\mathrm{Cov}(\boldsymbol{Z}) = \boldsymbol{\rho} = \begin{bmatrix} 1.00 & 0.63 & 0.45 \\ 0.63 & 1.00 & 0.35 \\ 0.45 & 0.35 & 1.00 \end{bmatrix}$

$\mathrm{Cov}(\boldsymbol{L}F + \boldsymbol{\varepsilon}) = L\,\mathrm{Var}(F)L' + 2\,\mathrm{Cov}(\boldsymbol{\varepsilon}, \mathbf{F}) + \mathrm{Var}(\boldsymbol{\varepsilon})$

$=\mathrm{LL'}+\Psi = \begin{bmatrix} 1.00 & 0.63 & 0.45 \\ 0.63 & 1.00 & 0.35 \\ 0.45 & 0.35 & 1.00 \end{bmatrix} = \boldsymbol{\rho}$

$$\boldsymbol{L} = \sqrt{\lambda_1}\boldsymbol{e}_1 = \begin{bmatrix} 0.8750 \\ 0.8302 \\ 0.7098 \end{bmatrix}$$

The estimate specific variances are provided by the diagonal elements of the matrix $\boldsymbol{\rho} - \mathbf{LL}'$, so

$$\Psi = \begin{bmatrix} 0.234375 & 0 & 0 \\ 0 & 0.310768 & 0 \\ 0 & 0 & 0.496184 \end{bmatrix}$$

$F_1 = \frac{\boldsymbol{e}_1'\boldsymbol{Z}}{\sqrt{\lambda_1}}$, $Var(F_1) = \frac{1}{\lambda_1}\boldsymbol{e}_1' Cov(\boldsymbol{Z})\boldsymbol{e}_1 = 1$

$Cov(\boldsymbol{Z}, F_1) = \frac{1}{\sqrt{\lambda_1}} Cov(\boldsymbol{Z}, \boldsymbol{e_1'Z}) = \frac{1}{\sqrt{\lambda_1}}\boldsymbol{e_1'\rho} = \sqrt{\lambda_1}\boldsymbol{e}_1 = L$

Then $corr(Z_1, F_1) = \frac{cov(Z_1, F_1)}{\sqrt{Var(Z_1) Var F_1}} = L_1 = 0.8750$

$corr(Z_2, F_1) = \frac{cov(Z_2, F_1)}{\sqrt{Var(Z_2) Var F_1}} = L_2 = 0.8302$

$corr(Z_3, F_1) = \frac{cov(Z_3, F_1)}{\sqrt{Var(Z_3) Var F_1}} = L_2 = 0.7098$

Example2

Suppose the random vector $(X_1, X_2, X_3, X_4, X_5)'$ have the covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 0.75 & 0.75 & 0 & 0 \\ 0.75 & 1 & 0.75 & 0 & 0 \\ 0.75 & 0.75 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -0.75 \\ 0 & 0 & 0 & -0.75 & 1 \end{bmatrix}$$

(a) Find the first two principal components Y_1 and Y_2. (b) Find the correlation coefficients ρ_{Y_1,X_1} and ρ_{Y_2,X_4}. (c) Find the sum of $\mathrm{Var}\,(Y_1) + \mathrm{Var}\,(Y_2) + \mathrm{Var}\,(Y_3)$ (d) Determine an appropriate number of principle components.

Solutions

(a)Let $\Sigma_1 = \begin{bmatrix} 1 & 0.75 & 0.75 \\ 0.75 & 1 & 0.75 \\ 0.75 & 0.75 & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & -0.75 \\ -0.75 & 1 \end{bmatrix}$ Then
$\Sigma = \begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \Sigma_2 \end{bmatrix}$
$det(\Sigma - \lambda \boldsymbol{I}_5) = det(\Sigma_1 - \lambda \boldsymbol{I}_3) det(\Sigma_2 - \lambda \boldsymbol{I}_2) = (2.5 - \lambda)(1.75 - \lambda)(0.25 - \lambda)^3$
Then eigenvalues and eigenvectors are

$$\lambda_1 = 2.50, \lambda_2 = 1.75, \lambda_3 = \lambda_4 = \lambda_5 = 0.25$$

$$\boldsymbol{e}_1' = [-0.5773503, -0.5773503, -0.5773503, 0, 0]$$

$$\boldsymbol{e}_2' = [0, 0, 0, -0.7071068, 0.7071068]$$

$$\boldsymbol{e}_3' = [0.3175691, -0.8102156, 0.4926465, 0, 0]$$

$$\boldsymbol{e}_4' = [0.7522078, -0.1010810, -0.6511268, 0, 0]$$

$$\boldsymbol{e}_5' = [0, 0, 0 - 0.7071068, -0.7071068]$$

The first two principal components

$$Y_1 = \boldsymbol{e}_1' X$$

$$Y_2 = \boldsymbol{e}_2' X$$

(b)

$$\rho_{Y_1,X_1} = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{-0.5773503\sqrt{2.5}}{\sqrt{1}} = -0.912871$$

$$\rho_{Y_2,X_4} = \frac{e_{24}\sqrt{\lambda_2}}{\sqrt{\sigma_{44}}} = \frac{-0.7071068\sqrt{1.75}}{\sqrt{1}} = -0.9354144$$

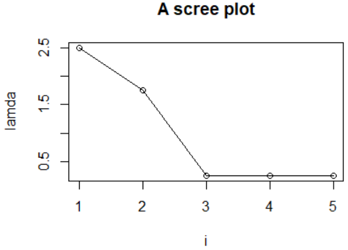
(c)

$$\text{Var}(Y_1) + \text{Var}(Y_2) + \text{Var}(Y_3) = \lambda_1 + \lambda_2 + \lambda_3 = 4.5$$

(d) $\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} = 0.5$

$\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} = 0.85$

$\frac{\lambda_1+\lambda_2+\lambda_3}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} = 0.9$



Example3

(a) Show that covariance matrix

$$\boldsymbol{\rho} = \begin{bmatrix} 1.00 & 0.63 & 0.45 & 0.27 \\ 0.63 & 1.00 & 0.35 & 0.21 \\ 0.45 & 0.35 & 1.00 & 0.15 \\ 0.27 & 0.21 & 0.15 & 1.00 \end{bmatrix}$$

for $p = 4$ standardized random variables Z_1, Z_2, Z_3 and Z_4 can be generated by the $m = 1$ factor model $\mathbf{Z} = \mathbf{LF} + \varepsilon$ with $\text{Var}(\mathbf{F}) = 1\, \text{Cov}(\varepsilon, \mathbf{F}) = 0$, and $\text{Var}(\epsilon) = \boldsymbol{\Psi}$

(b) The eigenvalues and eigenvectors of the correlation ρ above are

$$\begin{aligned} \lambda_1 &= 2.0888, \mathbf{e}_1' = [0.5993, 0.5606, 0.4722, 0.3218] \\ \lambda_2 &= 0.8832, \mathbf{e}_2' = [-0.1080, -0.1560, -0.3122, 0.9309] \\ \lambda_3 &= 0.6739, \mathbf{e}_3' = [0.2282, 0.5241, -0.8055, -0.1558] \\ \lambda_4 &= 0.3541, \mathbf{e}_4' = [-0.7597, 0.6219, 0.1749, 0.0747] \end{aligned}$$

Assume that an $m = 1$ factor model holds, calculate the loading matrix L and the matrix of the specific variance Ψ using the principal component solution method. What proportion of the total population variance is explained by the first common factor?

Solutions

(a)

$$Z_1 = 0.9F + \varepsilon_1$$
$$Z_2 = 0.7F + \varepsilon_2$$
$$Z_3 = 0.5F + \varepsilon_3$$
$$Z_4 = 0.3F + \varepsilon_4$$

Let

$$L = \begin{bmatrix} 0.9 \\ 0.7 \\ 0.5 \\ 0.3 \end{bmatrix}, \Psi = \begin{bmatrix} 0.19 & 0 & 0 & 0 \\ 0 & 0.51 & 0 & 0 \\ 0 & 0 & 0.75 & 0 \\ 0 & 0 & 0 & 0.91 \end{bmatrix}$$

Assume $\mathrm{Var}(F) = 1\,\mathrm{Cov}(\varepsilon, \mathbf{F}) = 0, \mathrm{Var}(\varepsilon) = \boldsymbol{\Psi}$

$$\mathrm{Cov}(\boldsymbol{Z}) = \boldsymbol{\rho} = \begin{bmatrix} 1.00 & 0.63 & 0.45 & 0.27 \\ 0.63 & 1.00 & 0.35 & 0.21 \\ 0.45 & 0.35 & 1.00 & 0.15 \\ 0.27 & 0.21 & 0.15 & 1.00 \end{bmatrix}$$

$$\mathrm{Cov}(\boldsymbol{L}F + \varepsilon) = L\,\mathrm{Var}(F)L' + 2\,\mathrm{Cov}(\varepsilon, \mathbf{F}) + \mathrm{Var}(\varepsilon)$$

$$=\text{LL'}+\Psi = \begin{bmatrix} 1.00 & 0.63 & 0.45 & 0.27 \\ 0.63 & 1.00 & 0.35 & 0.21 \\ 0.45 & 0.35 & 1.00 & 0.15 \\ 0.27 & 0.21 & 0.15 & 1.00 \end{bmatrix} = \boldsymbol{\rho}$$

Solutions

$$\text{(b)}L = \sqrt{\lambda_1} e_1 = \begin{bmatrix} 0.8661492 \\ 0.8102173 \\ 0.6824556 \\ 0.4650873 \end{bmatrix}$$

The estimate specific variances are provided by the diagonal elements of the matrix $\rho - \mathrm{LL}'$, so

$$\Psi = \begin{bmatrix} 0.2497856 & 0 & 0 & 0 \\ 0 & 0.3435479 & 0 & 0 \\ 0 & 0 & 0.5342543 & 0 \\ 0 & 0 & 0 & 0.7836938 \end{bmatrix}$$

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = 0.5222$$