MBA+

Artificial Intelligence & Machine Learning

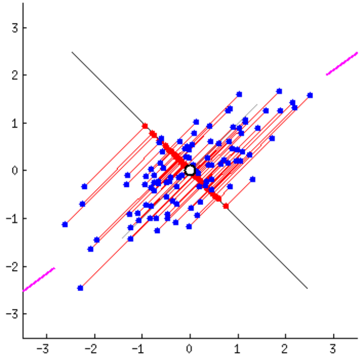
Prof. Dr. Vinicius F. Caridá

PCA

Principal Components Analysis

Análise de Componentes Principais (PCA)

A **análise dos componentes principais - ACP** ou **PCA** (do inglês *Principal Component Analysis*) é um método que tem por finalidade básica, a análise dos dados usados visando sua redução, eliminação de sobreposições e a **escolha das formas mais representativas de dados a partir de combinações lineares das variáveis originais**.



Análise de Componentes Principais (PCA)

Análise de Componentes Principais (PCA) é um dos métodos estatísticos de múltiplas variáveis mais simples. A PCA é considerada a **transformação linear ótima**, dentre as transformadas de imagens, sendo muito utilizada pela comunidade de reconhecimento de padrões.

Análise de Componentes Principais (PCA)

Análise dos Componentes Principais (PCA) é um método estatístico linear que encontra os autovalores e autovetores da matriz de covariância dos dados e, como esse resultado, pode-se realizar a redução dimensional dos dados e analisar os padrões principais de variabilidade presentes

Análise de Componentes Principais (PCA)

PCA é um método exploratório porque auxilia na elaboração de hipóteses gerais a partir dos dados coletados, contrastando com estudos direcionados nos quais hipóteses prévias são testadas. É também é capaz de separar a informação importante da redundante e aleatória.

A PCA também é muito utilizada em algoritmos de compressão de imagens. A característica básica da PCA é a redução do espaço necessário para a representação da imagem, já que a PCA promove uma compactação da energia.

Com o emprego da PCA a visualização de diversas variáveis em um determinado conjunto de dados torna-se mais produtiva, rápida, objetiva e eficiente.

Análise de Componentes Principais (PCA)

O PCA consiste em promover uma transformação linear nos dados de modo que os dados desta transformação tenham seus componentes mais relevantes nas primeiras dimensões, em eixos denominados principais. As figuras abaixo ilustram um conjunto bidimensional e o mesmo conjunto após a aplicação da PCA

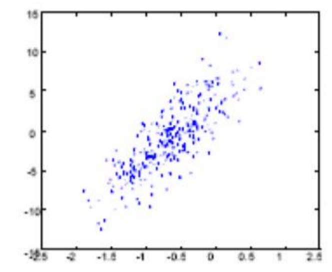


Figura 2 - Conjunto de dados

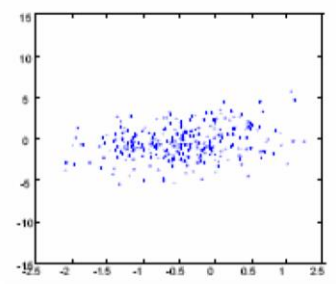
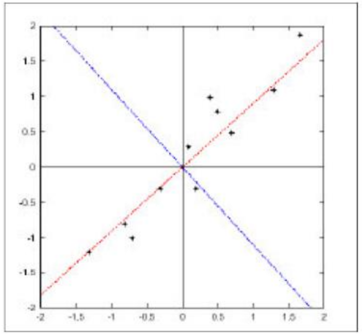


Figura 3 - Conjunto de dados após a PCA.

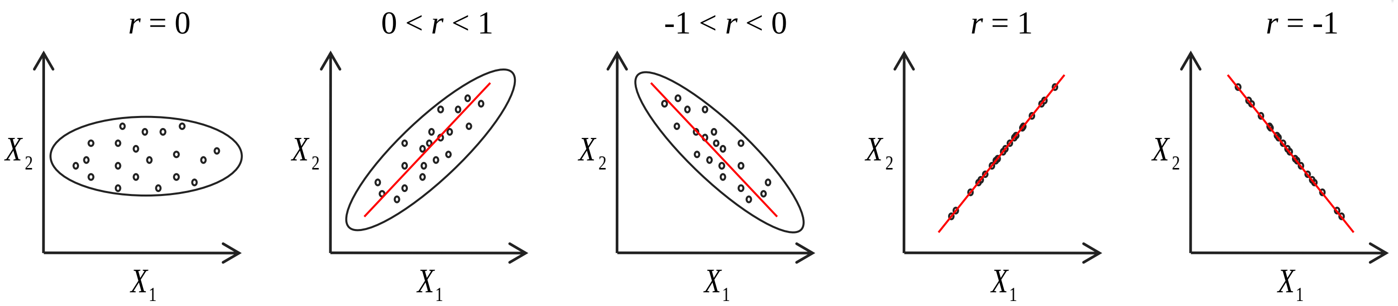
Análise de Componentes Principais (PCA)

A análise de componentes principais (PCA) é uma maneira de identificar a **relação entre características extraídas de dados**. É bastante útil quando os vetores de características têm muitas dimensões, quando uma representação gráfica não é possível, mas também pode ser útil em dimensões menores



A **componente principal** é o arranjo que melhor representa a distribuição dos dados (linha vermelha) e a **componente secundária** é perpendicular a componente principal (linha azul).

Associação entre Variáveis



$$r = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)Var(X_2)}}$$

Coeficiente de Correlação (de Pearson)

Quanto maior a variância, maior é a variabilidade e portanto maior a informação contida na variável. Num caso extremo, se a variância é zero, a variável não apresenta nenhuma informação a respeito do fenômeno por ela representada

Por sua vez, a covariância (e a correlação) é interpretada como uma redundância em análises múltiplas (2 ou mais variáveis), já que a informação contida numa variável está parcialmente representada em outra. Num caso extremo, para que utilizar duas variáveis perfeitamente correlacionadas ($|r| = 1$), uma vez que, conhecendo-se o valor de uma, pode-se inferir com precisão o valor da outra?

Análise de Componentes Principais (PCA)

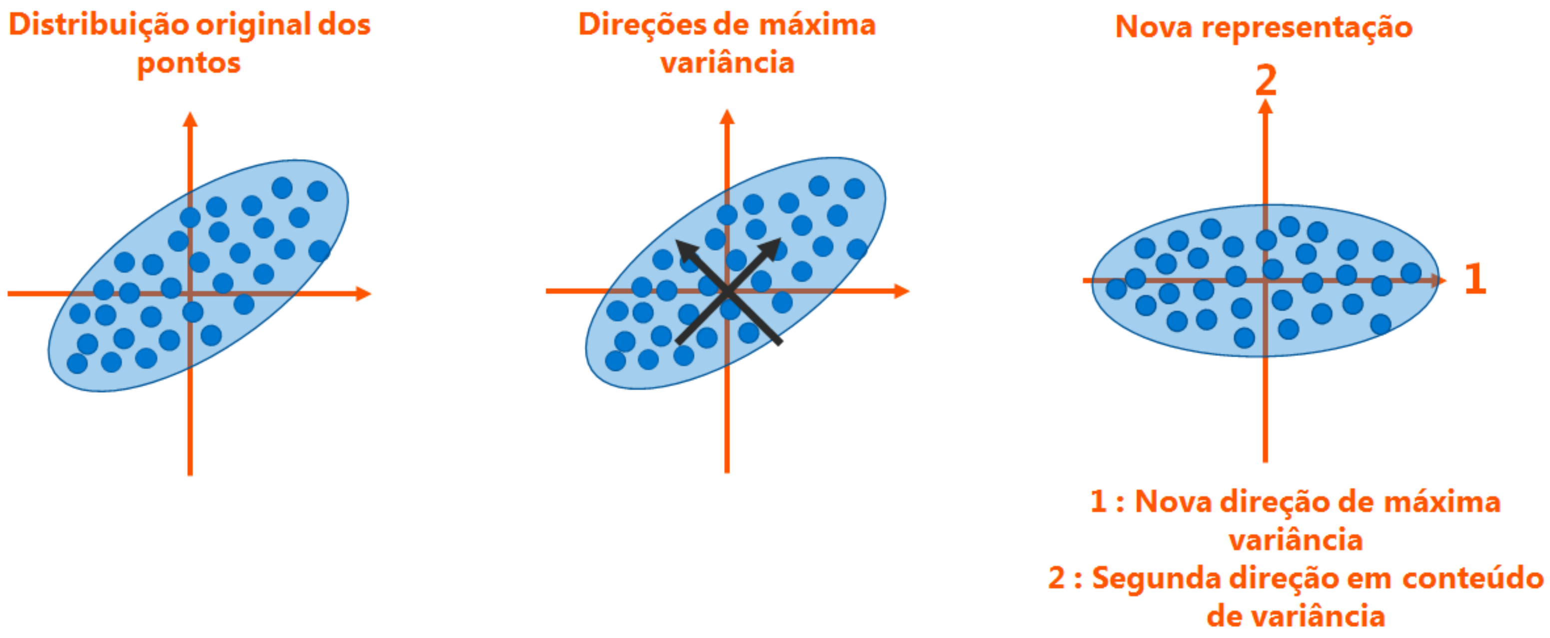
A covariância sempre é medida entre duas dimensões (calcular a covariância entre uma dimensão e ela mesma resulta na variância). A fórmula da covariância para dados de dimensão 2 (X e Y) é:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}[(X_i - \bar{X}) \cdot (Y_i - \bar{Y})]}{n}$$

Se os dados tiverem mais de duas dimensões, é necessário ter a covariância entre cada par de dimensões. A partir dessa idéia, surge a matriz de covariância. Se forem usadas três dimensões (*x*, *y* e *z*), a matriz de covariância terá este formato:

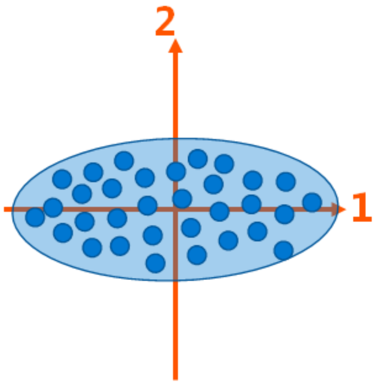
$$\boxed{matriz\_cov = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}}$$

Análise de Componentes Principais (PCA)



Análise de Componentes Principais (PCA)

Nova representação



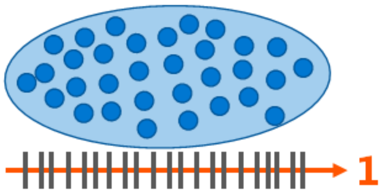
Ideia por trás de um PCA:

- Busca por direções (1) ortogonais entre si e (2) que exprimam consecutivamente um alto teor da variância dos dados

Parâmetros comuns de um PCA:

- Quantidade de variância: soma dos autovalores da matriz de covariância
- Número de componentes que desejamos reter

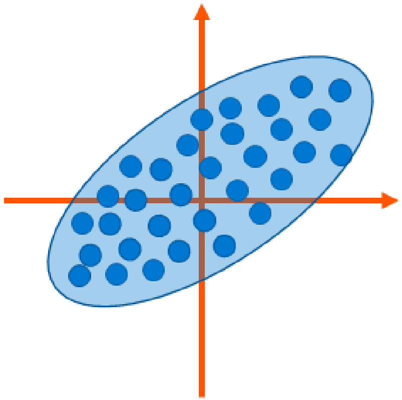
Retendo apenas a componente de maior quantidade de variância



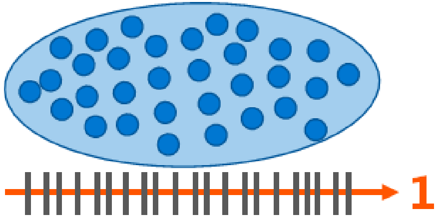
Ocorre uma redução de dimensionalidade de forma não-trivial

Análise de Componentes Principais (PCA)

Distribuição original dos pontos



Retendo apenas a componente de maior quantidade de variância



O que de fato aconteceu?

- Projeção dos dados em uma única direção
- Não é uma técnica de seleção de variáveis!

Problema: como recuperar os dados originais?

- Precisamos de todas as direções para fazer a transformação inversa!

Hipótese:

- Dados centrados em torno de 0

Autovalores e Autovetores

Os autovetores representam as transformações lineares aplicadas em m variáveis correlacionadas de modo a obter m variáveis não correlacionadas e podem ser representados na forma de uma matriz α ($m \times m$):

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{11} & \alpha_{21} & \cdots & \alpha_{m1} \\ \alpha_{12} & \alpha_{22} & \cdots & \alpha_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1m} & \alpha_{2m} & \cdots & \alpha_{mm} \end{bmatrix}$$

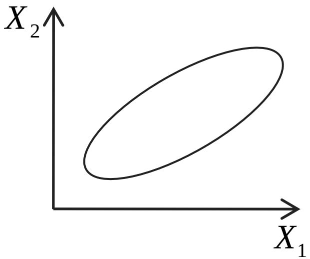
Nesta matriz, os autovetores estão organizados em colunas, ou seja, o 1º autovetor está na 1ª coluna, o 2º autovetor está na 2ª coluna e assim por diante.

Os autovalores representam a variância de cada componente (variável transformada) e podem ser representadas na forma de uma matriz diagonal λ ($m \times m$):

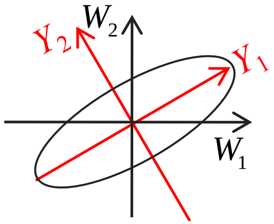
$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix} \qquad (\lambda_1 > \lambda_2 > \ldots > \lambda_m)$$

Transformação por Componentes Principais (exemplo 2D)

Como poderíamos eliminar a redundância entre X_1 e X_2?

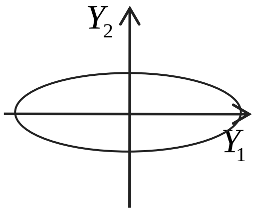


$Cov(X_1, X_2) > 0$



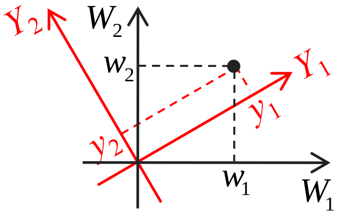
$Cov(W_1, W_2) = Cov(X_1, X_2)$

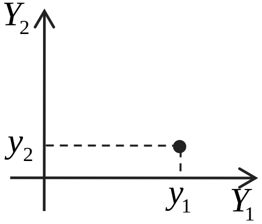
$W_j = X_j - \bar{X}_j$



$Cov(Y_1, Y_2) = 0$

rotação com preservação da ortogonalidade dos eixos





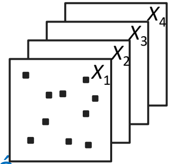
$Y_1 = \alpha_{11} W_1 + \alpha_{12} W_2$ $\quad Var(Y_1) = \lambda_1$ $\quad$ 1ª CP $\quad$ α são chamados de autovetores

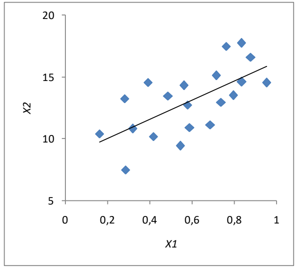
$Y_2 = \alpha_{21} W_1 + \alpha_{22} W_2$ $\quad Var(Y_2) = \lambda_2$ $\quad$ 2ª CP $\quad$ λ são chamados de autovalores ($\lambda_1 > \lambda_2$)

$Var(X_1) + Var(X_2) = \lambda_1 + \lambda_2$ $\quad$ informação total é preservada

Análise de Componentes Principais

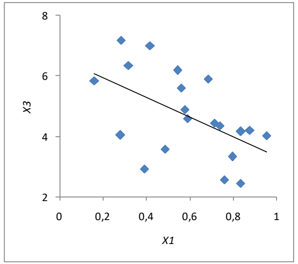
X_1	X_2	X_3	X_4
0,1596	10,4043	5,8457	11,7544
0,4160	10,1971	6,9976	14,3787
0,6841	11,1176	5,8942	11,9172
0,5421	9,4291	6,1867	12,8827
0,7133	15,0980	4,4443	9,1690
0,7587	17,4789	2,5728	5,2574
0,2824	7,4640	7,1848	15,1390
0,8336	14,5949	4,1581	9,2457
0,5752	12,7215	4,8841	10,5356
0,8308	17,7273	2,4392	5,8491
0,9510	14,5368	4,0345	8,5466
0,2796	13,2739	4,0397	8,3507
0,7357	12,9786	4,3470	8,9358
0,8759	16,5915	4,2100	9,0698
0,3904	14,5826	2,9220	6,4501
0,5584	14,3165	5,5950	11,6657
0,4849	13,4711	3,5743	7,4431
0,3169	10,8015	6,3426	13,0677
0,5869	10,9385	4,5867	9,6230
0,7940	13,5060	3,3238	7,4271

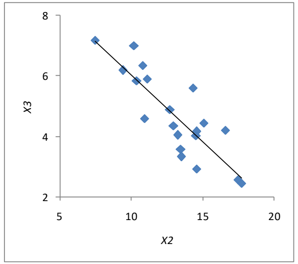




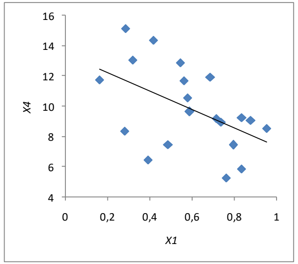
Matriz de Variância-Covariância

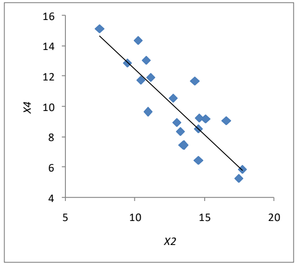
$\boxed{0,0516}$	0,4024	-0,1683	-0,3171
0,4024	$\boxed{7,3698}$	-3,2402	-6,3600
-0,1683	-3,2402	$\boxed{1,9737}$	3,8806
-0,3171	-6,3600	3,8806	$\boxed{7,6995}$

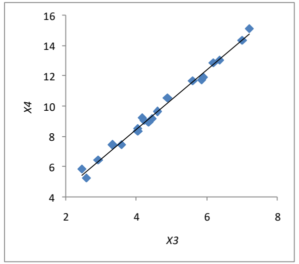




Autovalores			
15,7615	1,2900	0,0310	0,0121
Autovetores			
0,0340	0,0583	0,9481	0,3108
0,6485	0,7582	-0,0670	-0,0087
-0,3437	0,2787	-0,2837	0,8507
-0,6784	0,5865	0,1272	-0,4238

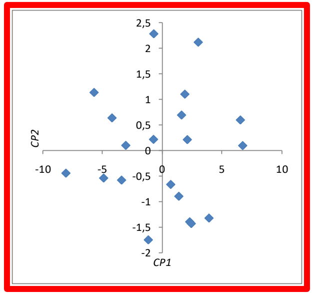






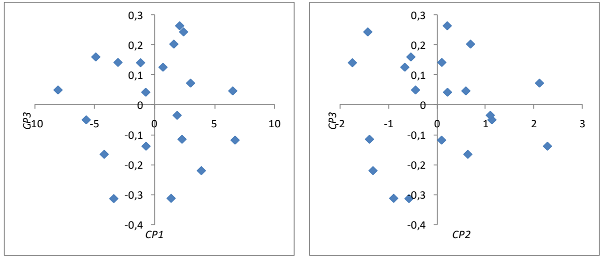
Análise de Componentes Principais

CP_1	CP_2	CP_3	CP_4
-3,4404	-0,5891	-0,3154	0,0689
-5,7422	1,1290	-0,0515	0,0181
-3,0872	0,0914	0,1410	0,1980
-4,9424	-0,5494	0,1594	0,0082
1,8577	1,0951	-0,0362	0,1037
6,7000	0,0872	-0,1194	0,1628
-8,0993	-0,4530	0,0486	-0,1626
1,5819	0,6859	0,2024	-0,1305
-0,7664	0,2092	0,0411	-0,1235
6,5079	0,5895	0,0455	-0,1814
2,0649	0,2042	0,2638	0,0977
1,3543	-0,9060	-0,3144	-0,0126
0,6757	-0,6745	0,1249	0,1452
2,9795	2,1135	0,0716	-0,0161
3,8802	-1,3335	-0,2217	-0,1349
-0,7436	2,2786	-0,1397	-0,0168
2,2647	-1,4066	-0,1165	0,0383
-4,2393	0,6300	-0,1669	-0,0195
-1,2009	-1,7602	0,1400	0,0294
2,3949	-1,4413	0,2433	-0,0723



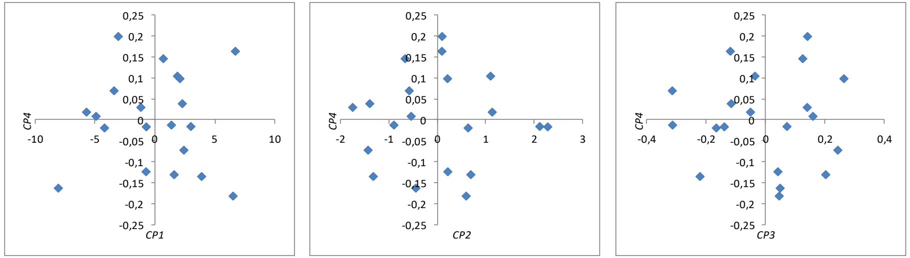
Matriz de Variância-Covariância

15,7615	0	0	0
0	1,2900	0	0
0	0	0,0310	0
0	0	0	0,0121



A primeira componente guarda 92,2% da variação total

As 2 primeiras CP, acumulam 99,7% da variação total



Análise de Componentes Principais (PCA)

Os **passos** para calcular as componentes principais são:

- **obter** os dados ou as ***M*** amostras de vetores de dimensão ***n***;
- calcular a **média** ou o **vetor médio** destes dados;
- **subtrair a média** de todos os itens de dados;
- calcular a **matriz de covariância** usando todas as subtrações. Ela é o resultado da média do produto de cada subtração por ela mesma e terá dimensão ***n*** x ***n***;
- calcular os **auto valores** e **auto vetores** da matriz de covariância.
- arranjar os a **matriz da Transformada de Hotelling** (cujas linhas são formadas a partir dos autovetores da matriz de covariância arranjados de modo que a primeira linha, o elemento (0,0), seja o auto vetor correspondente ao maior autovalor, e assim sucessivamente até que a última linha corresponda ao menor autovalor.

Observações

- A transformação prioriza as variáveis com maior variância

  Atenção: isso pode ocorrer quando as variáveis possuem diferentes grandezas

  Para neutralizar o efeito das variâncias de cada variável de modo que todas tenham o mesmo peso, os autovalores e autovetores devem ser calculados a partir da matriz de correlação (que equivale realizar a transformação sobre as variáveis padronizadas)

- Por usar covariância (ou correlação), esta transformação pressupõe relações lineares entre variáveis

- A interpretação dos valores de uma determinada componente principal pode ser bastante difícil, necessitando a avaliação do autovetor correspondente

Aplicações

- Diminuição da dimensionalidade do problema

  Ao invés de se trabalhar com 50 dimensões, escolhem-se as primeiras componentes que guardam a maior parte da informação total

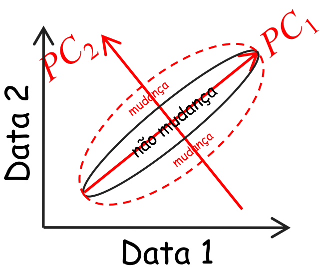
  Isso melhora o desempenho de classificadores que se baseiam em inversões de matrizes

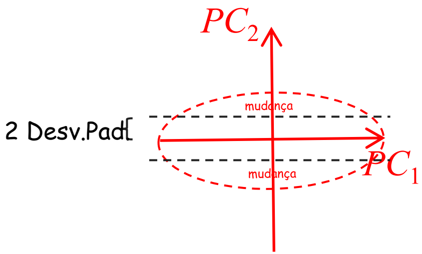
- Visualização de Dados

  A informação contida numa imagem hiperespectral pode ser visualizada usando-se as 3 primeiras componentes numa composição colorida RGB (obs: as cores podem ser de difícil interpretação)

- Detecção de Mudanças

  Numa comparação de 2 datas (mesma banda), os valores extremos da segunda componente evidenciam onde ocorreu mudança entre datas

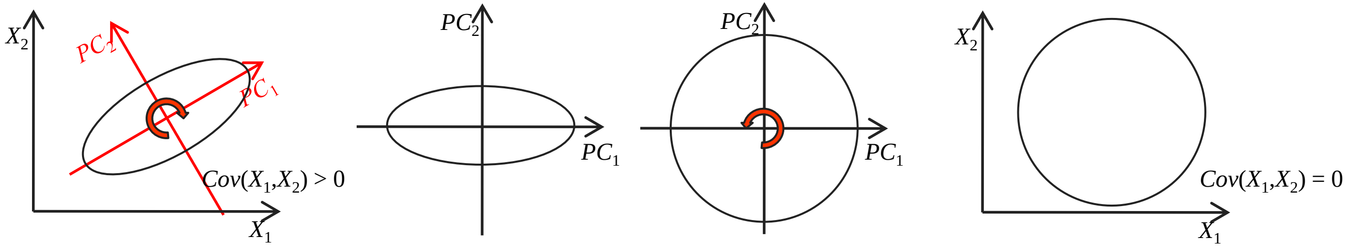




Aplicações

- Aumento de Contraste por Decorrelação
  Aplica-se a transformada por componentes principais, faz-se um aumento de contraste das componentes de modo que todas fiquem com a mesma variância e, por fim, faz-se a transformação inversa, restabelecendo-se as variáveis originais

- Simulação de dados correlacionados
  Calculam-se os autovalores e autovetores da matriz de variância-covariância. Simulam-se dados não-correlacionados com variância igual aos autovalores e faz-se a transformação inversa



Questions and Feedback

Thank you!

Obrigado !

V I N I

Vinicius Fernandes Caridá
vfcarida@gmail.com

@Vinicius Caridá

@Vinicius Caridá
@machine learning Brasil

@vfcarida

Anexos

Principal Components Analysis

Análise de Componentes Principais (PCA)

A matriz de transformação utilizada para o cálculo da PCA consiste em uma matriz cujas colunas são os autovetores da matriz de covariância estimada dos dados.

A matriz de transformação utilizada para o cálculo da PCA consiste em uma matriz cujas linhas são os autovetores da matriz de covariância estimada dos dados. A matriz de covariância Σ, é uma matriz simétrica e definida positiva, que possui informação sobre as variâncias em todos os eixos onde os dados estão distribuídos. Esta pode ser estimada como:

$$\Sigma \approx \frac{1}{N-1}\sum_i (\mathbf{x}_i - \mu)'(\mathbf{x}_i - \mu)$$

onde N é o número de amostras de dados xi, e μ é a média do conjunto.

Matriz de Variância-Covariância

Supondo que n amostras sejam avaliadas segundo m variáveis diferentes (atributos), podemos representar o conjunto de valores observados por um matriz $\mathbf{X}$, onde cada elemento x_{ij} representa o valor da i-ésima amostra para a j-ésima variável.

$$\mathbf{X} = \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & & X_m \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{nm} & \cdots & x_{nm} \end{bmatrix} \end{array}$$

A variância de cada variável e as covariâncias entre todos os pares de variáveis podem ser representadas através da matriz de variância-covariância $\Sigma_{\mathbf{X}}$ $(m \times m)$:

$$\Sigma_{\mathbf{X}} = \begin{array}{cc} \begin{array}{cccc} X_1 & X_2 & & X_m \end{array} & \\ \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_m) \\ Cov(X_1, X_2) & Var(X_2) & \cdots & Cov(X_2, X_m) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_1, X_m) & Cov(X_2, X_m) & \cdots & Var(X_m) \end{bmatrix} & \begin{array}{c} X_1 \\ X_2 \\ \\ X_m \end{array} \end{array}$$

$$Cov(X_1, X_2) = \frac{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{n-1}$$

$$Cov(X_k, X_l) = Cov(X_l, X_k)$$

$$Cov(X_k, X_k) = Var(X_k)$$

Cálculo dos Autovalores e Autovetores

Os autovalores (λ) e autovetores (α) são obtidos de modo a satisfazer a seguintes condições:

$\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}$ ou $\left(\boldsymbol{\Sigma}_{\mathbf{X}} - \lambda\mathbf{I}\right)\boldsymbol{\alpha} = \mathbf{0}$ sendo $\boldsymbol{\alpha}'\boldsymbol{\alpha} = \mathbf{I}$ (autovetores ortogonais)

Como os autovetores α são não nulos, então

$\left|\boldsymbol{\Sigma}_{\mathbf{X}} - \lambda\mathbf{I}\right| = 0$ gera um polinômio de grau m cujas raízes representam os m autovalores λ

Ordenam-se os autovalores do maior para o menor e para cada autovalor λ_k, calcula-se o autovetor α_k de modo que:

$$\left(\boldsymbol{\Sigma}_{\mathbf{X}} - \lambda_k\mathbf{I}\right)\boldsymbol{\alpha}_k = 0$$

Assim, os valores transformados para a componente principal k são calculados a partir da matriz $\mathbf{X}$:

$$\mathbf{Y}_k = \boldsymbol{\alpha}_k'\mathbf{X}'$$

Matriz de Correlação

Alternativamente, as relações entre as variáveis podem ser expressas através da matriz de correlação:

$$
r_{\mathbf{X}} = \begin{array}{c}
\begin{array}{cccc} X_1 & X_2 & & X_m \end{array} \\
\left[\begin{array}{cccc}
1 & r_{X_1,X_2} & \cdots & r_{X_1,X_m} \\
r_{X_1,X_2} & 1 & \cdots & r_{X_2,X_m} \\
\vdots & \vdots & \ddots & \vdots \\
r_{X_1,X_m} & r_{X_2,X_m} & \cdots & 1
\end{array}\right]
\end{array}
\begin{array}{c} X_1 \\ X_2 \\ \\ X_m \end{array}
$$

$$
r_{X_1,X_2} = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)Var(X_2)}}
$$

$$
r_{X_k,X_l} = r_{X_l,X_k}
$$

$$
r_{X_k,X_k} = 1
$$

Obs: em R, para computar a transformação por componentes principais utilizando a matriz de correlação, pode-se usar a função `pccor<-prcomp(x, scale=TRUE)`

Análise de Componentes Principais no R

```
x1<-c(0.1596,0.416,0.6841,0.5421,0.7133,0.7587,0.2824,0.8336,0.5752,0.8308,0.951,0.2796,0.7357,0.8759,0.3904,
    0.5584,0.4849,0.3169,0.5869,0.794)
x2<-c(10.4043,10.1971,11.1176,9.4291,15.098,17.4789,7.464,14.5949,12.7215,17.7273,14.5368,13.2739,12.9786,
    16.5915,14.5826,14.3165,13.4711,10.8015,10.9385,13.506)
x3<-c(5.8457,6.9976,5.8942,6.1867,4.4443,2.5728,7.1848,4.1581,4.8841,2.4392,4.0345,4.0397,4.347,4.21,2.922,
    5.595,3.5743,6.3426,4.5867,3.3238)
x4<-c(11.7544,14.3787,11.9172,12.8827,9.169,5.2574,15.139,9.2457,10.5356,5.8491,8.5466,8.3507,8.9358,9.0698,
    6.4501,11.6657,7.4431,13.0677,9.623,7.4271)
x<-cbind(x1,x2,x3,x4)
```

```
round(cov(x),4)
```

-       x1   x2   x3   x4
- x1 0.0516 0.4024 -0.1683 -0.3171
- x2 0.4024 7.3698 -3.2402 -6.3599
- x3 -0.1683 -3.2402 1.9737 3.8805
- x4 -0.3171 -6.3599 3.8805 7.6995

```
sum(diag(cov(x)))
```

- [1] 17.09458

```
pccov<-prcomp(x)
pccov
```

- Standard deviations:
- [1] 3.9700671 1.1358016 0.1759476 0.1101858 ← $\sqrt{\lambda}$

- Rotation:
-       PC1   PC2   PC3   PC4
- x1 0.03398365 -0.05831431 0.94807140 -0.310813742
- x2 0.64847514 -0.75822916 -0.06702796 0.008705756
- x3 -0.34373410 -0.27866559 -0.28371255 -0.850705320
- x4 -0.67835977 -0.58654366 0.12718124 0.423815393

```
sum(pccov$sdev^2)
```

- [1] 17.09458

Análise de Componentes Principais no R

summary(pccov)

- Importance of components:
- PC1 PC2 PC3 PC4
- Standard deviation 3.970 1.13580 0.17595 0.11019
- Proportion of Variance 0.922 0.07547 0.00181 0.00071
- Cumulative Proportion 0.922 0.99748 0.99929 1.00000

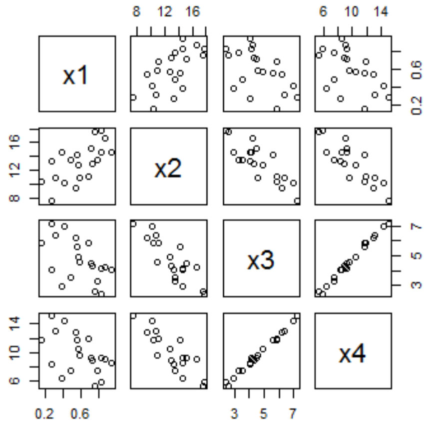
pc<-pccov$x #Calculando as componentes principais

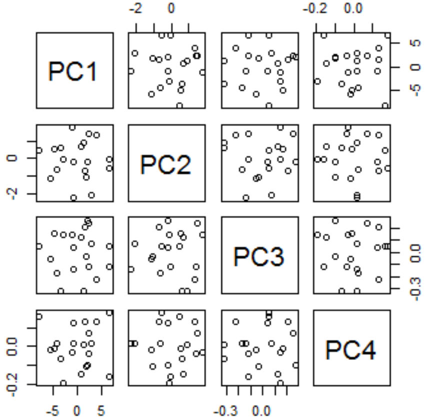
round(cov(pc),4)

- PC1 PC2 PC3 PC4
- PC1 15.7614 0.00 0.000 0.0000
- PC2 0.0000 1.29 0.000 0.0000
- PC3 0.0000 0.00 0.031 0.0000
- PC4 0.0000 0.00 0.000 0.0121

pairs(x)

pairs(pc)



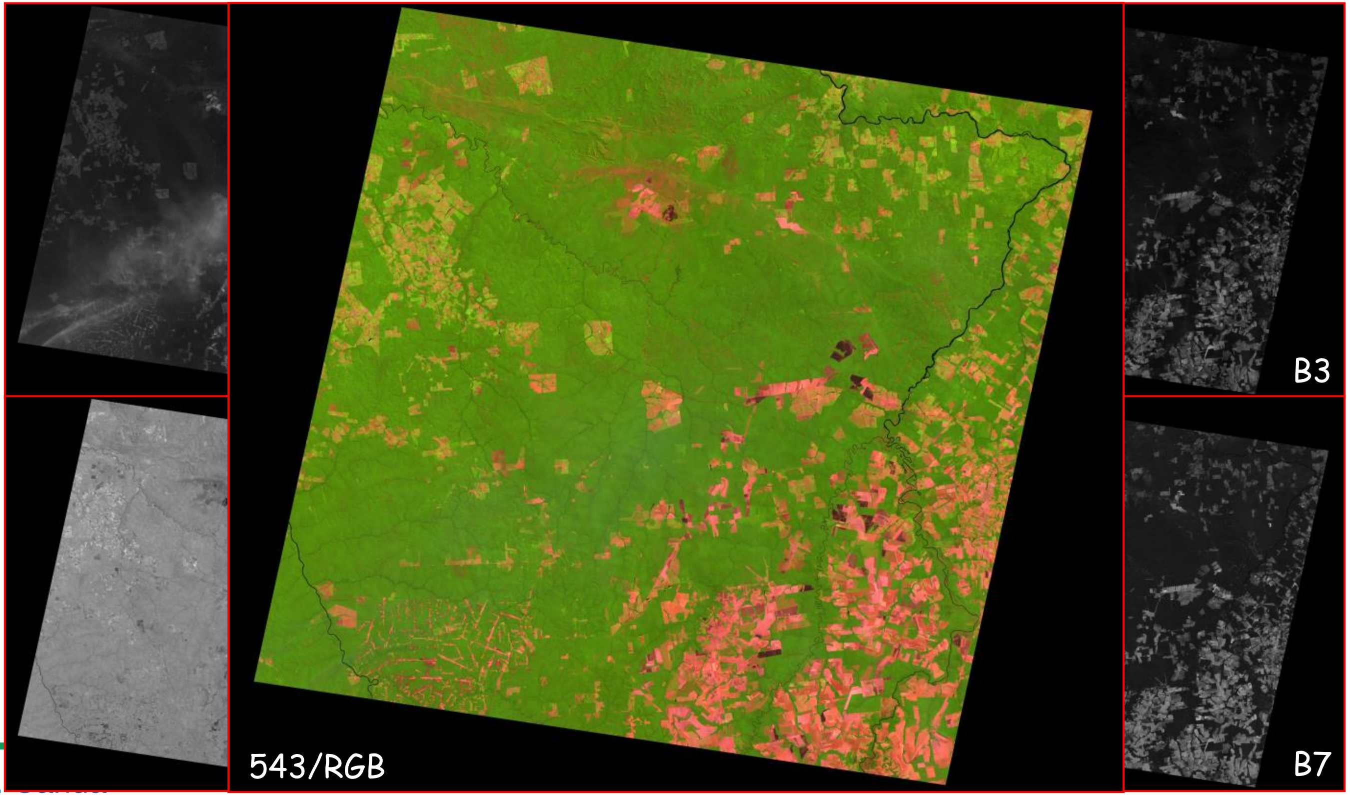


Visualização de Dados

TM/LANDSAT (Bandas 1 a 5 e 7) 227/68 ano 1999

Visualização de Dados

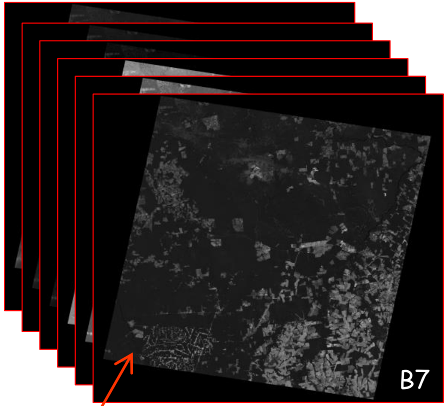
TM/LANDSAT (Bandas 1 a 5 e 7) 227/68 ano 1999 (ganho 2,45 e offset variável)



Visualização de Dados

TM/LANDSAT (Bandas 1 a 5 e 7) 227/68 ano 1999 (ganho 2,45 e offset variável)

		Média	D.Padrão				
	B1	45,556	19,300				
	B2	28,236	11,600				
	B3	25,484	20,837				
	B4	130,698	18,811				
	B5	108,052	44,184				
	B7	39,701	25,810				
Covariância		B1	B2	B3	B4	B5	B7
	B1	372,489	195,992	308,969	-62,672	493,634	295,327
	B2	195,992	134,554	228,131	-33,144	419,773	245,247
	B3	308,969	228,131	434,163	-107,410	837,561	499,226
	B4	-62,672	-33,144	-107,410	353,868	-161,005	-161,894
	B5	493,634	419,773	837,561	-161,005	1952,258	1094,010
	B7	295,327	245,247	499,226	-161,894	1094,010	666,157
		Autovetor					
Autovalor		B1	B2	B3	B4	B5	B7
3235,547	CP1	0,230	0,180	0,351	-0,088	0,766	0,444
345,967	CP2	0,079	-0,022	0,055	-0,966	-0,215	0,104
272,900	CP3	0,842	0,291	0,241	0,132	-0,328	-0,153
35,210	CP4	-0,333	0,169	0,549	0,167	-0,487	0,543
19,145	CP5	0,312	-0,291	-0,563	0,107	-0,150	0,683
4,720	CP6	0,157	-0,877	0,444	0,047	0,010	-0,085

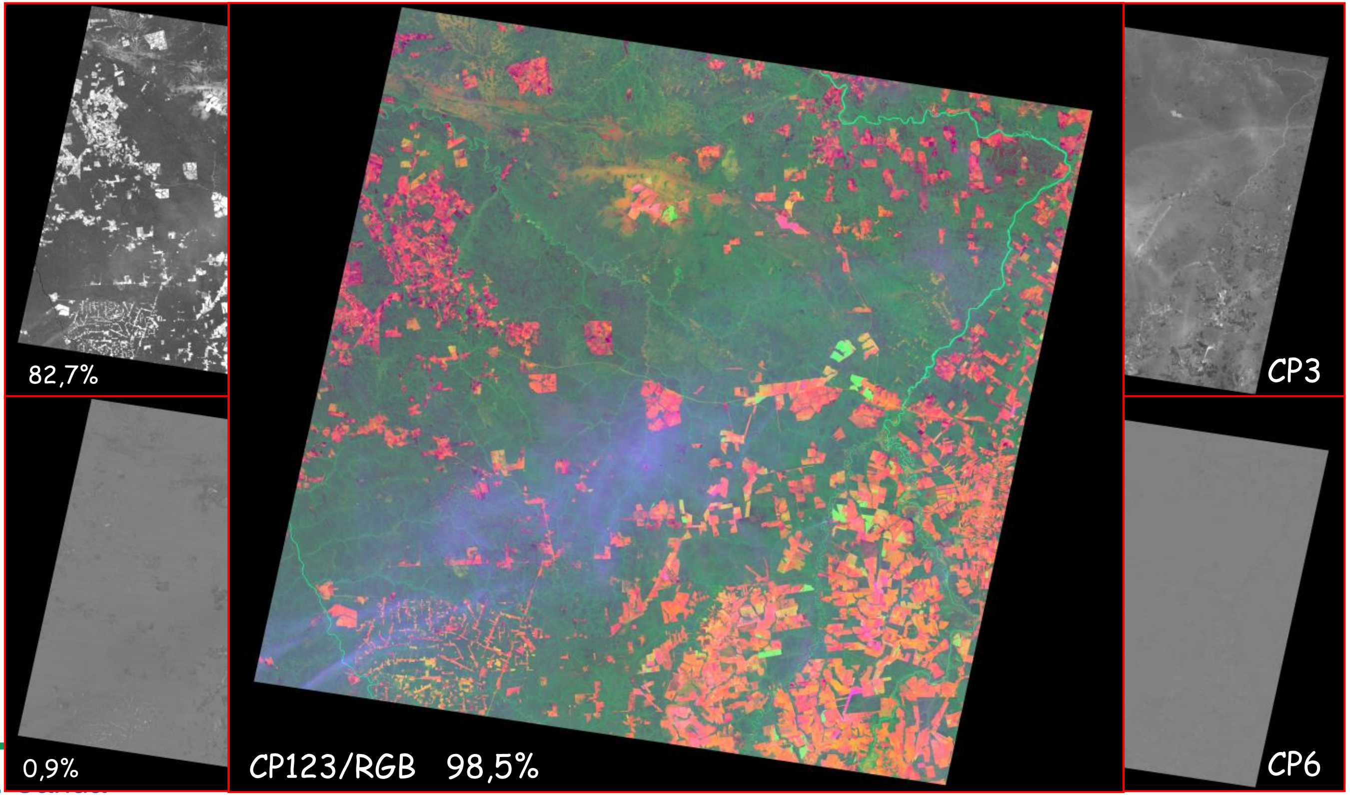


Importante: Para obtenção das estatísticas é necessário desconsiderar a região "sem informação"

OBS: Em geral, as imagens resultantes não são representadas em bytes (0 a 255), tendo valores positivos e negativos

Visualização de Dados

Componentes Principais

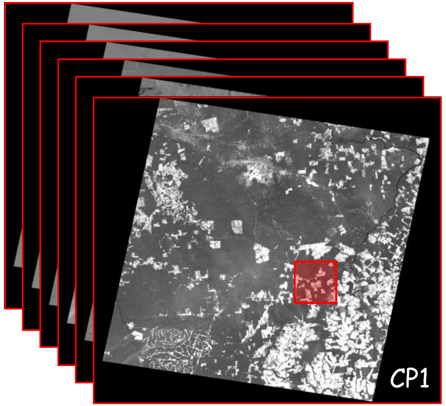


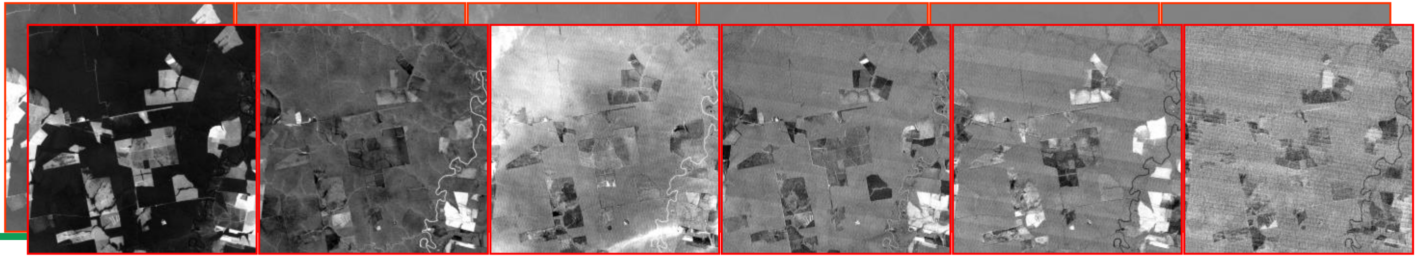
(mesmo ganho e offset da CP1)

Visualização de Dados

Componentes Principais

		Média	D.Padrão				
	CP1	0	56,882				
	CP2	0	18,600				
	CP3	0	16,520				
	CP4	0	5,934				
	CP5	0	4,375				
	CP6	0	2,172				
Covariância		CP1	CP2	CP3	CP4	CP5	CP6
	CP1	3235,55	0	0	0	0	0
	CP2	0	345,967	0	0	0	0
	CP3	0	0	272,900	0	0	0
	CP4	0	0	0	35,210	0	0
	CP5	0	0	0	0	19,145	0
	CP6	0	0	0	0	0	4,720





(ganho e offset arbitrários)