Análise Multivariada
Análise de Componentes Principais

Prof. George von Borries

Departamento de Estatística
Universidade de Brasília

2023

Karl Pearson

Otimização Geométrica - *Procedimento para encontrar linhas e planos para melhor ajustar um conjunto de pontos no espaço p-dimensional - Philosophical Magazine, 1901.*

Harold Hotelling

Otimização Algébrica - *Procedimento para encontrar um número reduzido de variáveis independentes com boa representação do conjunto original de variáveis - Journal of Educational Psychology, 1933.*

Advento da Computação

Década de 60 - *Explosão de aplicações e desenvolvimento da técnica.*
70 em diante - *Refinamento para tratar com bases super dimensionadas e mineração de dados.*
Atual - *pesquisa para aplicação em dados esparsos* ($n << p$).

Ideia: Reduzir a dimensão de um conjunto de dados através de componentes principais que retem a maior parte da variação existente nas variáveis originais.

Objetivos

- Redução de dados.
- Interpretação de dados através da estrutura de $\mathbf{\Sigma}$, $\mathbf{S}$ ou $\mathbf{R}$.
- Detecção de *outliers* e/ou *clusters* nos dados.
- Construir grupos reduzidos de variáveis para utilizar em outros procedimentos de análise.
- Investigar normalidade multivariada nos dados.
- Criar índices e pesos para grupos de variáveis.

Exemplo: Iris Data contém dados de um género de plantas (IRIS) com flor, muito apreciado por suas diversas espécies, que apresentam flores de cores muito vivas. O nome popular é lírio.

O problema é classificar as flores nos três tipos denominados setosa, versicolor e virginica. As características disponíveis são: largura e comprimento da sépala (parte da flor que dá sustentação a pétala) e pétala ($p = 4$).

Três tipos de flores Iris: (a) setosa, (b) versicolor, (c) virginica.
Fonte: Murphy, K.P. *Machine Learning*, 2012.

Os dados apresentam uma estrutura de variação que deve ser preservada ao máximo quando projetamos e rotacionamos.

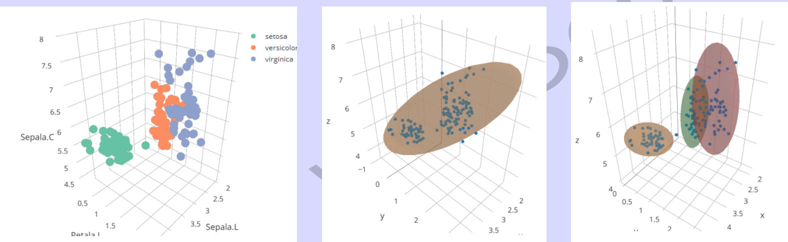

Figura: Iris - variáveis Sepala.L, Petala.L e Sepala.C.

A combinação linear de variáveis aleatórias $X_1, X_2, \ldots, X_p$ determina um novo sistema de coordenadas através da rotação do sistema original.

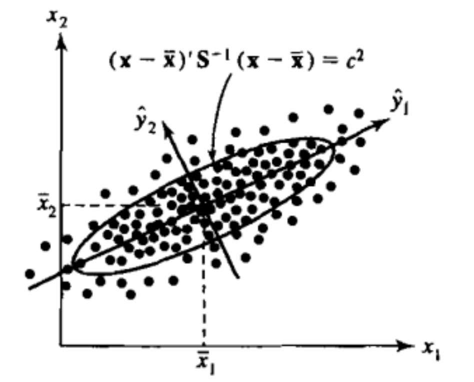

Figura: Rotação de eixos (J&W).

Definição 1: PCA Populacional

Seja um vetor aleatório p-dimensional $\mathbf{X}^\top = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p]$ com matriz de variância-covariâncias $\mathbf{\Sigma}$, autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$ e autovetores $\mathbf{e}_i$, $i = 1, \ldots, p$.

A i-ésima componente principal (CP) é representada por

$$\mathrm{Y}_i = \mathbf{e}_i^\top \mathbf{X} = \mathrm{e}_{i1} X_1 + \mathrm{e}_{i2} X_2 + \ldots + \mathrm{e}_{ip} X_p \quad i = 1, \ldots, p$$

tal que

$$\mathrm{Var}(\mathrm{Y}_i) = \mathbf{e}_i^\top \Sigma \mathbf{e}_i = \lambda_i \quad i = 1, \ldots, p$$

$$\mathrm{Cov}(\mathrm{Y}_i, \mathrm{Y}_k) = \mathbf{e}_i^\top \Sigma \mathbf{e}_k = 0 \quad i \neq k$$

Resultados:

- As **Componentes Principais** são combinações lineares não correlacionadas que maximizam $\mathrm{Var}\{\mathbf{Y}_i\}$.
- **CPs** podem ser representadas por qualquer combinação linear tal que $\mathbf{Y}_i = \mathbf{a}_i^\top \mathbf{X}$ maximiza $\mathrm{Var}\{\mathbf{Y}_i\}$, sujeito a $\mathbf{a}_i^\top \mathbf{a}_i = 1$ e $\mathrm{Cov}(\mathbf{a}_i^\top \mathbf{X}, \mathbf{a}_k^\top \mathbf{X}) = 0$ para todo $k < i$.
- A primeira CP maximiza $\mathbf{a}_1^\top \mathbf{X}$, sujeito a $\mathbf{a}_1^\top \mathbf{a}_1 = 1$.

Teorema 1: O critério a ser maximizado é $\mathrm{Var}\{\mathbf{Y}\} = \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}$ que é obtido com λ_1 quando $\mathbf{a} = \mathbf{e}_1$.

Teorema 2: Sejam $\mathbf{Y}_1, \ldots, \mathbf{Y}_p$ as componentes principais obtidas na Definição (1). Então,

$$\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \sum_{i=1}^{p} \mathrm{Var}(\mathbf{X}_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(\mathbf{Y}_i)$$

Proporção da Variância Total explicada pela k-ésima CP:

$$\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}, \quad k = 1, \ldots, p.$$

A ideia é substituir $\mathbf{X}$ por $\mathbf{Y}$ sem perda significativa de informação.

- $\mathbf{Y}_i$ e $\mathbf{Y}_k$ são não correlacionados para $i \neq k$.

Teorema 3: A correlação entre $\mathbf{Y}_i$ e $\mathbf{X}_k$ pode ser expressa como,

$$\rho_{\mathbf{Y}_i, \mathbf{X}_k} = \frac{\mathbf{e}_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

onde $\mathrm{Var}(\mathbf{X}_k) = \sigma_{kk}$, $\mathbf{Y}_i = \mathbf{e}_i^\top \mathbf{X}$, $(\lambda_i, \mathbf{e}_i)$ é o i-ésimo autovalor-autovetor de $\mathbf{\Sigma}$.

Resumo e Observações:

1. CPs representam a seleção de um novo sistema de coordenadas. Este sistema é obtido pela rotação dos eixos originais em novos eixos que fornecem uma estrutura mais simples de variância-covariância dos dados.
    - A primeira CP representa a direção de maior variabilidade.
    - A segunda CP representa a direção de maior variabilidade que é ortogonal a primeira CP.
    - O processo de construção das CPs continua até obtermos a CP de menor variabilidade, ortogonal a todas as outras CPs.
2. Podemos ver o problema como a minimização da soma de quadrados das distâncias entre os pontos no espaço e os eixos definidos pelas CPs.
3. CPs só exigem suposições sobre a distribuição das variáveis se estivermos interessados em inferências (testes e intervalos) com dados amostrais.
4. O Biplot de Gabriel é muito útil na ACP.

PCA com Matriz de Correlações S

- Indicado quando as variáveis possuem diferentes escalas de medida ou as variâncias são muito diferentes.

  Cuidado: Segundo Khattree e Naik[1], o uso da matriz de correlações $\boldsymbol{\rho}$ pode prejudicar o objetivo da indentificação de variáveis que mais contribuem com a variação total.

  Importante: CPs obtidas da matriz de covariâncias produz resultados diferentes da obtida com a matriz de correlações.

- Seja $\mathrm{Z}_i = \frac{\mathrm{X}_i - \mu_i}{\sigma_{ii}}$, i.e., $\mathbf{Z} = (\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \boldsymbol{\mu})$, com

$$\mathbf{V} = \begin{pmatrix} \sigma_{11} & 0 & \dots & 0 \\ 0 & \sigma_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{pp} \end{pmatrix}$$

[1] Multivariate Data Reduction and Discrimination with SAS Software, 2000.

Temos que

$$\mathbf{Z} = (\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \boldsymbol{\mu})$$

Assim,

$$\mathrm{E}(\mathbf{Z}) = (\mathbf{V}^{1/2})^{-1}(\mathrm{E}(\mathbf{X}) - \boldsymbol{\mu}) = \mathbf{0}$$

e

$$\mathrm{Cov}(\mathbf{Z}) = (\mathbf{V}^{1/2})^{-1}\mathrm{Cov}(\mathbf{X})(\mathbf{V}^{1/2})^{-1} = (\mathbf{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\mathbf{V}^{1/2})^{-1} = \boldsymbol{\rho}$$

Então,

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

implica em

$$\sum_{i=1}^{p} \mathrm{Var}(\mathrm{Y}_i) = \sum_{i=1}^{p} \mathrm{Var}(\mathrm{Z}_i) = p$$

e

$$\rho_{\mathrm{Y}_i,\mathrm{Z}_k} = e_{ik}\sqrt{\lambda_i} \quad i, k = 1, \ldots, p$$

A proporção total da variância explicada pela CP_i será $\lambda_i/p, \; i = 1, \ldots, p$.

Em relação as duas abordagens,

- O percentual da variância de cada componente de **R** difere do percentual de **S**.
- Os coeficientes das componentes são diferentes nos dois casos.
- **R** e as respectivas CPs são invariantes em escala.
- Os componentes de uma dada matriz **R** podem ser também de outra matriz $\mathbf{R}_2$. Logo a contribuição percentual da primeira componente não é uma observação muito útil.
- O uso de CPs com matriz de correlações deve ser restrito a problemas com variáveis em escalas muito diferentes.

Quantas componentes reter no estudo?

- Método 1: gráfico (**Scree Plot**) de $\lambda_i \times i$ com os autovalores em ordem decrescente.
- Método 2: Reter componentes suficientes para explicar um certo percentual da variância total.
- Método 3: Reter as componentes com $\lambda_i > \sum_{i=1}^{p} \lambda_i / p$. Para a matriz de correlações esta média é igual a 1.
- Método 4: (**Teste da significância das componentes principais**) este teste assume normalidade multivariada dos dados (amostrais) e utiliza a abordagem de razão de verssimilhança.

  $\mathrm{H}_{0k} : \gamma_{p-k+1} = \ldots = \gamma_p$, sendo $\gamma_{i's}$ os autovalores populacionais.

  Faça $\bar{\lambda} = \sum_{i=p-k+1}^{p} \frac{\lambda_i}{k}$ e utilize a estatística do teste

$$u = \left(n - \frac{2p+11}{6}\right)\left(k \ln \bar{\lambda} - \sum_{i=p-k+1}^{p} \ln \lambda_i\right),$$

  que tem dist. aproximada χ^2_ℓ, aonde $\ell = (k-1)(k+2)/2$.

  Começar com $H_0 : \gamma_{p-1} = \gamma_p$. Se não rejeitar H_0, teste $\mathrm{H}_0 : \gamma_{p-2} = \gamma_{p-1} = \gamma_p$ e continue testando até rejeitar H_0

PCA de Dados Amostrais

- Mesmo procedimento substituindo $\boldsymbol{\mu}$ por $\bar{\mathbf{X}}$ e $\boldsymbol{\Sigma}$ por $\mathbf{S}$.
- Neste caso podemos estudar a dist. dos λ_i. Neste curso iremos assumir que $\mathbf{S}$ é um bom estimador de $\boldsymbol{\Sigma}$.
- Devemos verificar se $\mathbf{S}$ é positiva definida.
- Se $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{S} \longrightarrow \boldsymbol{\Sigma}$ para $n \longrightarrow \infty$ e

$$(\mathbf{X} - \bar{\mathbf{X}})^\intercal \mathbf{S}^{-1} (\mathbf{X} - \bar{\mathbf{X}}) = c^2$$

estima o contorno da f.d.p. da respectiva normal multivariada.

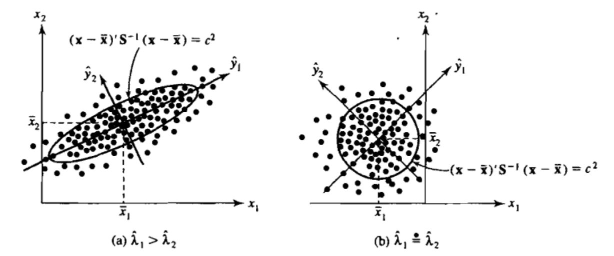

Figura 8.1 de J&W, pág. 449, sexta edição.

Sobre o cálculo de CPs

- PCA realiza a mesma redução de dimensionalidade da SVD quando aplicada numa matriz de dados centralizada (nas colunas).
- Em matrizes esparsas, a centralização dos dados pode destruir a esparsidade dos dados e por isso é preferivel aplicar SVD.
- O `R` apresenta várias funções para PCA:
  - `prcomp`: o cálculo é feito com base na SVD da matriz centralizada (e possivelmente padronizada).
  - `princomp`: o cálculo é feito utilizando a decomposição spectral da matriz de dados. Permite o uso da matriz de covariâncias ou correlações.
  - `eigen`, `PCA`, `pca`: ver Anderson[2] (2013) no site do curso.

[2] Anderson, G.B. (2013) *Principal Component Analysis in R: An examination of the different functions and methods to perform PCA*.

Exemplos

1. pca.R: Problema 8.10 de Johnson & Wichern: taxas de retorno semanais para ações de 5 empresas.
2. pca.R: Problema 8.18 de Johnson & Wichern: recordes nacionais em provas de corrida para mulheres.
3. pca.R: Quantas componentes reter no estudo?
4. Poster: Sousa, T.R. Análise de Componentes Principais Aplicada a Representação de Imagens 2D. Mestrado em Estatística, UnB, 2011.
5. Poster: von Borries, G.; Coutinho, M.; von Borries, R. e Miosso, C.J. PCA of Brain-Generated Biopotencial Measurements - IX Clatse, 2010.
6. TCC - Ana Carolina da Cruz (2018): A análise de componentes principais aplicada na análise de motilidade gástrica por cintilografia. Bacharelado em Estatística, UnB.