

Aprendizagem de Dados Simbólicos e/ou Numéricos

Francisco de A.T. de Carvalho

Análise de Componentes Principais

✉ Introdução

✉ Problema

- ✉ Ponto de partida: observações descritas por p variáveis
- ✉ Meta: encontrar k ($k \leq p$) novas variáveis (*componentes principais*)
 - ✉ cada uma delas combinação linear das p variáveis iniciais
- ✉ preservando o melhor possível a variabilidade

Análise de Componentes Principais

✉ Componentes principais populacional

Vetor aleatório $\mathbf{X}_{1 \times p}^T = [X_1, \dots, X_p]$

Matriz de covariâncias $\mathbf{S}_{p \times p}$

Auto-valores $l_1 \geq \dots \geq l_p \geq 0$

Auto-vetores $\mathbf{e}_1, \dots, \mathbf{e}_p$

Análise de Componentes Principais

✉ Componentes principais populacional
Considere as combinações lineares:

$$Y_1 = \mathbf{a}_1^T \mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

...

$$Y_p = \mathbf{a}_p^T \mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p$$

Então:

$$Var(Y_i) = \mathbf{a}_i^T \mathbf{S} \mathbf{a}_i, i = 1, \dots, p$$

$$Cov(Y_i, Y_k) = \mathbf{a}_i^T \mathbf{S} \mathbf{a}_k, i, k = 1, \dots, p$$

Análise de Componentes Principais

✉ Componentes principais populacional

As componentes principais $Y_1, \dots, Y_p$ tem variância tão grande quanto possível:

✉ Primeira Componente Principal: combinação linear $\mathbf{a}_1^T \mathbf{X}$ que maximiza $\text{Var}(\mathbf{a}_1^T \mathbf{X})$ sujeito a $\mathbf{a}_1^T \mathbf{a}_1 = 1$

✉ ...

✉ i-ésima Componente Principal: combinação linear $\mathbf{a}_i^T \mathbf{X}$ que maximiza $\text{Var}(\mathbf{a}_i^T \mathbf{X})$ sujeito a $\mathbf{a}_i^T \mathbf{a}_i = 1$

Análise de Componentes Principais

✉ Resultado 1

Seja S a matriz de co-variâncias associada a $\mathbf{X}' = [X_1, \dots, X_p]$. Seja $(\lambda_1, \mathbf{e}_1), \dots, (\lambda_p, \mathbf{e}_p)$ com $\lambda_1 \geq \dots \geq \lambda_p$ os pares de autovalores e auto-vetores de S . A i -ésima componente principal é:

$$Y_i = \mathbf{e}_i^T \mathbf{X} = e_{i1}X_1 + \dots + e_{ip}X_p \quad i = 1, \dots, p$$

Temos também:

$$\text{Var}(Y_i) = \mathbf{e}_i^T \mathbf{S} \mathbf{e}_i = l_i \quad i = 1, \dots, p$$
$$\text{Cov}(Y_i, Y_k) = \mathbf{e}_i^T \mathbf{S} \mathbf{e}_k \quad i \neq k$$

Análise de Componentes Principais

✉ Resultado 2

Seja $\mathbf{X}' = [X_1, \dots, X_p]$ de matriz de co-variâncias S cujos pares de autovalores e auto-vetores são $(\lambda_1, \mathbf{e}_1), \dots, (\lambda_p, \mathbf{e}_p)$ onde $\lambda_1 \geq \dots \geq \lambda_p$. Seja $Y_1 = \mathbf{e}_1^T \mathbf{X}', \dots, Y_p = \mathbf{e}_p^T \mathbf{X}'$ as componentes principais. Então:

$$s_{11} + \dots + s_{pp} = \sum_{i=1}^p \text{Var}(X_i) = l_{11} + \dots + l_{pp} = \sum_{i=1}^p \text{Var}(Y_i)$$

Análise de Componentes Principais

✉ Importância de uma Componente Principal
É medida pelo aporte de sua variação à variação total

$$\left(\begin{array}{l} \text{Proporção da Variância} \\ \text{Total explicada por } Y_k \end{array} \right) = \frac{l_k}{l_1 + \dots + l_p} \quad k = 1, \dots, p$$

Análise de Componentes Principais

✉ Resultado 3

Se $Y_1 = \mathbf{e}_1^T \mathbf{X}', \dots, \mathbf{e}_i^T \mathbf{X}', \dots, Y_p = \mathbf{e}_p^T \mathbf{X}'$ com $\mathbf{e}_i^T = [e_{1i}, \dots, e_{ki}, \dots, e_{pi}]$ são as componentes principais obtidas a partir de S , então

$$r_{Y_i, X_k} = \frac{e_{ik} \sqrt{l_i}}{\sqrt{s_{kk}}} \quad i, k = 1, \dots, p$$

são os coeficientes de correlação entre as componentes principais Y_i e as variáveis X_k

Análise de Componentes Principais

✉ Variáveis padronizadas

Sejam as variáveis padronizadas:

$$Z_1 = \frac{(X_1 - m_1)}{\sqrt{s_{11}}}, \dots, Z_p = \frac{(X_p - m_p)}{\sqrt{s_{pp}}}$$

Isto é:

$$\underset{p \times 1}{\mathbf{Z}} = \left(\underset{p \times p}{\mathbf{V}^{1/2}} \right)^{-1} \left(\underset{p \times 1}{\mathbf{X}} - \underset{p \times 1}{\boldsymbol{\mu}} \right)$$

$$E(\mathbf{Z}) = \mathbf{0} \quad \text{e} \quad \text{Cov}(\mathbf{Z}) = \left(\mathbf{V}^{1/2} \right)^{-1} \mathbf{S} \left(\mathbf{V}^{1/2} \right)^{-1} = ?$$

Análise de Componentes Principais

✉ Resultado 4

A i -ésima componente principal das variáveis padronizadas $\mathbf{Z}'=[Z_1, \dots, Z_p]$ de matriz de covariâncias $\text{cov}(\mathbf{Z}) = \mathbf{r}$ é

$$Y_i = \mathbf{e}_i^T \mathbf{Z} = \mathbf{e}_i^T (\mathbf{V}^{1/2})^{-1} (\mathbf{X} - \boldsymbol{\mu}) \quad i = 1, \dots, p$$

Além disso:

$$1 + \dots + 1 = p = \sum_{i=1}^p \text{Var}(Z_i) = l_1 + \dots + l_p = \sum_{i=1}^p \text{Var}(Y_i)$$

$$r_{Y_i, Z_k} = e_{ki} \sqrt{l_i} \quad i, k = 1, \dots, p$$

$(\lambda_1, \mathbf{e}_1), \dots, (\lambda_p, \mathbf{e}_p)$ com $\lambda_1 \geq \dots \geq \lambda_p$ são os pares de autovalores e auto-vetores de $\mathbf{r}$

Análise de Componentes Principais



$$\left(\begin{array}{l} \text{Proporção da Variância} \\ \text{Total explicada por } Y_k \end{array} \right) = \frac{l_k}{1 + \dots + 1} = \frac{l_k}{p} \quad k = 1, \dots, p$$

λ_k são os autovalores de r

Análise de Componentes Principais

✉ Observações

- ✉ As componentes principais obtidas a partir de S não são as mesmas obtidas a partir de r
- ✉ A importância relativa de cada componente também será diferente
- ✉ O uso da matriz de correlações é indicado quando as variáveis
 - ✉ são medidas em unidades diferentes
 - ✉ são medidas em escalas com ampla diferença de amplitudes

Análise de Componentes Principais

✉ Componentes Principais Amostrais

Seja $\mathbf{x}_1, \dots, \mathbf{x}_n$ n observações independentes de $\mathbf{X}^T = [X_1, \dots, X_n]$ de vetor de médias populacional μ e matriz de co-variâncias populacional S
Seja

$\bar{\mathbf{x}}$: vetor de média amostral

$\mathbf{S}$: Matriz de co - variâncias amostral

$\mathbf{R}$: Matriz de correlações amostral

Análise de Componentes Principais

✉ Componentes Principais Amostrais

Os n valores da combinação linear

$$\mathbf{a}_1^T \mathbf{x} = a_{11}x_{1j} + \dots a_{p1}x_{pj} \quad j = 1, \dots, n$$

tem média $\mathbf{a}_1^T \bar{\mathbf{x}}$ e variância $\mathbf{a}_1^T \mathbf{S} \mathbf{a}_1$

Da mesma forma, todos os pares $(\mathbf{a}_1^T \mathbf{x}_j, \mathbf{a}_2^T \mathbf{x}_j)$ tem covariância amostral $\mathbf{a}_1^T \mathbf{S} \mathbf{a}_2$

Análise de Componentes Principais

✉ Componentes principais amostrais:

Primeira componente principal amostral

combinação linear $\hat{y}_1 = \mathbf{a}_1^T \mathbf{x}_j$ que maximiz a variância

amostral $\text{var}(\hat{y}_1) = \text{var}(\mathbf{a}_1^T \mathbf{x}_j) = \mathbf{a}_1^T \mathbf{S} \mathbf{a}_1$ sujeito a $\mathbf{a}_1^T \mathbf{a}_1 = 1$

...

i - ésima componente principal amostral

combinação linear $\hat{y}_i = \mathbf{a}_i^T \mathbf{x}_j$ que maximiz a variância amostral

$\text{var}(\hat{y}_i) = \text{var}(\mathbf{a}_i^T \mathbf{x}_j) = \mathbf{a}_i^T \mathbf{S} \mathbf{a}_i$ sujeito a $\mathbf{a}_i^T \mathbf{a}_i = 1$ e

$\text{cov}(\hat{y}_i, \hat{y}_k) = \text{cov}(\mathbf{a}_i^T \mathbf{x}_j, \mathbf{a}_k^T \mathbf{x}_j) = \mathbf{a}_i^T \mathbf{S} \mathbf{a}_k = 0 \quad k < i$

Análise de Componentes Principais

✉ Resultado 5

$\mathbf{S}=\{s_{ik}\}$ é a matriz ($p \times p$) de co-variâncias amostral cujos pares de autovalores e auto-vetores são

$$(\hat{l}_1, \hat{\mathbf{e}}_1), \dots, (\hat{l}_p, \hat{\mathbf{e}}_p) \quad \text{onde} \quad \hat{l}_1 \geq \dots \geq \hat{l}_p$$

A i -ésima componente principal amostral é

$$\hat{y}_i = \mathbf{e}_i^T \mathbf{x} = \hat{e}_{i1}x_1 + \dots + \hat{e}_{ip}x_p \quad i = 1, \dots, p$$

onde $\mathbf{x}^T = [x_1, \dots, x_p]$ é uma observação de $\mathbf{X}^T = [X_1, \dots, X_p]$

Análise de Componentes Principais

✉ Resultado 5

Além disso

$$\text{Variância amostral}(\hat{y}_k) = \hat{l}_k \quad k = 1, \dots, p$$

$$\text{Co - variância amostral}(\hat{y}_i, \hat{y}_k) = 0 \quad i \neq k$$

$$\text{Variância total amostral} = \sum_{i=1}^p s_{ii} = \hat{l}_1 + \dots + \hat{l}_p$$

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ki} \sqrt{\hat{l}_i}}{\sqrt{s_{ii}}} \quad i, k = 1, \dots, p$$

Análise de Componentes Principais

✉ Resultado 5 Também

Proporção da variância total amostral explicada por $\hat{y}_k$

$$\frac{\hat{l}_k}{\hat{l}_1 + \dots + \hat{l}_p} \quad k = 1, \dots, p$$

Análise de Componentes Principais

✉ Componentes Principais Amostrais: variáveis padronizadas

Sejam os dados padronizados:

$$\mathbf{z}_j = \mathbf{D}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}}) = \left[\frac{x_{1j} - \bar{x}_1}{\sqrt{s_{11}}}, \dots, \frac{x_{pj} - \bar{x}_p}{\sqrt{s_{pp}}} \right]^T \quad j = 1, \dots, n$$

Análise de Componentes Principais

✉ Componentes Principais Amostrais: variáveis padronizadas

Matriz de dados de observações padronizadas:

$$\mathbf{Z}_{n \times p} = \begin{bmatrix} \mathbf{z}_1^T \\ \mathbf{z}_2^T \\ \vdots \\ \mathbf{z}_n^T \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}$$

Análise de Componentes Principais

✉ Componentes Principais Amostrais: variáveis padronizadas

Matriz de dados de observações padronizadas:

$$\mathbf{Z}_{n \times p} = \begin{bmatrix} \frac{x_{11} - \bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{12} - \bar{x}_1}{\sqrt{s_{11}}} & \dots & \frac{x_{1p} - \bar{x}_1}{\sqrt{s_{11}}} \\ \frac{x_{21} - \bar{x}_2}{\sqrt{s_{22}}} & \frac{x_{22} - \bar{x}_2}{\sqrt{s_{22}}} & \dots & \frac{x_{2p} - \bar{x}_2}{\sqrt{s_{22}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1} - \bar{x}_p}{\sqrt{s_{pp}}} & \frac{x_{n2} - \bar{x}_p}{\sqrt{s_{pp}}} & \dots & \frac{x_{np} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix}$$

Análise de Componentes Principais

✉ Componentes Principais Amostrais: variáveis padronizadas

Vetor de médias amostral:

$$\bar{\mathbf{z}} = \frac{1}{n} (\mathbf{1}^T \mathbf{Z})^T = \frac{1}{n} \mathbf{Z}^T \mathbf{1} = \frac{1}{n} \begin{bmatrix} \sum_{j=1}^n \frac{x_{j1} - \bar{x}_1}{\sqrt{s_{11}}} \\ \sum_{j=1}^n \frac{x_{j2} - \bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \sum_{j=1}^n \frac{x_{jp} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} = \mathbf{0}$$

Análise de Componentes Principais

✉ Componentes Principais Amostrais: variáveis padronizadas

Matriz de co-variâncias amostral:

$$\begin{aligned} \mathbf{S} &= \frac{1}{n-1} \left(\mathbf{Z} - \frac{1}{n} \mathbf{1} \mathbf{1}^T \mathbf{Z} \right)^T \left(\mathbf{Z} - \frac{1}{n} \mathbf{1} \mathbf{1}^T \mathbf{Z} \right) \\ &= \frac{1}{n-1} \left(\mathbf{Z} - \mathbf{1} \bar{\mathbf{z}}^T \right)^T \left(\mathbf{Z} - \mathbf{1} \bar{\mathbf{z}}^T \right) \\ &= \frac{1}{n-1} \mathbf{Z}^T \mathbf{Z} \end{aligned}$$

Análise de Componentes Principais

✉ Componentes Principais Amostrais: variáveis padronizadas

Matriz de co-variâncias amostral:

$$\mathbf{S} = \begin{bmatrix} \frac{(n-1)s_{11}}{s_{11}} & \frac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \dots & \frac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} \\ \frac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \frac{(n-1)s_{22}}{s_{22}} & \dots & \frac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{2p}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} & \frac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} & \dots & \frac{(n-1)s_{pp}}{s_{pp}} \end{bmatrix}$$

Análise de Componentes Principais

✉ Resultado 6:

Se $\mathbf{z}_1, \dots, \mathbf{z}_n$ são observações padronizadas com matriz de co-variâncias $\mathbf{R}$ a i -ésima componente principal é

$$\hat{y}_i = \hat{\mathbf{e}}_i^T \mathbf{z} = \hat{e}_{i1}z_1 + \dots + \hat{e}_{ip}z_p \quad i = 1, \dots, p$$

onde $(\hat{l}_i, \hat{\mathbf{e}}_i)$ é o i -ésimo par de autovalores e auto-vetores de $\mathbf{R}$ com

$$\hat{l}_1 \geq \dots \geq \hat{l}_p \geq 0$$

Análise de Componentes Principais

✉ Resultado 6:

Também:

Variância amostral($\hat{y}_i$) = $\hat{l}_i$ $i = 1, \dots, p$

Co - variância amostral($\hat{y}_i, \hat{y}_k$) = 0 $i \neq k$

Variância total amostral padronizada = $tr(\mathbf{R}) = p = \hat{l}_1 + \dots + \hat{l}_p$

$r_{\hat{y}_i, z_k} = \hat{e}_{ki} \sqrt{\hat{l}_i}$ $i, k = 1, \dots, p$

Proporção da variância total amostral explicada por $\hat{y}_k$

$$\frac{\hat{l}_k}{p} \quad k = 1, \dots, p$$