AULA2 - REDUÇÃO DE DIMENSIONALIDADE

PCA - Principal Component Analysis

MBA EM IA e BIGDATA

Profa. Dra. Roseli Aparecida Francelin Romero
SCC - ICMC - USP

2021

Sumário

Análise de Componentes Principais (PCA)

- Dado um conjunto de amostras, cada qual com um conjunto finito de variáveis.
- Derivar novas componentes que produzam uma descrição mais simples do sistema.
- Reduzir as variáveis originais a um número menor de variáveis ortogonais (não correlacionadas).
- Mudança de espaço de variáveis.

- É necessário apenas 01 dimensão

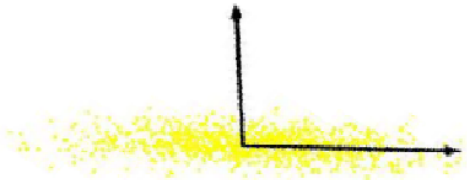


Figura 1: Dados bi-dimensionais

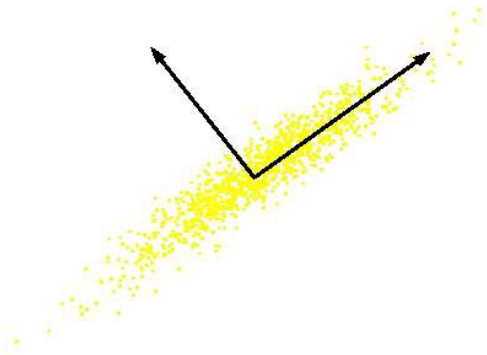


Figura 2: Dados bi-dimensionais rotacionados

- No caso em que os dados ficam sobre on perto de um subespaço d-dimensional, de dimensão mais baixa, os eixos deste subespaco são os ideais para representar os dados
- A identificação dos eixos é conhecida como **Análise de Componentes Principais** e pode ser obtida usando ferramentas clássicas de computação em matriz (**Decomposição de Valor Singular ou Auto Valor**).

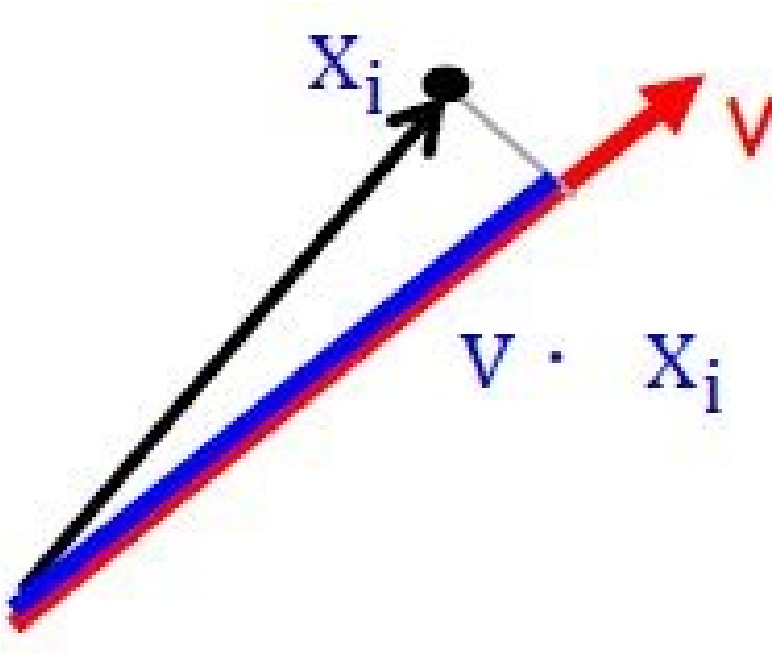


Figura 3: Projeção de xi sobre v

- $||v|| = 1$, Point xi (D-dimensional vector); Projection of xi sobre v is $v \times xi$

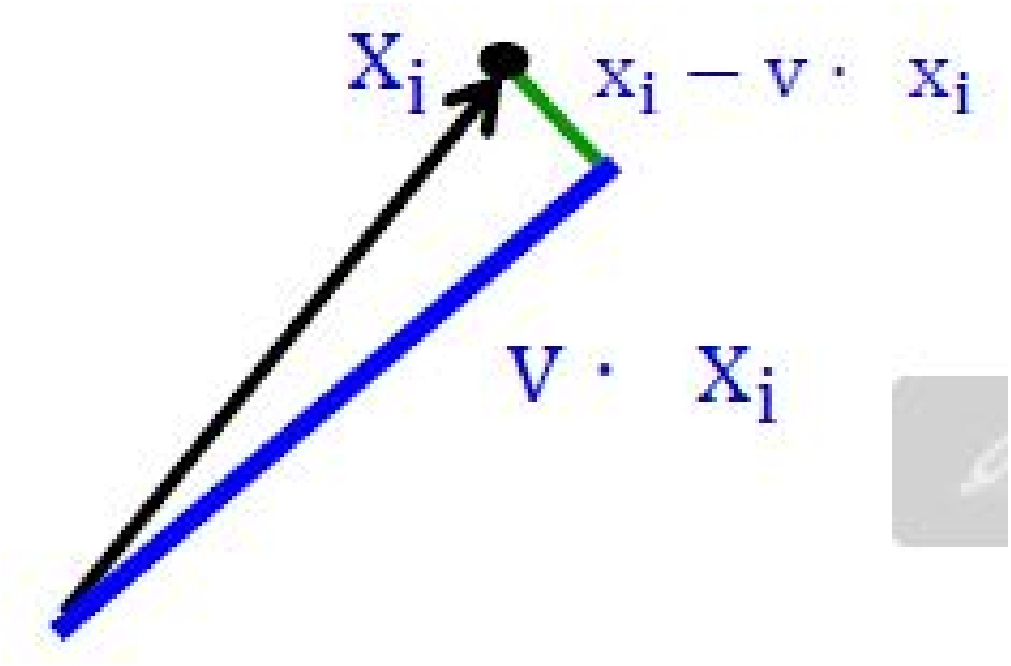


Figura 4: Componente principal é ortogonal

Componentes principais (PC) são direções ortogonais que capturam a maior parte da variação nos dados.

- Primeira PC - direção de maior variabilidade nos dados.
- A projeção de pontos de dados no primeiro PC discrimina os dados em qualquer direção (os pontos estão mais espalhados quando projetamos os dados nessa direção em comparação com outras direções).

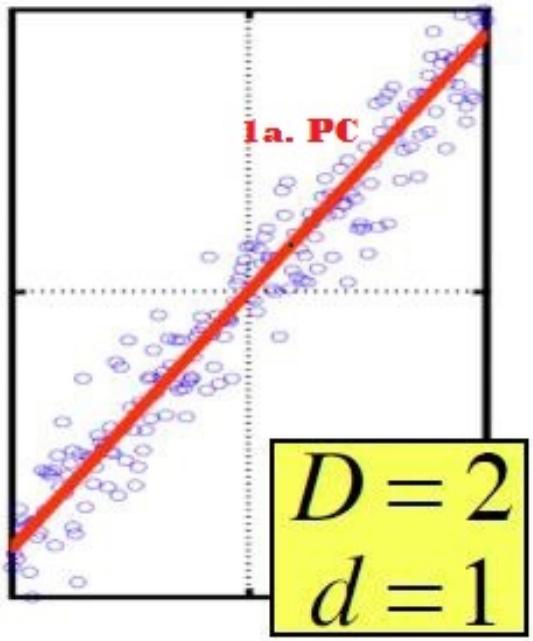
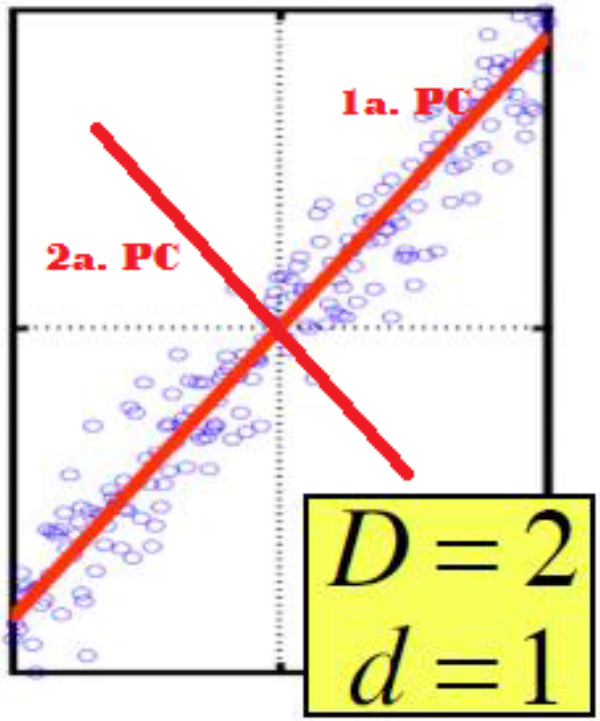


Figura 5: Primeira componente principal

- 2º PC - Próxima direção ortogonal (não correlacionada) de maior variabilidade (remova toda a variabilidade na primeira direção e encontre a próxima direção de maior variabilidade)
- E assim por diante



Análise de Componentes Principais (PCA)

Objetivo

Dadas p variáveis, $X_1, X_2, ..., X_p$, deseja-se achar combinações lineares dessas para produzir índices que **não** sejam correlacionados, de tal forma que:

- Índices Z: componentes principais.

Análise de Componentes Principais (PCA)

- i-ésima componente principal.

$$Z_i = v_{i1}X_1 + v_{i2}X_2 + \cdots + v_{ip}X_p$$

- Com restrição:

$$v_{i1}^2 + v_{i2}^2 + \cdots + v_{ip}^2 = 1$$

- E com $Z_1, Z_2, \ldots, Z_{i-1}, Z_i$ não correlacionados.

Análise de Componentes Principais (PCA)

- **PCA:** resume-se em encontrar os autovalores e autovetores da matriz C de covariância dos dados.

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pp} \end{bmatrix}$$

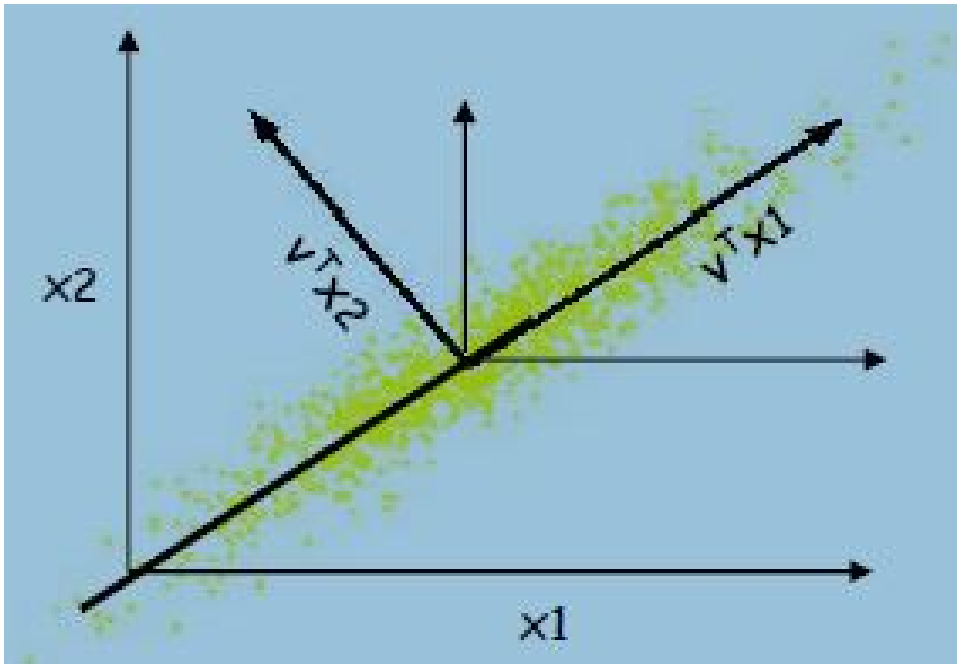
Análise de Componentes Principais (PCA)

- Supondo que os autovalores da matriz C estejam ordenados da seguinte forma:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_j \geq \cdots \geq \lambda_p$$

- Os autovetores associados são **não-correlacionados**:

$$v_1, v_2, \cdots, v_j, \cdots, v_p$$



Análise de Componentes Principais (PCA)

- **Propriedades:**

$$v_i^T v_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

- Para:

$$Z_i = v_{i1}X_1 + v_{i2}X_2 + \cdots + v_{ip}X_p$$

- $(v_{i1}, v_{i2}, \ldots, v_{ip})$ são os elementos do i-ésimo autovetor correspondente.

Análise de Componentes Principais (PCA)

- A soma dos autovalores corresponde ao traço da matriz covariância C:

$$\lambda_1 + \lambda_2 + \cdots + \lambda_p = c_{11} + c_{22} + \cdots + c_{pp}$$

- $var(Z_1) \geq var(Z_2) \geq \cdots \geq var(Z_p)$

$$var(Z_i) = \lambda_i$$

Sumário

Exemplo: conjunto Iris.dat

Tabela 1: Aplicando PCA na base de dados Iris.dat

		Autovetores (coeficientes)			
Componente	Autovalor	X_1	X_2	X_3	X_4
1	2.91082	0.522371	-0.263356	0.581254	0.565611
2	0.92122	0.372320	0.925556	0.021094	0.065417
3	0.14735	-0.721015	0.242033	0.140889	0.633804
4	0.02061	-0.261998	0.124137	0.801155	-0.523543

Exemplo: conjunto Iris.dat

- **Conclusões:**
  - Z_1 é responsável por 72.77% do total da variância.
  - Z_2 é responsável por 23.03% do total da variância.
  - Z_3 é responsável por 3.68% do total da variância.
  - Z_4 é responsável por 0.52% do total da variância.

Reconstrução dos dados originais

$$
\begin{aligned}
Z &= [Z_1, Z_2, \ldots, Z_p]^T \\
&= [X^T v_1, X^T v_2, \ldots, X^T v_{p-1}]^T \\
&= A^T X
\end{aligned}
$$

$$\text{matriz ortogonal} \rightarrow A^T = A^{-1}$$

$$X = A \cdot Z = \sum_{i=1}^{p} Z_i v_i$$

Sumário

Redução de dimensionalidade

- Sejam $\lambda_1, \lambda_2, \ldots, \lambda_m$ os m autovalores da matriz C.
- Então, $X' \sim X$, onde:

$$X' = \sum_{i=1}^{m} Z_i v_i \qquad m < p$$

- Erro: $e = X - X'$, de modo que:

$$e = \sum_{i=m+1}^{p} Z_i v_i$$

Redução de dimensionalidade

- O vetor de erro e é ortogonal ao vetor X', que aproxima X.
- **Princípio da ortogonalidade:** $e^T X' = 0$