

Análise de componentes principais

Patrícia de Siqueira Ramos

UNIFAL-MG, *campus* Varginha

4 de Junho de 2018

Métodos de aprendizado

- Supervisionado:
 - Há p variáveis medidas em n observações e uma resposta \mathbf{Y} também medida nas n observações
 - Objetivo: prever \mathbf{Y} usando $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p$
 - Exemplos: regressão, redes neurais

Métodos de aprendizado

- Supervisionado:
 - Há p variáveis medidas em n observações e uma resposta \mathbf{Y} também medida nas n observações
 - Objetivo: prever \mathbf{Y} usando $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p$
 - Exemplos: regressão, redes neurais
- Não supervisionado:
 - Há apenas as observações $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p$ e não há variável resposta \mathbf{Y}
 - Objetivo: descobrir coisas interessantes sobre \mathbf{X}
 - Não há como avaliar os resultados, pois não há uma resposta correta
 - Exemplos: análise de componentes principais, análise de agrupamento

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

$$\mathbf{X}^T = [X_1 \quad X_2 \quad \dots \quad X_p]$$

em um conjunto de variáveis não correlacionadas

$$\mathbf{Y}^T = [Y_1 \quad Y_2 \quad \dots \quad Y_p],$$

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

ACP

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

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p$$

$$Y_2 = a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p$$

$$\vdots$$

$$Y_p = a_{p1}X_1 + a_{p2}X_2 + \dots + a_{pp}X_p.$$

Cada componente é uma soma ponderada dos X s, 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, \dots, 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 **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 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

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

cujas 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

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

cujas 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

$$\mathbf{a}_2^T \mathbf{a}_2 = 1 \quad \text{e} \quad \mathbf{a}_2^T \mathbf{a}_1 = 0$$

(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 ($j = 1, \dots, p$) é

$$Y_j = \mathbf{a}_j^T \mathbf{X} = a_{j1}X_1 + a_{j2}X_2 + \dots + a_{jp}X_p,$$

Obtenção dos CPs amostrais

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

$$Y_j = \mathbf{a}_j^T \mathbf{X} = a_{j1}X_1 + a_{j2}X_2 + \dots + a_{jp}X_p,$$

que tem a maior variância sujeito a

$$\mathbf{a}_j^T \mathbf{a}_j = 1 \quad \text{e} \quad \mathbf{a}_j^T \mathbf{a}_i = 0 \quad (j > i)$$

Obtenção dos CPs amostrais

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

$$Y_j = \mathbf{a}_j^T \mathbf{X} = a_{j1}X_1 + a_{j2}X_2 + \dots + a_{jp}X_p,$$

que tem a maior variância sujeito a

$$\mathbf{a}_j^T \mathbf{a}_j = 1 \quad \text{e} \quad \mathbf{a}_j^T \mathbf{a}_i = 0 \quad (j > i)$$

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 $\lambda_1, \lambda_2, \dots, \lambda_p$, 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:

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

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

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

Variâncias

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

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

Variâncias

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

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

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

Variâncias

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

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

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

Variâncias

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

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

Variâncias

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

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

- 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:

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

- 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:

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

Ilustração da ACP para $p = 2$

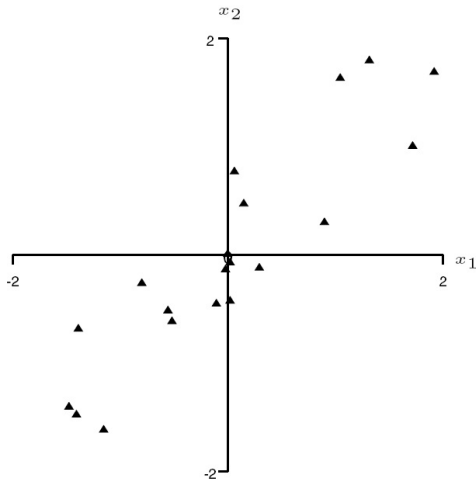


Ilustração da ACP para $p = 2$

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

Ilustração da ACP para $p = 2$

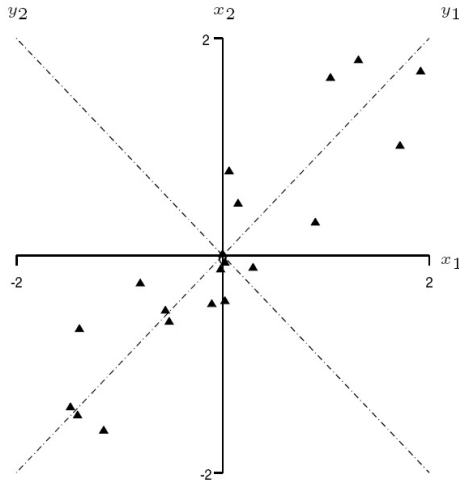


Ilustração da ACP para $p = 2$

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

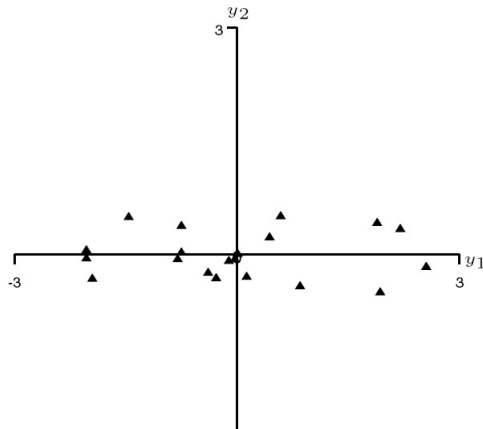


Ilustração da ACP para $p = 2$

- Quanto maior a correlação entre X_1 e X_2 , mais próximas as observações estarão do 1º CP (Y_1) e maior a variação explicada por ele
- No exemplo, se $V(X_1) \gg V(X_2)$, 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 **R** no lugar de **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 **R** no lugar de **S** (veremos depois)
- Do exemplo,

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

$$V(Y_1) = \lambda_1 = 1,90$$

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

$$V. \text{ total de } \mathbf{Y} = 1,90 + 0,10 = 2.$$

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

$$V(Y_2) = \lambda_2 = 0,10$$

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 **R** no lugar de **S** (veremos depois)
- Do exemplo,

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

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

$$V(Y_1) = \lambda_1 = 1,90$$

$$V(Y_2) = \lambda_2 = 0,10$$

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

$$V. \text{ total de } \mathbf{Y} = 1,90 + 0,10 = 2.$$

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

Ilustração da ACP para $p = 2$

- Proporção da variação explicada pelo primeiro componente:

$$\frac{V(Y_1)}{\text{Variância total}(\mathbf{X})} = \frac{\lambda_1}{\text{traço}(\mathbf{S})} = \frac{1,90}{2} = 0,95$$

- ou seja, o CP1 (Y_1) explica 95% da variação total

Ilustração da ACP para $p = 2$

- Proporção da variação explicada pelo segundo componente:

$$\frac{V(Y_2)}{\text{Variância total}(\mathbf{X})} = \frac{\lambda_2}{\text{traço}(\mathbf{S})} = \frac{0,10}{2} = 0,05$$

- ou seja, o CP2 (Y_2) explica 5% da variação total

Ilustração da ACP para $p = 2$

- Dessa forma, a partir da análise usando a matriz de correlações **S** (que não é a ideal, é melhor usar **R**), com o primeiro componente já explicaríamos 95% da variação total e não precisaríamos do CP2.