Data Science and Big Data

Taconeli, C.A.

23 de novembro, 2018

Análise de componentes principais

Análise de componentes principais

- A análise de componentes principais (ACP) é uma técnica multivariada, aplicada a um conjunto de p variáveis aleatórias, que tem como principais objetivos:
  - Reduzir a dimensão dos dados, projetando-os em uma dimensão $k < p$;
  - Explorar a estrutura de variância das variáveis;
  - Determinar índices e produzir escores com base nos resultados avaliados para as p variáveis.
- A ACP consiste na obtenção de novas variáveis, determinadas a partir das variáveis originais, de tal forma que um pequeno número de novas variáveis (componentes principais) seja capaz de explicar a maior parte possível da variação presente nos dados.

Análise de componentes principais

- As novas variáveis obtidas (denominadas componentes) são tais que o primeiro (principal) componente é aquele capaz de explicar a maior parte possível da variação dos dados, o segundo é aquele que explica a maior parte possível não explicada pelo primeiro...

- Algebricamente, os componentes $(Z_1, Z_2, ..., Z_p)$ correspondem a combinações lineares das variáveis originais $Y_1, Y_2, ..., Y_p$.

- Geometricamente, as combinações lineares representam um novo sistema de coordenadas, obtido por meio da rotação do sistema original.

Análise de componentes principais

- Motivação de uma ACP, em que y_1 e y_2 são as variáveis originais e z_1 e z_2 os componentes

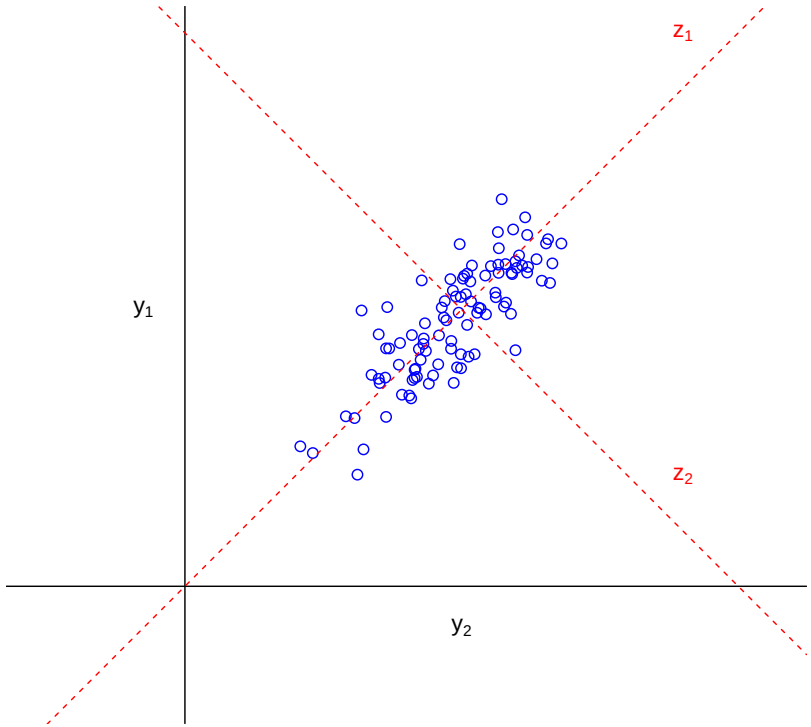

Figura 1: Ilustração - componentes principais.

Análise de componentes principais

- Em algumas aplicações, os componentes da ACP configuram o objetivo final do estudo.

- Em outras situações, os componentes obtidos servem como passo intermediário para realização de outras análises, como regressão, classificação, agrupamento...

Análise de componentes principais

- Considere o vetor aleatório $\boldsymbol{Y}' = (Y_1, Y_2, ..., Y_p)$ com matriz de covariâncias $\boldsymbol{\Sigma}$ de autovalores $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$.

- Considere as combinações lineares:

$$Z_1 = \boldsymbol{a_1'Y} = a_{11}Y_1 + a_{12}Y_2 + ... + a_{1p}Y_p$$
$$Z_2 = \boldsymbol{a_2'Y} = a_{21}Y_1 + a_{22}Y_2 + ... + a_{2p}Y_p$$
$$\vdots$$
$$Z_p = \boldsymbol{a_p'Y} = a_{p1}Y_1 + a_{p2}Y_2 + ... + a_{pp}Y_p$$

Análise de componentes principais

- Sabemos que:

$$Var(Z_i) = \boldsymbol{a_i'\Sigma a_i}, \qquad i = 1, 2, ..., p$$

$$Cov(Z_i, Z_k) = \boldsymbol{a_i'\Sigma a_k}, \qquad i, k = 1, 2, ..., p$$

- Os componentes principais são as combinações lineares de $Y_1, Y_2, ..., Y_p$ **não correlacionadas** e que tenham **maior variância possível**.

- No entanto, $Var(Z_i) = \boldsymbol{a_i'\Sigma a_i}$ pode crescer indefinidamente multiplicando $\boldsymbol{a_i}$ por alguma constante. Assim, aplicamos a restrição $\boldsymbol{a_i'a_i} = 1$.

Análise de componentes principais

- Resumo da análise de componentes principais:

Passo 1 - Primeiro componente principal = combinação linear $\boldsymbol{a}_1'\boldsymbol{Y}$ que maximiza $Var(\boldsymbol{a}_1'\boldsymbol{Y})$ sujeita a $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$;

Passo 2 - Segundo componente principal = combinação linear $\boldsymbol{a}_2'\boldsymbol{Y}$ que maximiza $Var(\boldsymbol{a}_2'\boldsymbol{Y})$ sujeita a $\boldsymbol{a}_2'\boldsymbol{a}_2 = 1$ e $Cov(\boldsymbol{a}_1'\boldsymbol{Y}, \boldsymbol{a}_2'\boldsymbol{Y}) = 0$;

$\vdots$

Passo i - i-ésimo componente principal = combinação linear $\boldsymbol{a}_i'\boldsymbol{Y}$ que maximiza $Var(\boldsymbol{a}_p'\boldsymbol{Y})$ sujeita a $\boldsymbol{a}_i'\boldsymbol{a}_i = 1$ e $Cov(\boldsymbol{a}_i'\boldsymbol{Y}, \boldsymbol{a}_k'\boldsymbol{Y}) = 0$, para todo $k < i$.

Análise de componentes principais

- Para determinação dos componentes principais, com base no que foi exposto, usaremos o seguinte teorema:
- **Maximização de formas quadráticas** Seja $\boldsymbol{B}$ uma matriz positiva definida com autovalores $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ e autovetores associados normalizados $\boldsymbol{e_1}, \boldsymbol{e_2}, ..., \boldsymbol{e_p}$. Então:

$$\max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x'Bx}}{\boldsymbol{x'x}} = \lambda_1, \quad \text{obtido quando} \quad \boldsymbol{x} = \boldsymbol{e_1};$$

$$\min_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x'Bx}}{\boldsymbol{x'x}} = \lambda_p, \quad \text{obtido quando} \quad \boldsymbol{x} = \boldsymbol{e_p}.$$

Adicionalmente,

$$\max_{\boldsymbol{x} \perp \boldsymbol{e_1}, \boldsymbol{e_2}, ..., \boldsymbol{e_k}} \frac{\boldsymbol{x'Bx}}{\boldsymbol{x'x}} = \lambda_{k+1}, \quad \text{obtido quando} \quad \boldsymbol{x} = \boldsymbol{e_{k+1}}.$$

Análise de componentes principais

- Assim, no contexto de componentes principais, seja $\boldsymbol{Y'} = (Y_1, Y_2, ..., Y_p)$ um vetor aleatório. Seja $\boldsymbol{\Sigma}$ a matriz de variâncias e covariâncias e $(\lambda_1, \boldsymbol{e_1}), (\lambda_2, \boldsymbol{e_2}), ..., (\lambda_p, \boldsymbol{e_p})$ seus autovalores e autovetores, tal que $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Então:

$$\max_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{(\boldsymbol{a'\Sigma a})}{\boldsymbol{a'a}} = \max_{\boldsymbol{a} \neq \boldsymbol{0}} (\boldsymbol{a'\Sigma a}) = \lambda_1, \quad \text{obtido quando} \quad \boldsymbol{a = e_1};$$

$$\min_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{(\boldsymbol{a'\Sigma a})}{\boldsymbol{a'a}} = \min_{\boldsymbol{a} \neq \boldsymbol{0}} (\boldsymbol{a'\Sigma a}) = \lambda_p, \quad \text{obtido quando} \quad \boldsymbol{a = e_p}.$$

Adicionalmente,

$$\max_{\boldsymbol{a} \perp \boldsymbol{e_1}, \boldsymbol{e_2}, ..., \boldsymbol{e_k}} \frac{(\boldsymbol{a'\Sigma a})}{\boldsymbol{a'a}} = \max_{\boldsymbol{a} \perp \boldsymbol{e_1}, \boldsymbol{e_2}, ..., \boldsymbol{e_k}} (\boldsymbol{a'\Sigma a}) = \lambda_{k+1}, \quad \text{obtido quando} \quad \boldsymbol{a = e_{k+1}}$$

Análise de componentes principais

- O i-ésimo componente principal é definido por:

$$Z_i = \boldsymbol{e_i'Y} = e_{i1}Y_1 + e_{i2}Y_2 + ... + e_{ip}Y_p, \quad i = 1, 2, ..., p.$$

- Como consequências:

$$Var(Z_i) = \boldsymbol{e_i'\Sigma e_i} = \lambda_i, \quad i = 1, 2, ..., p.$$

$$Cov(Z_i, Z_k) = \boldsymbol{e_i'\Sigma e_k} = 0, \quad i \neq k.$$

Análise de componentes principais

- Adicionalmente, como resultado da decomposição espectral:

$$\sum_{i=1}^{p} Var(Y_i) = \sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \lambda_1 + \lambda_2 + \lambda_p = \sum_{i=1}^{p} Var(Z_i)$$

- Assim, a proporção da variação total dos dados explicada pelo i-ésimo componente pode ser calculada por:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_p}, \quad i = 1, 2, ..., p.$$

Análise de componentes principais

- A correlação entre o i-ésimo componente (Z_i) e a k-ésima variável (Y_k) é dada por:

$$\rho_{Z_i,Y_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sigma_{kk}}.$$

Análise de componentes principais

- Em geral, observamos as proporções explicadas pelos componentes principais e retemos um pequeno número de componentes capazes de explicar boa parte da variabilidade dos dados;

- Os coeficientes (**cargas**) dos componentes (elementos dos autovetores), seus sinais e magnitudes, permitem interpretar os componentes e avaliar a importância das variáveis em sua constituição;

- As correlações entre componentes e variáveis podem ser usadas também pra avaliar importância das variáveis na constituição dos componentes e produção de gráficos;

Análise de componentes principais

- Chamamos **escores** os valores dos componentes (resultados das combinações lineares) calculados para um particular indivíduo;

- Os escores podem ser usados para efeito de ranqueamento dos indivíduos, agrupamento, classificação...

Análise de componentes principais

- Nas situações em que as variáveis originais têm escalas diferentes, é natural esperarmos que as variâncias, consequentemente, sejam diferentes.

- Para evitar distorções nos resultados, ocasionadas pela predominância de variáveis com maior variação, é recomendável considerar a análise de componentes principais com base nas variáveis padronizadas:

$$Y_1^* = \frac{Y_1 - \mu_1}{\sqrt{\sigma_{11}}}, \quad Y_2^* = \frac{Y_2 - \mu_2}{\sqrt{\sigma_{22}}}, ..., Y_p^* = \frac{Y_p - \mu_p}{\sqrt{\sigma_{pp}}}$$

Análise de componentes principais

- Os componentes principais para $Y_1^*, Y_2^*, ..., Y_p^*$ são determinados pela matriz de variâncias e covariâncias das variáveis padronizadas, o que equivale à matriz de correlações de $Y_1, Y_2, ..., Y_p$.

- Assim, teremos $Z_i^* = \boldsymbol{e_i'} \boldsymbol{Y_i^*}, i = 1, 2, ..., p$, com:

$$\sum Var\, Z_i^* = \sum Var\, Y_i^* = p$$

e

$$\rho_{Y_i^*, Z_k^*} = e_{ik}\sqrt{\lambda_i}, \quad i, k = 1, 2, ..., p,$$

sendo $(\lambda_1, \boldsymbol{e_1}), (\lambda_2, \boldsymbol{e_2}), ..., (\lambda_p, \boldsymbol{e_p})$ os autovalores e autovetores da matriz de correlações.

Análise de componentes principais

- Na prática, aplicaremos a ACP a uma amostra, de tal forma que a análise é conduzida com base em uma amostra $\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_n}$, com estatísticas amostrais $\bar{\boldsymbol{y}}$, $\boldsymbol{S}$ e $\boldsymbol{R}$.

- Seja $\boldsymbol{S}$ a matriz de covariâncias amostral, com $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), (\hat{\lambda}_2, \hat{\boldsymbol{e}}_2), \ldots, (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ os pares de autovalores e autovetores, tal que $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$.

- O i-ésimo componente principal é dado por:

$$\hat{z}_i = \hat{\boldsymbol{e}}_{\boldsymbol{i}}' \boldsymbol{y} = \hat{e}_{i1} y_1 + \hat{e}_{i2} y_2 + \ldots + \hat{e}_{ip} y_p, \qquad i = 1, 2, \ldots, p$$

Análise de componentes principais

- A variância amostral de $\hat{z}_i$ é igual a $\hat{\lambda}_i$, e a covariância amostral entre $\hat{z}_i$ e $\hat{z}_i$ é igual a zero;

- A variância amostral total fica dada por:

$$\sum_{i=1}^{n} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + ... + \hat{\lambda}_p$$

- A correlação entre o i-ésimo componente e a k-ésima variável é dada por:

$$r_{\hat{z}_i, y_k} = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}, i, k = 1, 2, ..., p.$$

Análise de componentes principais

- Um resultado fundamental da ACP é a produção de escores, que correspondem a valores calculados para os indivíduo nos componentes obtidos.

- Mesmo que a ACP seja realizada a partir das variáveis em sua escala original, é usual calcular os escores com base nas variáveis centradas nas respectivas médias. Por exemplo, para um indivíduo g, no *i-ésimo* componente:

$$\hat{z}_{g\,i} = \hat{\mathbf{e}}_{\boldsymbol{i}}'(\boldsymbol{y} - \bar{\boldsymbol{y}}) = e_{i1}(y_{g1} - \bar{y}_1) + e_{i2}(y_{g2} - \bar{y}_2) + \ldots + e_{ip}(y_{gp} - \bar{y}_p).$$

- Como resultado, os componentes principais obtidos terão suas médias amostrais iguais a 0.

Análise de componentes principais

- Já no caso em que a análise é conduzida com base nas variáveis padronizadas (matriz de correlações), os escores são calculados a partir delas:

$$\hat{z}_{g\,i} = e_{i1}\frac{y_{g1} - \bar{y}_1}{\sqrt{s_{11}}} + e_{i2}\frac{y_{g2} - \bar{y}_2}{\sqrt{s_{22}}} + \ldots + e_{ip}\frac{y_{gp} - \bar{y}_p}{\sqrt{s_{pp}}}.$$

Análise de componentes principais

- Um dos principais atrativos da ACP é a possibilidade de visualização dos resultados em gráficos de duas ou três dimensóes, como resultado da projeção. Algumas alternativas:
  - Plotar os escores de cada indivíduo em cada componente, a fim de identificar grupos de indivíduos com escores (e características) semelhantes, identificar indivíduos extremos,...
  - Plotar as correlações entre variáveis e componentes, a fim de visualizar as associações entre variáveis e componentes, as intercorrelações entre variáveis...
  - Plotar conjuntamente dados e variáveis num mesmo gráfico, permitindo visualizar globalmente relações entre variáveis e indivíduos (gráfico biplot, veremos adiante).

Análise de componentes principais

- Uma das questões mais recorrentes em ACP é o número de componentes a reter e explorar na análise.
- Um dos aspectos a se levar em consideração é a interpretação prática dos componentes, possível justificativa à luz da teoria da área...
- Obviamente, no entanto, aspectos referentes ao desempenho da técnica devem ser considerados.

Análise de componentes principais

- Alguns critérios que fundamentam a definição do número de componentes a serem "retidos":
  - Reter o menor número k de componentes principais capaz de explicar uma porcentagem mínima desejável da variação dos dados (Ex: 70 ou 80%);
  - Reter os componentes cujos autovalores são maiores que a média dos autovalores, $(\sum_{i=1}^{p} \lambda_i / p)$. Para a matriz de correlações a média é igual a 1;
  - Decidir o número de componentes com base na apreciação do *scree plot*;
  - Testar a significância dos componentes principais.

Análise de componentes principais

- Um teste a ser considerado em ACP é o teste de completa independência entre as variáveis.
- Caso as variáveis sejam conjuntamente independentes, não há porque fazer uma ACP, uma vez que, nesse caso, cada variável individualmente dominará um componente.
- Nesse caso, consideramos a hipótese nula $H_0 : \boldsymbol{P} = \boldsymbol{I}$, sendo $\boldsymbol{P}$ a matriz de correlações populacional. A estatística do teste é dada por:

$$u = -\left(n - 1 - \frac{2p + 5}{6}\right) * ln\left|\boldsymbol{R}\right|,$$

que segue, sob H_0, distribuição χ^2_f, com $f = \frac{1}{2}p(p-1)$. Rejeitaremos H_0, a um nível α, se $u > \chi^2_f(\alpha)$ (Teste de Bartlett).