Análise de componentes principais

Patrícia de Siqueira Ramos

UNIFAL-MG, *campus* Varginha

5 de Maio de 2016

Introdução

- Técnica exploratória
- Objetivo: substituir p variáveis métricas correlacionadas por um número menor de k variáveis não correlacionadas que contenham a maior parte das informações contidas nas variáveis originais
- Há redução na dimensionalidade dos dados (é mais simples interpretar duas ou três variáveis não correlacionadas do que 20 ou 30 que apresentam interrelações complicadas)

ACP

ACP transforma o conjunto de variáveis correlacionadas

$$\mathbf{X}^T = \begin{bmatrix} X_1 & X_2 & \dots & X_p \end{bmatrix}$$

em um conjunto de variáveis não correlacionadas

$$\mathbf{Y}^T = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_p \end{bmatrix},$$

em que cada Y_j (componente principal) é uma combinação linear das variáveis em $\mathbf{X}$.

ACP

Os componentes principais $Y_1, \ldots, Y_p$ são, então, as combinações lineares:

$$\begin{aligned}
Y_1 =& a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 =& a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\vdots \\
Y_p =& a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p.
\end{aligned}$$

Cada componente é uma soma ponderada dos Xs, em que os a_{ji}s são os pesos ou coeficientes.

ACP

Os CPs são obtidos em ordem decrescente de importância:

Y_1 explica o máximo possível da variância total,

Y_2 explica o máximo da variância restante,

Y_3 etc.

ACP

Os CPs são obtidos em ordem decrescente de importância:

Y_1 explica o máximo possível da variância total,

Y_2 explica o máximo da variância restante,

Y_3 etc.

O conjunto $Y_1, Y_2, \ldots, Y_p$ explica a variância total presente em $\mathbf{X}$:

$$\sum_{j=1}^{p} V(Y_j) = \sum_{i=1}^{p} V(X_i).$$

ACP

- Espera-se que poucos dos primeiros componentes já contabilizem boa parte da variação em $\mathbf{X}$ e os restantes possam ser descartados sem grande perda de informação

ACP

- Espera-se que poucos dos primeiros componentes já contabilizem boa parte da variação em $\mathbf{X}$ e os restantes possam ser descartados sem grande perda de informação
- ACP é análoga à análise de correspondência (AC), ambas servem para reduzir a dimensionalidade dos dados, mas ACP é para variáveis contínuas e AC é para categóricas

ACP

- Espera-se que poucos dos primeiros componentes já contabilizem boa parte da variação em $\mathbf{X}$ e os restantes possam ser descartados sem grande perda de informação
- ACP é análoga à análise de correspondência (AC), ambas servem para reduzir a dimensionalidade dos dados, mas ACP é para variáveis contínuas e AC é para categóricas
- ACP é muito utilizada como uma etapa na análise de regressão linear quando:
  (i) há muitas variáveis independentes em relação ao número de observações
  (ii) há multicolinearidade (correlação entre as variáveis independentes)

Obtenção dos CPs amostrais

O primeiro CP das observações é a combinação linear

$$Y_1 = \mathbf{a_1}^T \mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p,$$

cuja variância amostral é $\mathbf{a_1}^T \mathbf{S}\mathbf{a_1}$ e é a maior dentre todas as outras combinações lineares.

Obtenção dos CPs amostrais

O primeiro CP das observações é a combinação linear

$$Y_1 = \mathbf{a_1}^T\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p,$$

cuja variância amostral é $\mathbf{a_1}^T\mathbf{S}\mathbf{a_1}$ e é a maior dentre todas as outras combinações lineares.

- Restrição: $\mathbf{a_1}^T\mathbf{a_1} = 1$ (SQ deve ser igual a 1, $\sum_{i=1}^{p} a_{1i}^2 = 1$)
- Para obter os coeficientes de Y_1, ou seja, $\mathbf{a_1}$, devemos escolher os valores de $\mathbf{a_1}$ de forma a maximizar a variância de Y_1 sujeito à restrição

Obtenção dos CPs amostrais

O primeiro CP das observações é a combinação linear

$$Y_1 = \mathbf{a_1}^T\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p,$$

cuja variância amostral é $\mathbf{a_1}^T\mathbf{S}\mathbf{a_1}$ e é a maior dentre todas as outras combinações lineares.

- Restrição: $\mathbf{a_1}^T\mathbf{a_1} = 1$ (SQ deve ser igual a 1, $\sum_{i=1}^{p} a_{1i}^2 = 1$)
- Para obter os coeficientes de Y_1, ou seja, $\mathbf{a_1}$, devemos escolher os valores de $\mathbf{a_1}$ de forma a maximizar a variância de Y_1 sujeito à restrição
- Solução: $\mathbf{a_1}$ é o autovetor de $\mathbf{S}$ associado ao seu maior autovalor

Obs.: Lembrando que os autovalores e autovetores de uma matriz quadrada $\mathbf{A}$ são tais que $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$

Obtenção dos CPs amostrais

O segundo CP das observações é a combinação linear

$$Y_2 = \mathbf{a_2}^T \mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p,$$

Obtenção dos CPs amostrais

O segundo CP das observações é a combinação linear

$$Y_2 = \mathbf{a_2}^T \mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p,$$

que tem a maior variância sujeito a

$$\mathbf{a_2}^T \mathbf{a_2} = 1 \qquad \text{e} \qquad \mathbf{a_2}^T \mathbf{a_1} = 0$$

(a segunda garante que Y_1 e Y_2 são ortogonais, ou seja, são não correlacionados)

Obtenção dos CPs amostrais

O j-ésimo CP das observações ($j = 1, \ldots, p$) é

$$Y_j = \mathbf{a_j}^T \mathbf{X} = a_{j1}X_1 + a_{j2}X_2 + \cdots + a_{jp}X_p,$$

Obtenção dos CPs amostrais

O j-ésimo CP das observações ($j = 1, \ldots, p$) é

$$Y_j = \mathbf{a_j}^T \mathbf{X} = a_{j1}X_1 + a_{j2}X_2 + \cdots + a_{jp}X_p,$$

que tem a maior variância sujeito a

$$\mathbf{a_j}^T \mathbf{a_j} = 1 \qquad \text{e} \qquad \mathbf{a_j}^T \mathbf{a_i} = 0 \quad (j > i)$$

Obtenção dos CPs amostrais

O j-ésimo CP das observações ($j = 1, \ldots, p$) é

$$Y_j = \mathbf{a_j}^T \mathbf{X} = a_{j1}X_1 + a_{j2}X_2 + \cdots + a_{jp}X_p,$$

que tem a maior variância sujeito a

$$\mathbf{a_j}^T \mathbf{a_j} = 1 \qquad \text{e} \qquad \mathbf{a_j}^T \mathbf{a_i} = 0 \quad (j > i)$$

O vetor $\mathbf{a_j}$ de coeficientes é o autovetor de $\mathbf{S}$ associado ao seu j-ésimo maior autovalor

Obtenção dos CPs amostrais

Se os p autovalores de $\mathbf{S}$ são $\lambda_1, \lambda_2, \ldots, \lambda_p$, a variância do j-ésimo CP é dada por λ_j.

Então, a variância total dos p CPs será igual à variância total das variáveis originais:

$$\sum_{j=1}^{p} \lambda_j = S_1^2 + S_2^2 + \cdots + S_p^2,$$

em que S_i^2 é a variância amostral de X_i, ou

$$\sum_{j=1}^{p} \lambda_j = \mathrm{traço}(\mathbf{S}),$$

Variâncias

- A proporção da variância total de **X** explicada pelo j-ésimo CP é

$$\frac{V(Y_j)}{\text{V.total de } \mathbf{X}} = \frac{\lambda_j}{tr(\mathbf{S})} = \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

Variâncias

- A proporção da variância total de $\mathbf{X}$ explicada pelo j-ésimo CP é

$$\frac{V(Y_j)}{\text{V.total de } \mathbf{X}} = \frac{\lambda_j}{tr(\mathbf{S})} = \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

- Variância generalizada de $\mathbf{X}$ e $\mathbf{Y} = |\mathbf{S}| = \prod_{j=1}^{p} \lambda_j$

Variâncias

- A proporção da variância total de **X** explicada pelo j-ésimo CP é

$$\frac{V(Y_j)}{\text{V.total de } \mathbf{X}} = \frac{\lambda_j}{tr(\mathbf{S})} = \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

- Variância generalizada de **X** e **Y** $= |\mathbf{S}| = \prod_{j=1}^{p} \lambda_j$
- Assim, **X** e **Y** são equivalentes em relação às variâncias.

Variâncias

- Os primeiros k componentes ($k < p$) explicam uma proporção da variação total:

$$\frac{\sum_{j=1}^{k} V(Y_j)}{\text{V.total de } \mathbf{X}} = \frac{\sum_{j=1}^{k} V(Y_j)}{tr(\mathbf{S})} = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

Variâncias

- Os primeiros k componentes ($k < p$) explicam uma proporção da variação total:

$$\frac{\sum_{j=1}^{k} V(Y_j)}{\text{V.total de } \mathbf{X}} = \frac{\sum_{j=1}^{k} V(Y_j)}{tr(\mathbf{S})} = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

- A correlação entre Y_j e X_i é dada por:

$$r_{Y_j, X_i} = \frac{a_{ji}\sqrt{\lambda_j}}{\sqrt{S_{ii}}}.$$

Variâncias

- Os primeiros k componentes ($k < p$) explicam uma proporção da variação total:

$$\frac{\sum_{j=1}^{k} V(Y_j)}{\text{V.total de } \mathbf{X}} = \frac{\sum_{j=1}^{k} V(Y_j)}{tr(\mathbf{S})} = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

- A correlação entre Y_j e X_i é dada por:

$$r_{Y_j, X_i} = \frac{a_{ji}\sqrt{\lambda_j}}{\sqrt{S_{ii}}}.$$

Obs.: As correlações são úteis para interpretação dos CPs.

Ilustração da ACP para $p = 2$

- Imagine que X_1 e X_2 estão na mesma escala, são altamente correlacionadas positivamente e ainda:

$$V(X_1) = V(X_2) = 1 \qquad \text{e} \qquad r_{12} = 0,90$$

Ilustração da ACP para $p = 2$

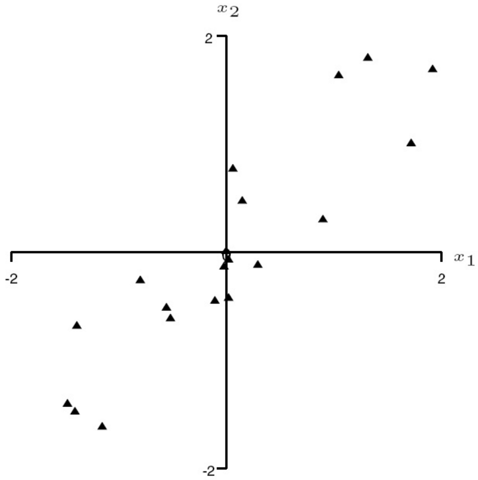

Ilustração da ACP para $p = 2$

- Obter os CPs envolve uma rotação ortogonal dos eixos
- O 1^o CP (Y_1) estará na direção de maior variância (minimizando-se as distâncias quadradas das observações ao 1^o CP). O 2^o CP é fixo pois deve ser ortogonal ao 1^o
- Se $V(X_1) \neq V(X_2)$, o 1^o CP ficará próximo ao eixo de maior variância

Ilustração da ACP para $p = 2$

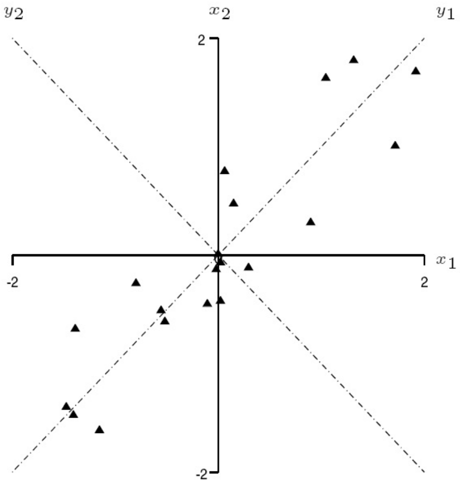

Ilustração da ACP para $p = 2$

- Plotam-se as observações em relação aos novos eixos Y_1 e Y_2:

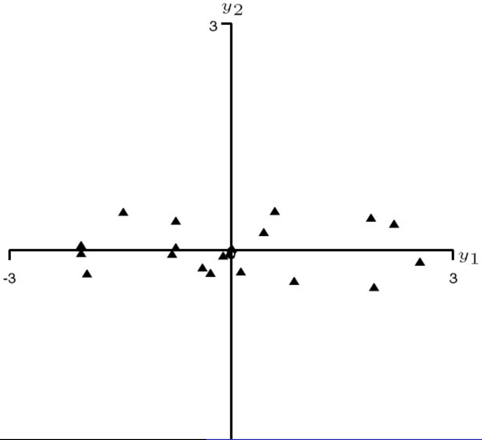

Ilustração da ACP para $p = 2$

- Quanto maior a correlação entre X_1 e X_2, mais próximas as observações estarão do 1^o CP (Y_1) e maior a variação explicada por ele
- No exemplo, se $V(X_1) >> V(X_2)$, a inclinação seria bem menor, o que faria Y_1 bem mais parecido com X_1
- Em geral, quanto maior a variância de uma variável, mais dominante ela é

Ilustração da ACP para $p = 2$

- Em situações reais, escalas são arbitrárias, então é melhor padronizar as variáveis para que tenham variância igual a 1, o que corresponde a utilizar a matriz de correlações $\mathbf{R}$ no lugar de $\mathbf{S}$ (veremos depois)

Ilustração da ACP para $p = 2$

- Em situações reais, escalas são arbitrárias, então é melhor padronizar as variáveis para que tenham variância igual a 1, o que corresponde a utilizar a matriz de correlações $\mathbf{R}$ no lugar de $\mathbf{S}$ (veremos depois)
- Do exemplo,

$$Y_1 = X_1/\sqrt{2} + X_2/\sqrt{2} \qquad Y_2 = X_2/\sqrt{2} - X_1/\sqrt{2}$$

$$V(Y_1) = 1,90 \qquad V(Y_2) = 0,10$$

$$\text{V. total de } \mathbf{X} = 1 + 1 = 2.$$

Ilustração da ACP para $p = 2$

- Em situações reais, escalas são arbitrárias, então é melhor padronizar as variáveis para que tenham variância igual a 1, o que corresponde a utilizar a matriz de correlações $\mathbf{R}$ no lugar de $\mathbf{S}$ (veremos depois)
- Do exemplo,

$$Y_1 = X_1/\sqrt{2} + X_2/\sqrt{2} \qquad Y_2 = X_2/\sqrt{2} - X_1/\sqrt{2}$$

$$V(Y_1) = 1,90 \qquad V(Y_2) = 0,10$$

$$\text{V. total de } \mathbf{X} = 1 + 1 = 2.$$

- A variância total $(= 2)$ fica distribuída de forma desigual, com a maior parte alocada no 1^o CP, Y_1.