

Análise de Componentes Principais

Aluno: Marcus Freire

Orientador: Ricardo Rios

Data: 26/10/2015

O que é Análise de Componentes Principais (PCA em inglês)?

- É um método multivariada que nos permite estudar e explorar um conjunto de variáveis quantitativas medidas em um conjunto de objetos de dados
- Inventado por Pearson (1901) e Hotelling (1933)

Para que serve?

- Redução da Dimensão
- Visualização
- Extração de Características
- Compressão de Dados
- Suavização dos dados

Onde é usado?

➤ **Química**

Na Quimiometria a Análise dos Componentes Principais (PCA) é uma das ferramentas mais utilizadas, que visa principalmente à redução do número de variáveis, eliminação de dados redundantes e facilitar a interpretação dos dados.

➤ **Processamento Digital de Imagens**

Utilização da Análise de Componentes Principais na compressão de imagens digitais

➤ **Mercado de Finanças**

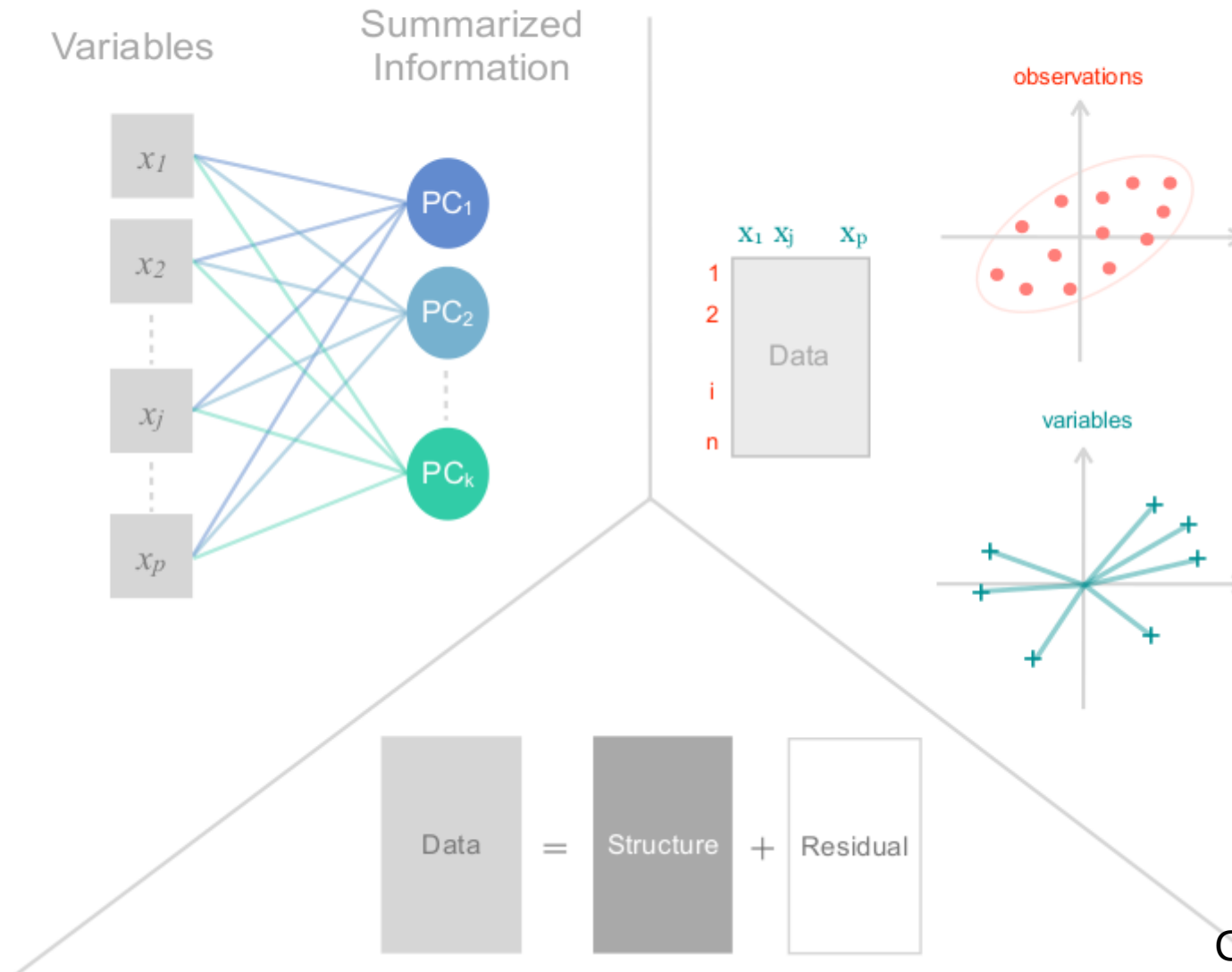
Mercado brasileiro : aplicação de análise de componentes principais no cálculo de VAR para carteiras de renda fixa

O PCA nos permite:

- Um "melhor resumo", as informações importantes contidas em uma tabela de dados.
- Encontrar uma "representação gráfica" da informação essencial contida dentro de um conjunto de dados.
- Para encontrar uma "aproximação ideal" de um conjunto de dados com uma perda mínima de informação.

O PCA nos permite:

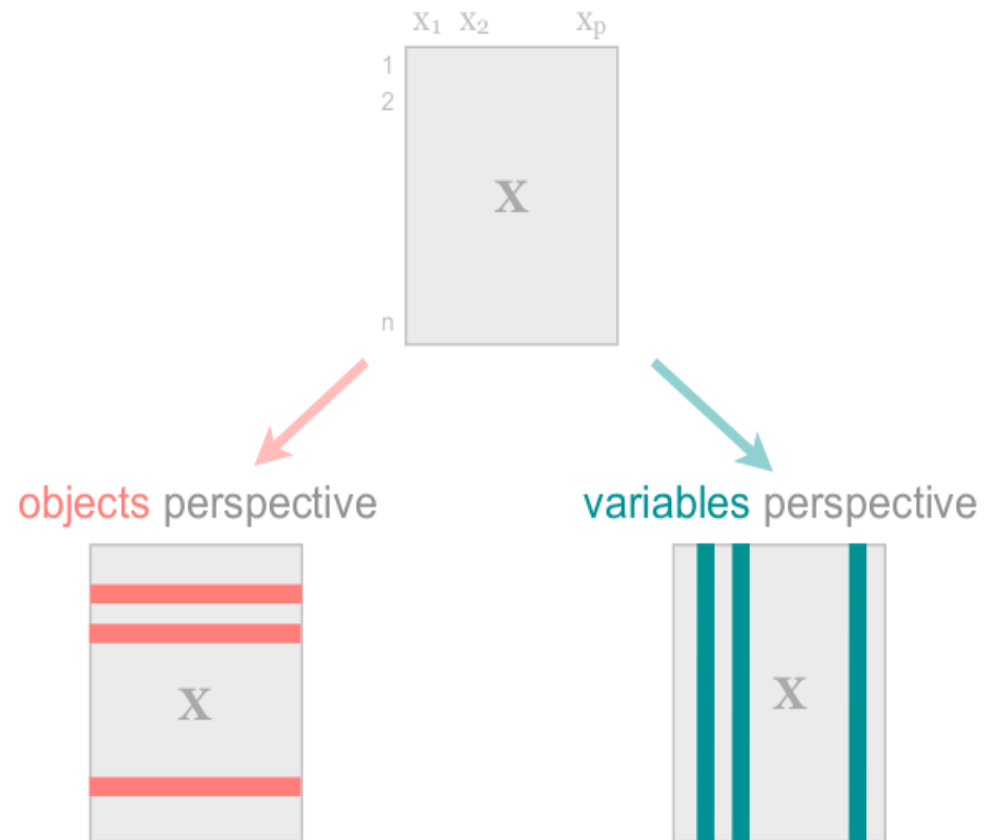
Three Perspectives for PCA



Obs.: Ler Nota

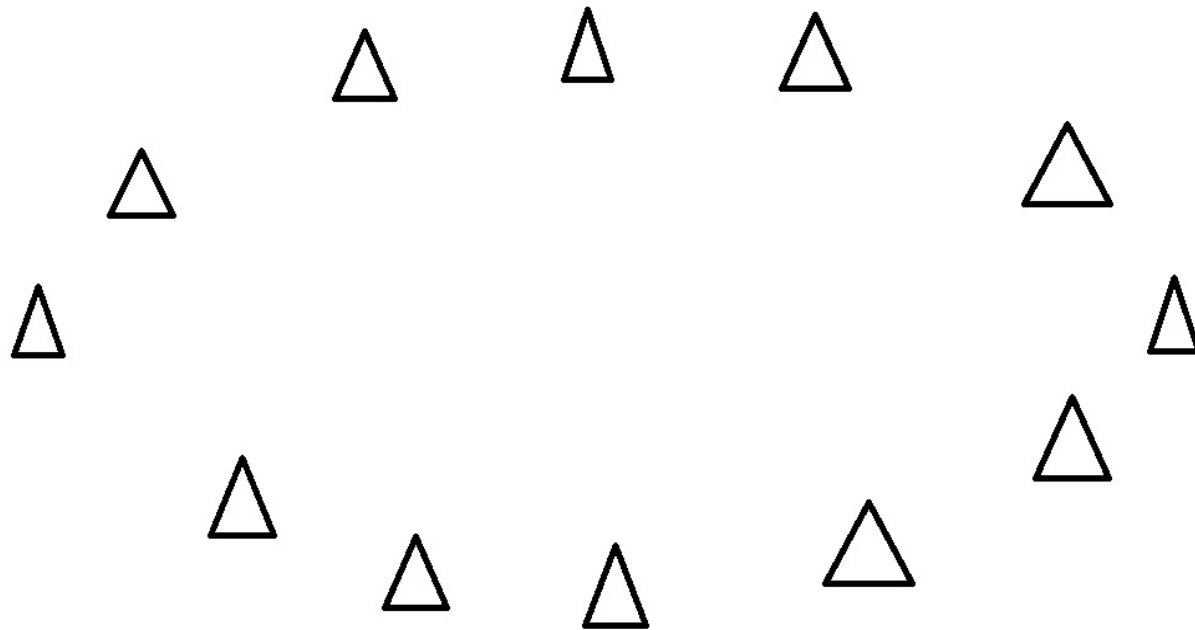
Analizando os dados

- Conjunto de Dados na forma retangular e centrado, ou seja, a soma de suas variáveis tem média ZERO.
- As linhas representam objetos (isto é, observações, os indivíduos, amostras).
- As colunas representam variáveis (ou seja, recursos, características, atributos).



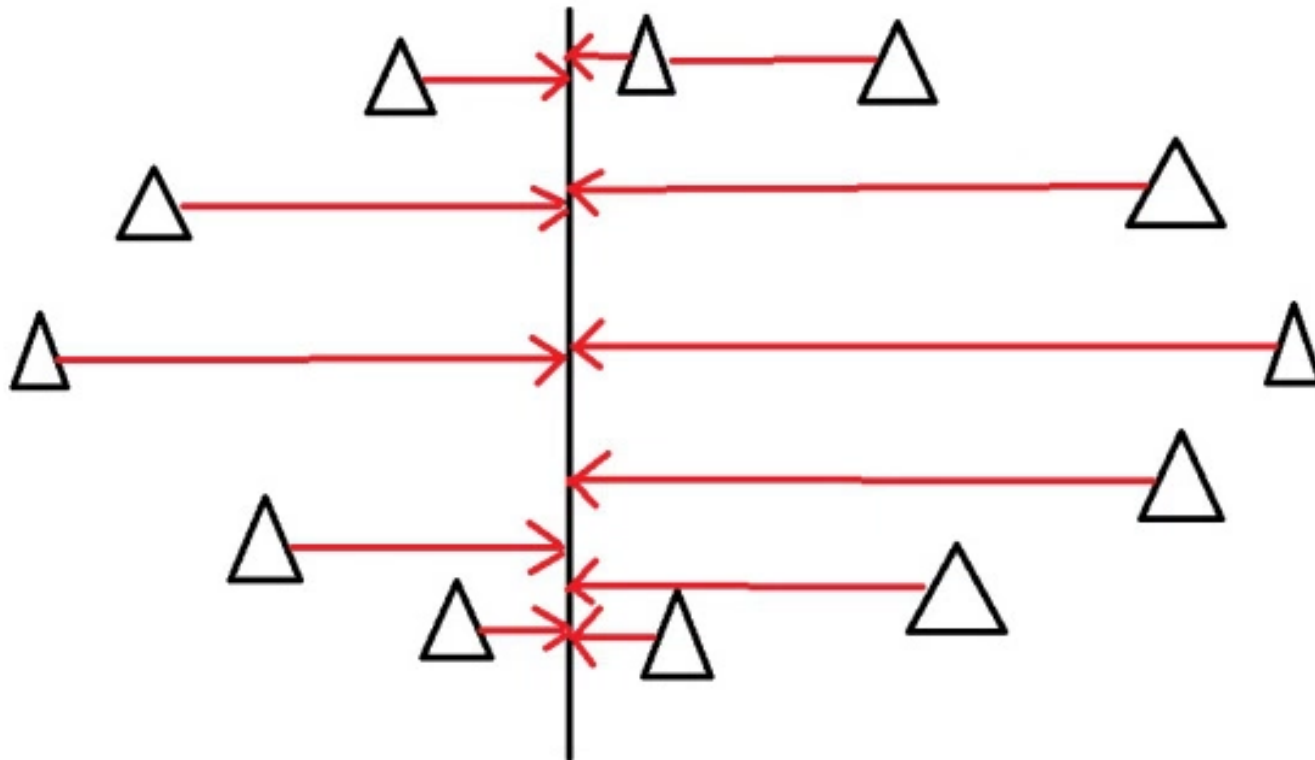
O que é a Análise de Componente Principal?

- Imagine que os triângulos são pontos dos dados. Para encontrar a direção onde há mais variância, encontrar uma linha reta onde os dados estão mais espalhados ou projetados sobre ela.



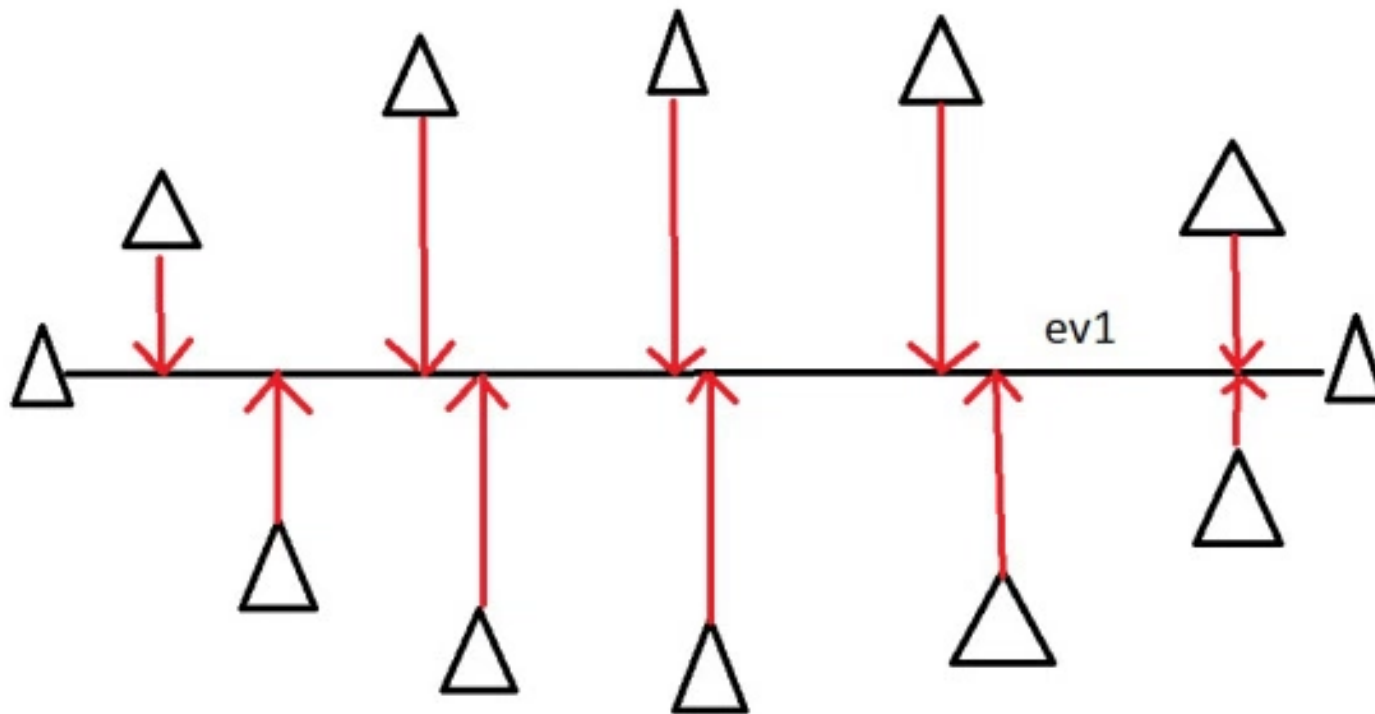
O que é a Análise de Componente Principal?

- Os dados não estão muito espalhados aqui, portanto, ela não tem uma grande variação. Provavelmente não é o componente principal.



O que é a Análise de Componente Principal?

Nesta linha de dados estão de maneira mais espalhados, que tem uma grande variância.



- Felizmente, podemos usar a matemática para encontrar o componente principal, em vez de desenhar linhas e triângulos em forma desigual. Este é o lugar onde autovetores e autovalores irão entrar.

Componentes Principais (Pcs em inglês)

Os PC usa um conjunto de dados representado por uma matriz de n registros por p atributos, que podem estar correlacionados, e *sumariza esse conjunto por **eixos não correlacionados*** (componentes principais) que são uma combinação linear das p variáveis originais

$$PC_1 \longrightarrow Z_1 = w_{11}X_1 + w_{12}X_2 + \cdots + w_{1p}X_p$$

$$PC_2 \longrightarrow Z_2 = w_{21}X_1 + w_{22}X_2 + \cdots + w_{2p}X_p$$

$$\vdots \qquad \qquad \qquad \vdots$$

$$PC_k \longrightarrow Z_k = w_{k1}X_1 + w_{k2}X_2 + \cdots + w_{kp}X_p$$

- Os PCs são obtidas como combinações lineares (ou seja, uma soma ponderada) das variáveis originais.

Componentes Principais (PCs em inglês)

- Para evitar um PCs capturando a mesma variação de outros PCs (isto é, evitando a informação redundante), eles necessitam ser mutuamente ortogonal, para que não estejam correlacionado uns com os outros.
- Olhando para PCs implica que captura a maior parte da variação nos dados, ou seja em Termos estatístico, Queremos obter PCs com variação máxima

Raciocínio geométrico da PCA

- O **centroide** dos pontos é definido pela média de cada atributo
- A **variância** de cada atributo é média dos quadrados da diferença dos n pontos com relação a média de cada atributo

$$V_i = \frac{1}{n-1} \sum_{m=1}^n (X_{im} - \bar{X}_i)^2$$

Raciocínio geométrico da PCA

- Grau com que cada variável é linearmente correlacionado é representado pela sua **covariância**.

The diagram shows the formula for covariance C_{ij} with five red arrows pointing to specific parts of the equation, each with a text label below it:

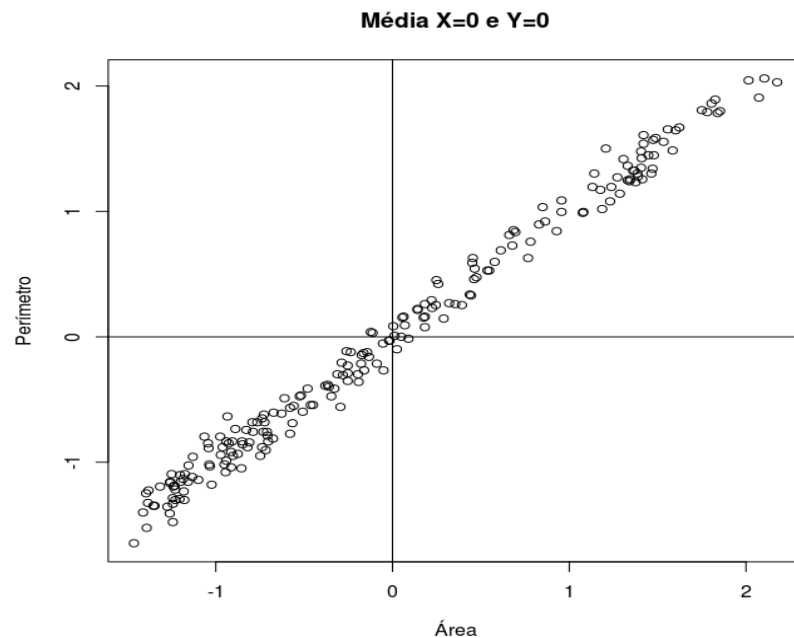
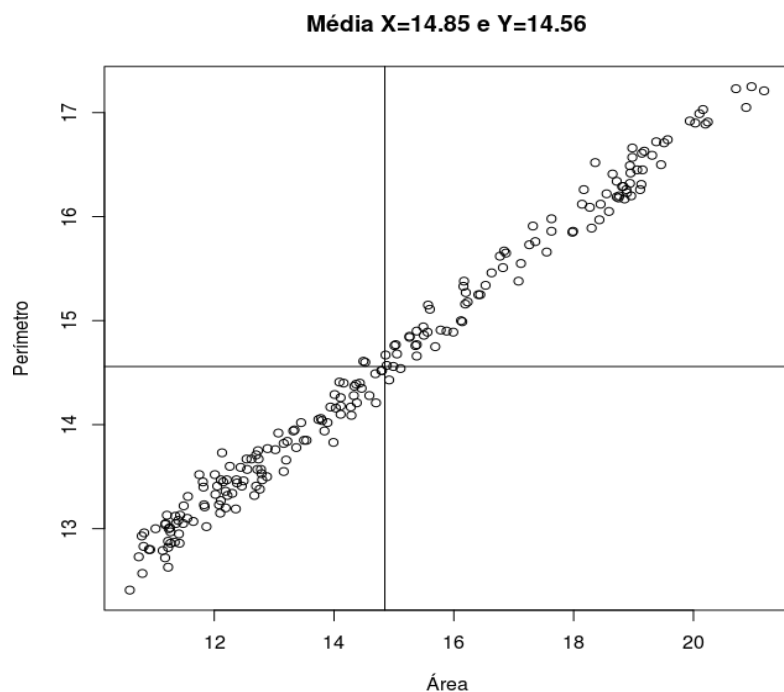
- An arrow points to C_{ij} with the label "Covariance of variables i and j".
- An arrow points to the denominator $n - 1$ with the label "Sum over all n objects".
- An arrow points to X_{im} with the label "Value of variable i in object m".
- An arrow points to \bar{X}_i with the label "Mean of variable i".
- An arrow points to X_{jm} with the label "Value of variable j in object m".
- An arrow points to \bar{X}_j with the label "Mean of variable j".

$$C_{ij} = \frac{1}{n - 1} \sum_{m=1}^n (X_{im} - \bar{X}_i)(X_{jm} - \bar{X}_j)$$

Raciocínio geométrico da PCA

- Caso seu conjunto de dados não seja centrado poderá usar o comando do R, para centralizar, ou deixar as variáveis com média ZERO;

```
> X = scale(Conj.DADOS, center = TRUE, scale = TRUE)
```



Raciocínio geométrico da PCA

- O objetivo da PCA é **rotacionar** rigidamente os eixos desse espaço p -dimensional para novas posições (eixos principais) que tem a seguinte propriedade:
 - Ordenado de tal maneira que o **eixo principal 1 tem a maior variância**, o eixo 2 tem menor variância que o 1º, ..., e o último eixo tem a menor variância
 - Covariância entre cada par de eixos é zero (**os eixos principais não são correlacionados**).

COMO VAMOS ROTACIONAR OS DADOS NAS COMPONENTES PRINCIPAIS?

Como encontrar os Autovalores e Autovetores?

- Em uma Matriz $A_{n \times n}$, podemos encontrar os autovalores λ e autovetores v pela função característica definida como:
 - $p(\lambda) = \det(A - \lambda I)$, onde:
 - $p(\lambda)$ é chamado de polinômio característico de A ;
 - I é a matriz identidade.

Como encontrar os Autovalores e Autovetores?

- Dada a matriz $A = \begin{pmatrix} 4 & 5 \\ 2 & 1 \end{pmatrix}$

- Não foi Normalizado, pois o intuito é só mostrar os cálculos para achar os autovalores e autovetores

- **Cálculo dos autovalores:** $\det(A - \lambda I) = 0$

$$\det(A - \lambda I) = \det\left(\begin{bmatrix} 4 & 5 \\ 2 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = \det\begin{pmatrix} 4 - \lambda & 5 \\ 2 & 1 - \lambda \end{pmatrix}$$

- $\det(A - \lambda I) = 0 \Leftrightarrow (4 - \lambda)(1 - \lambda) - 10 = 0 \Leftrightarrow \lambda^2 - 5\lambda - 6 = 0$
- Os **autovalores** são $\lambda_1 = -1$ e $\lambda_2 = 6$.

Como encontrar os Autovalores e Autovetores?

- Para cada autovalor encontrado, resolvemos o sistema linear $(A - \lambda I) v = 0$.

- Para $\lambda = -1$:

$$\begin{pmatrix} 4 - (-1) & 2 \\ 2 & 1 - (-1) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0$$

- $4x + 2y = 0 \Rightarrow x = -1/2y$
- $2x + 2y = 0 \Rightarrow x = -y$

- Para $\lambda = 6$:

$$\begin{pmatrix} 4 - 6 & 2 \\ 2 & 1 - 6 \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = 0$$

- $-2x_1 + 2y_1 = 0 \Rightarrow x_1 = y_1$
- $2x_1 - 5y_1 = 0 \Rightarrow x_1 = 5/2y_1$

Explicação Algébrica

Teorema 2.2C. Seja A uma matriz $n \times p$. Então $A'A$ e AA' têm as seguintes propriedades:

- (i) $A'A$ é $p \times p$ e é obtida como produto das *colunas* de A .
- (ii) AA' é $n \times n$ e é obtida como produto das *linhas* de A .
- (iii) Ambas as matrizes $A'A$ e AA' são simétricas.
- (iv) Se $A'A = \Phi$ então $A = \Phi$.

- Poderemos dizer que (i) Associação de Variáveis e (ii) Associação de objetos, como elas são simétricas, obedece o Teorema 2.12C

Teorema 2.12C. Seja A ($n \times n$) uma matriz simétrica

- (i) Os autovalores de A são números reais.
- (ii) Os autovetores $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ são mutuamente ortogonais; isto é, $\mathbf{x}_i' \mathbf{x}_j = 0$ para $i \neq j$

- Dada a matriz $A = \begin{pmatrix} 4 & 5 \\ 2 & 1 \end{pmatrix}$

- Centroide

```
> A
      [,1] [,2]
[1,]    4    5
[2,]    2    1
> A=scale(A,scale=F)
> A
      [,1] [,2]
[1,]    1    2
[2,]   -1   -2
```

```
> #Matriz Simétrica
> A.simetrica = t(A) %*% A
> A.simetrica
      [,1] [,2]
[1,]    2    4
[2,]    4    8
```

Decomposição de Matrizes

- A decomposição da matriz é apenas um meio de expressar uma matriz como um produto de duas ou mais matrizes simples.
- Há muitos tipos de decomposições matriciais mas para nossos propósitos compreensão PCA, estamos interessados em duas decomposições:
 - Decomposição de autovalores (EVD em inglês)
 - Decomposição Valor Singular (SVD em inglês)

Explicação Algébrica

Teorema 2.12D. Se \mathbf{A} é uma matriz simétrica com autovalores $\lambda_1, \lambda_2, \dots, \lambda_n$ e autovetores normalizados $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ então \mathbf{A} pode ser expressa como

$$\mathbf{A} = \mathbf{C}\mathbf{D}\mathbf{C}' \quad (2.102)$$

$$= \sum_{i=1}^n \lambda_i \mathbf{x}_i \mathbf{x}_i' \quad (2.103)$$

onde $\mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ e \mathbf{C} é a matriz ortonormal $\mathbf{C} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$. O resultado mostrado em (2.102) ou (2.103) é chamado de *decomposição espectral* de \mathbf{A} .

Ver prova nas págs. 46-47.

Corolário 1. Se \mathbf{A} é uma matriz simétrica e \mathbf{C} e \mathbf{D} são definidas como no Teorema 2.12D, então \mathbf{C} diagonaliza \mathbf{A} , isto é,

$$\mathbf{C}'\mathbf{A}\mathbf{C} = \mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) \quad (2.105)$$

Teorema 2.12E. Se \mathbf{A} é uma matriz com autovalores $\lambda_1, \lambda_2, \dots, \lambda_n$ então

$$(i) \det(\mathbf{A}) = |\mathbf{A}| = \prod_{i=1}^n \lambda_i \quad (2.106)$$

$$(ii) \text{tr}(\mathbf{A}) = \sum_{i=1}^n \lambda_i \quad (2.107)$$

EVD

- O atrativo do EVD é quando aplicado a simétrica
- Uma matriz $M_{n \times n}$, pode ser decomposta em:

- $M = UDU'$

- $U_{n \times p}$, coluna ortonormal que contem os autovetores de M
- $D_{p \times p}$, matriz Diagonal que contém os auto valores de M

EVD

- Função `eigen()`
 - R fornece a função `eigen()`, para realizar uma decomposição autovalor de uma dada matriz
- `eigen` (saída)

Uma lista com os seguintes componentes:

- `Values`: Vector que contém os Autovalores
- `Vectors`: matriz cujas colunas contêm os AutoVetores

EVD

```
#EVD
A = matrix(c(4,2,5,1),nrow=2)
#Matriz Simétrica
A.simetrica = t(A) %*% A
# Decomposição de AutoValores
EVD = eigen(A.simetrica)
#AutoValores
lambda = EVD$values
#AutoVetores
U = EVD$vectors
```

```
> U %*% diag(lambda) %*% t(U)
      [,1] [,2]
[1,]    2    4
[2,]    4    8
> A.simetrica
      [,1] [,2]
[1,]    2    4
[2,]    4    8
```

```
> U
      [,1] [,2]
[1,] 0.4472136 -0.8944272
[2,] 0.8944272  0.4472136
> lambda
[1] 10  0
```

- $U = t(U) \%*\% U$ – Matriz simétrica dos autovetores
- $Y = U \%*\% \text{diag}(\text{lambda}) \%*\% t(U)$ – Y pode ser decomposto na diagonal dos autovalores e na matriz de autovetores

SVD

- Se aplica a qualquer matriz retangular, o que significa que podemos aplicar SVD para qualquer tabela de dados X .
- Uma Matriz $X_{n \times p}$, pode ser decomposta como:

- $X = U\Lambda V$

Onde:

- $U_{n \times p}$ é uma matriz coluna ortonormal contendo os vetores singulares esquerda
- $\Lambda_{p \times p}$ é matriz diagonal contendo os valores singulares de X
- $V_{p \times p}$ é matriz coluna ortonormal contendo os vetores singulares direita

SDV

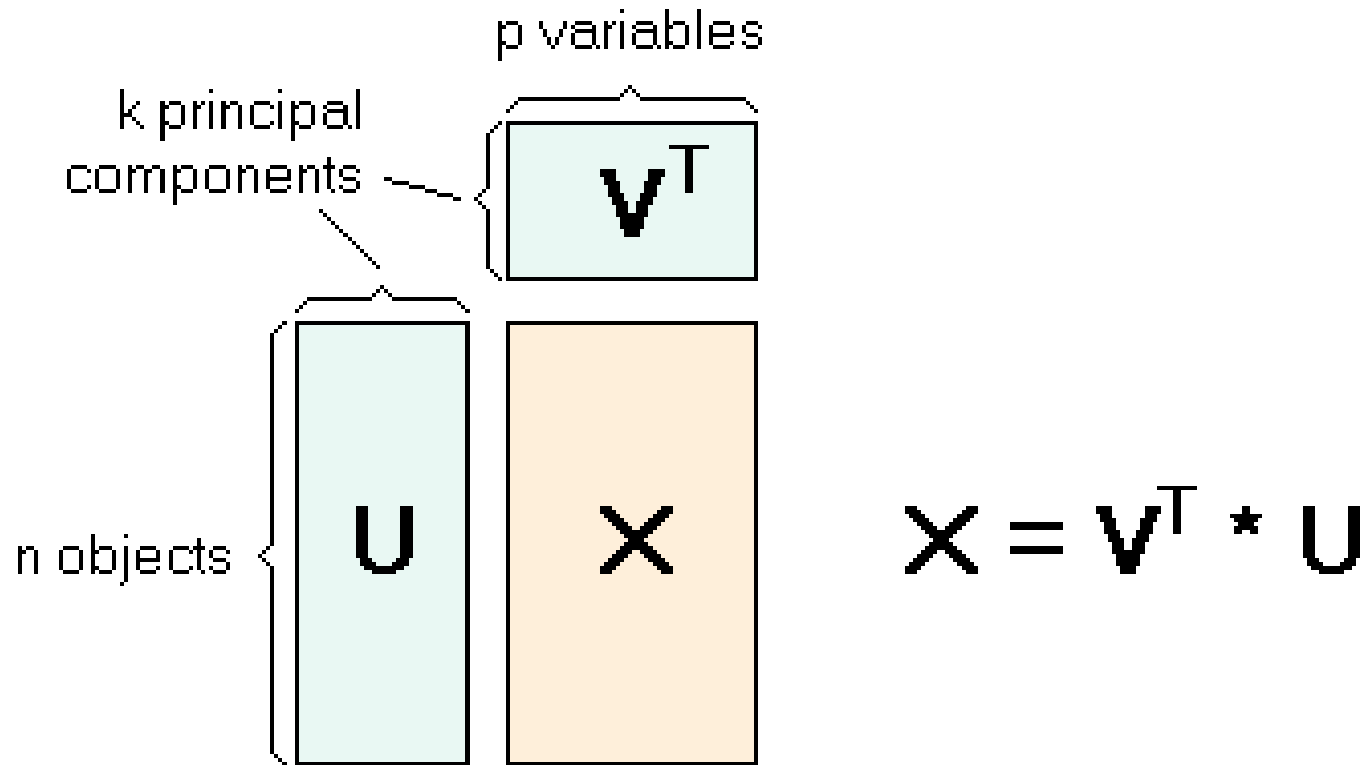
- Valor singular de decomposição (SVD), utilizando a função `svd()`
- A função `SVD(X, nu, nv)`, `nu` e `nv` são calculados automaticamente, são os mínimos da direita e esquerda da matriz
 - `nu` - o número de vetores esquerdos singulares para ser computada. Estes dados devem estar entre 0 e $n = \text{nrow}(x)$.
 - `nv` - o número de vetores direitos singulares para ser computada. Estes dados devem estar entre 0 e $p = \text{Ncol}(x)$.
- Função SVD retorna:
 - `d` - um vector que contém os valores singulares de `X`, de comprimento $\min(n, p)$.
 - `u` - a matriz cujas colunas contém os vetores esquerdo singulares de `x`, se presente `nu > 0`. Dimensão $c(n, nu)$.
 - `v` - a matriz cujas colunas contém os vetores direito singulares de `x`, presente se `nv > 0`. Dimensão $c(p, nv)$.

SDV

SDV

```
set.seed(20)
M = matrix(rnorm(9),3,3)
#Matriz Simétrica
> Y = t(M) %*% M
# Decomposição de valores singulares
SVD = svd(Y)
# Orthornomal de U
t(SDV$u) %*% SDV$u
# Ortonormal de V
t(SDV$v) %*% SDV$v
# Y é igual U D V'
U %*% diag(SDV$d) %*% t(V)
```

Entendendo a matriz de dados



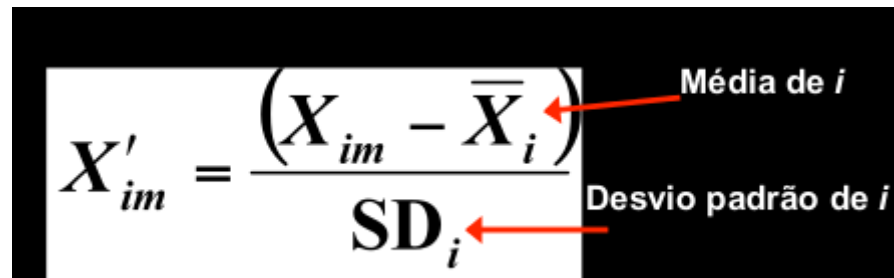
PCA é baseado em uma decomposição da matriz de dados X em duas matrizes ortogonais V e U

Entendendo a matriz de dados

- A matriz V é geralmente chamado de matriz de **Loadings**, podem ser entendidos como os pesos para cada variável original para o cálculo do componente principal.
 - São os Autovetores
- A matriz U é chamado de matriz de **Scores**, contém os dados originais em um sistema de coordenadas rodado.
 - O Produto matricial dos dados originais com os autovetores

Covariância

- Usar covariância entre variáveis somente faz sentido se elas estão representadas na mesma unidade.
- **Variáveis com alta variância vão dominar as componentes principais** (problema)
- Esses problemas são geralmente contornados normalizando os atributos



The diagram shows the formula for z-score normalization:
$$X'_{im} = \frac{(X_{im} - \bar{X}_i)}{SD_i}$$
 The formula is presented on a white background with a black border. Red arrows point from text labels to parts of the formula: one arrow points from "Média de i " to \bar{X}_i , and another arrow points from "Desvio padrão de i " to SD_i .

Correlação

- Covariâncias entre variáveis normalizadas são as correlações.
- Depois da normalização, cada variável tem Variância 1
- Correlações também podem ser calculadas a partir de variâncias e covariâncias:

Correlation between variables i and j → $r_{ij} = \frac{C_{ij}}{\sqrt{V_i V_j}}$

Covariance of variables i and j → C_{ij}

Variance of variable i → V_i

Variance of variable j → V_j

Calculando os PCs pelo EVD

```
pca_evd <- function(dataset, center = TRUE, scale = TRUE) {  
  # Media Central e Normalização  
  X = scale(dataset, center = center, scale = scale)  
  # Decomposição de AutoValores  
  if (nrow(X) >= ncol(X)) {  
    EVD = eigen(t(X) %*% X)  
  } else {  
    EVD = eigen(X %*% t(X))  
  }  
  # scores  
  scores = X %*% EVD$vectors  
  rownames(scores) = rownames(dataset)  
  # loadings  
  loadings = EVD$vectors  
  rownames(loadings) = colnames(dataset)  
  # results  
  list(  
    values = EVD$values / (nrow(X) - 1),  
    scores = scores,  
    loadings = loadings  
  )  
}
```

Calculando os Pcs pelo SVD

```
pca_svd <- function(dataset, center = TRUE, scale = TRUE) {  
  # Media central e normalização  
  X = scale(dataset, center = center, scale = scale)  
  # Valor singular de decomposição  
  SVD = svd(X)  
  # scores  
  scores = SVD$u %*% diag(SVD$d)  
  rownames(scores) = rownames(dataset)  
  # loadings  
  loadings = SVD$v  
  rownames(loadings) = colnames(dataset)  
  # Resultado - Multilicação de lagrangian  
  list(  
    values = SVD$d^2 / (nrow(X) - 1),  
    scores = scores,  
    loadings = loadings  
  )  
}
```

Redução de Dimensionalidade

- Escolher as componentes e formar o vetor de características
 - O autovetor associado ao maior autovalor é a componente principal mais relevante. Quanto maior o autovalor associado, maior é a importância do autovetor (componente). Assim, a redução de dimensionalidade ocorre ao retirar as componentes menos significantes.

Construindo o Vetor Característica (FeatureVector)

- O novo conjunto de dados é dado multiplicando a transposta do vetor de características pela transposta dos dados ajustados
- $\text{FinalData} = t(\text{FeatureVector}) \times (\text{DataAdjust})$

```
> f=t(U[,1])%*%A.simetrica
> f
      [,1]      [,2]
[1,] 4.472136 8.944272
```

Conjuntos de dados Sementes

- O grupo examinado consiste em três diferentes variedades de trigo: Kama, Rosa and Canadian
 - 70 elementos cada, selecionados aleatoriamente para o experimento.
- Alta qualidade na visualização da estrutura interna do núcleo foi detectada utilizando um suave técnica de Raio-X.
 - É não-destrutivo e consideravelmente mais barato do que outras técnicas de imagem mais sofisticados, como microscopia de varredura ou tecnologia laser.
- As imagens foram gravadas em 13x18 cm placas KODAK raios-X. Os estudos foram realizados utilizando combinar grãos de trigo colhido proveniente de campos experimentais, exploradas no Instituto de agrofísica da Academia Polonesa de Ciências, em Lublin.
- <http://mlr.cs.umass.edu/ml/datasets/seeds>

Conjuntos de dados Sementes

Informações de Atributo:

Para construir os dados, foram medidos sete parâmetros geométricos de grãos de trigo:

1. Área A,
2. Perímetro P
3. Compacidade $C = 4 * \pi * A / P^2$,
4. Comprimento de kernel (Núcleo),
5. largura do kernel (Núcleo),
6. coeficiente assimétrica
7. comprimento do sulco kernel (Núcleo).

Todos esses parâmetros contínua de valor real.

Conjuntos de dados Sementes

```
> seed=read.table("seeds_dataset.txt",sep="\t",row.names=NULL,header=T)
> head(seed,3)
  Área Perímetro Compactação Comp.Nucl Larg.Nucl Coef.assime Comp.Nucl.Sulco
1 15.26    14.84    0.8710    5.763    3.312    2.221    5.220
2 14.88    14.57    0.8811    5.554    3.333    1.018    4.956
3 14.29    14.09    0.9050    5.291    3.337    2.699    4.825
Vari.Trigo
1      1
2      1
3      1
```

```
> seed.center=scale(seed[,1:7],center=T,scale=T)
> head(seed.center,3)
      Área      Perímetro  Compactação  Comp.Nucl  Larg.Nucl  Coef.assime
[1,]  0.14175904  0.214948819 6.045733e-05  0.3034930  0.1413640  -0.9838010
[2,]  0.01116136  0.008204153 4.274938e-01 -0.1682227  0.1969616  -1.7839036
[3,] -0.19160873 -0.359341919 1.438945e+00 -0.7618171  0.2075516  -0.6658882
Comp.Nucl.Sulco
[1,] -0.3826631
[2,] -0.9198156
[3,] -1.1863572
```


Conjuntos de datos Sementes

```
[1] 71
> seed.pca=prcomp(seed.center)
> summary(seed.pca)
Importance of components:
```

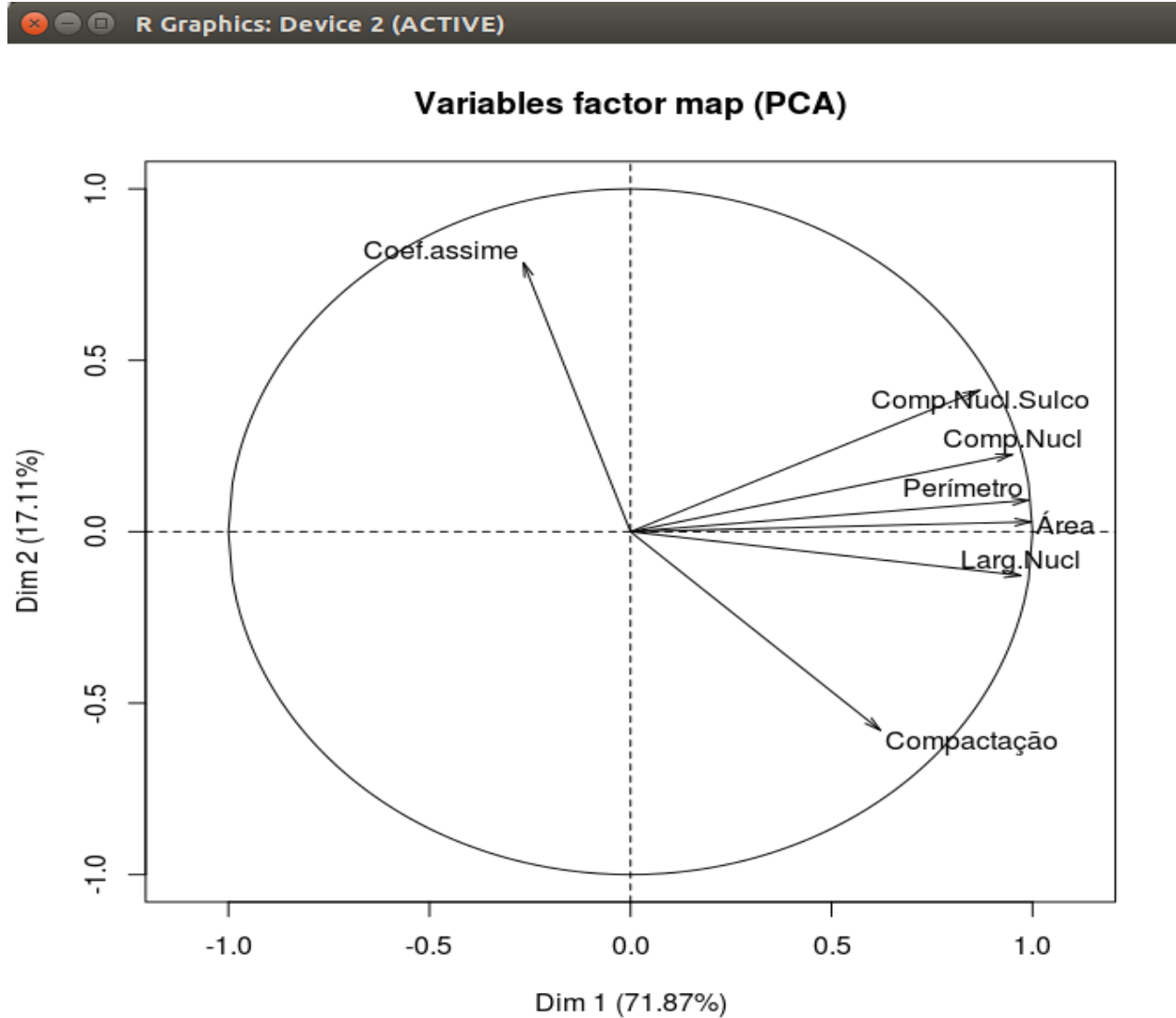
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.2430	1.0943	0.82341	0.26147	0.13680	0.07302	0.02850
Proportion of Variance	0.7187	0.1711	0.09686	0.00977	0.00267	0.00076	0.00012
Cumulative Proportion	0.7187	0.8898	0.98668	0.99645	0.99912	0.99988	1.00000

```
> seed.pca$sdev^2
[1] 5.0312011860 1.1975728470 0.6780034386 0.0683644770 0.0187136090
[6] 0.0053320457 0.0008123968
> seed.pca$sdev[1]^2/sum(seed.pca$sdev^2)
[1] 0.718743
```

Conjuntos de dados Sementes

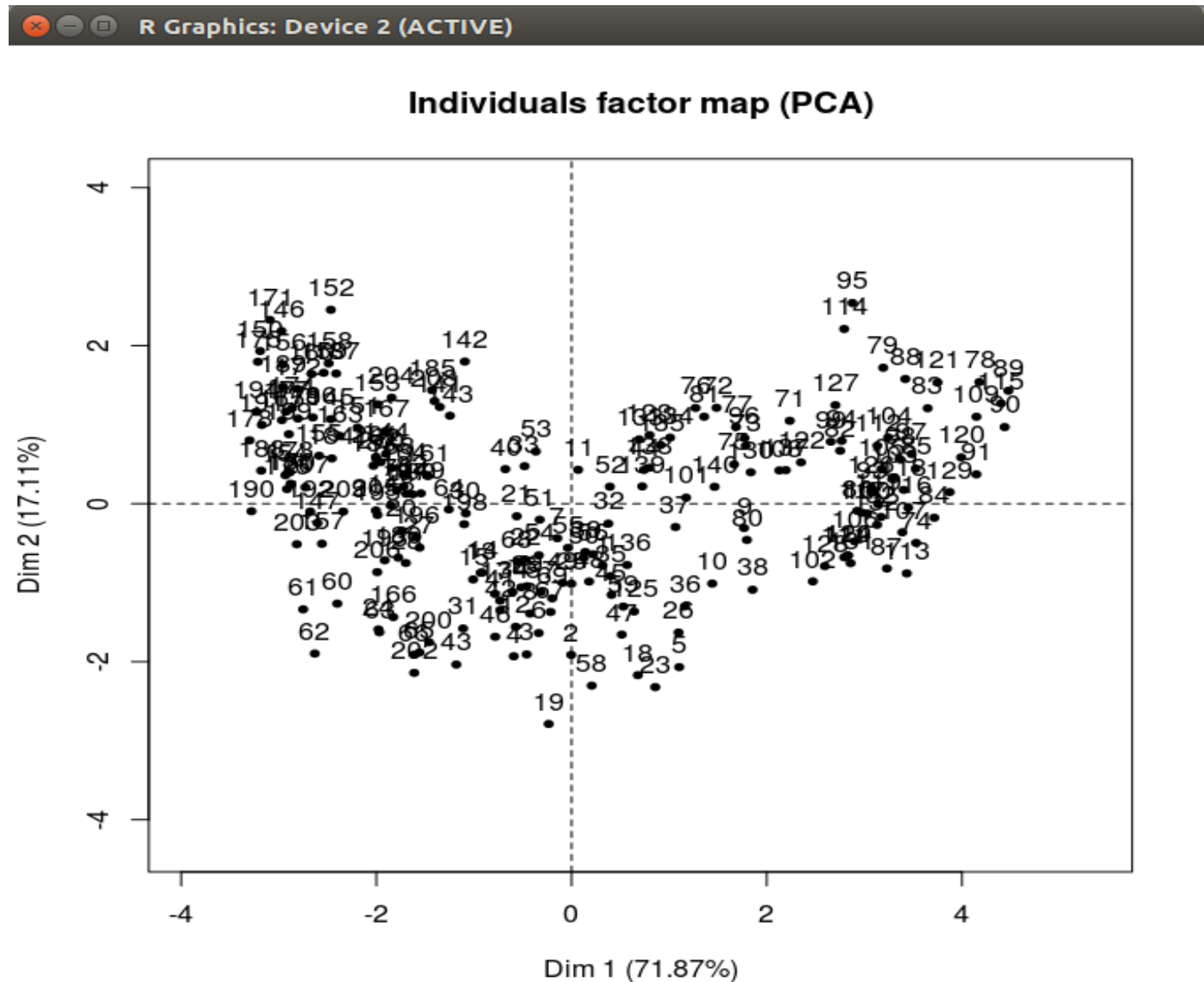
Quanto mais perto uma seta é a circunferência do círculo, melhor a sua representação nos eixos indicados.

Além disso, observe como as variáveis são agrupadas.



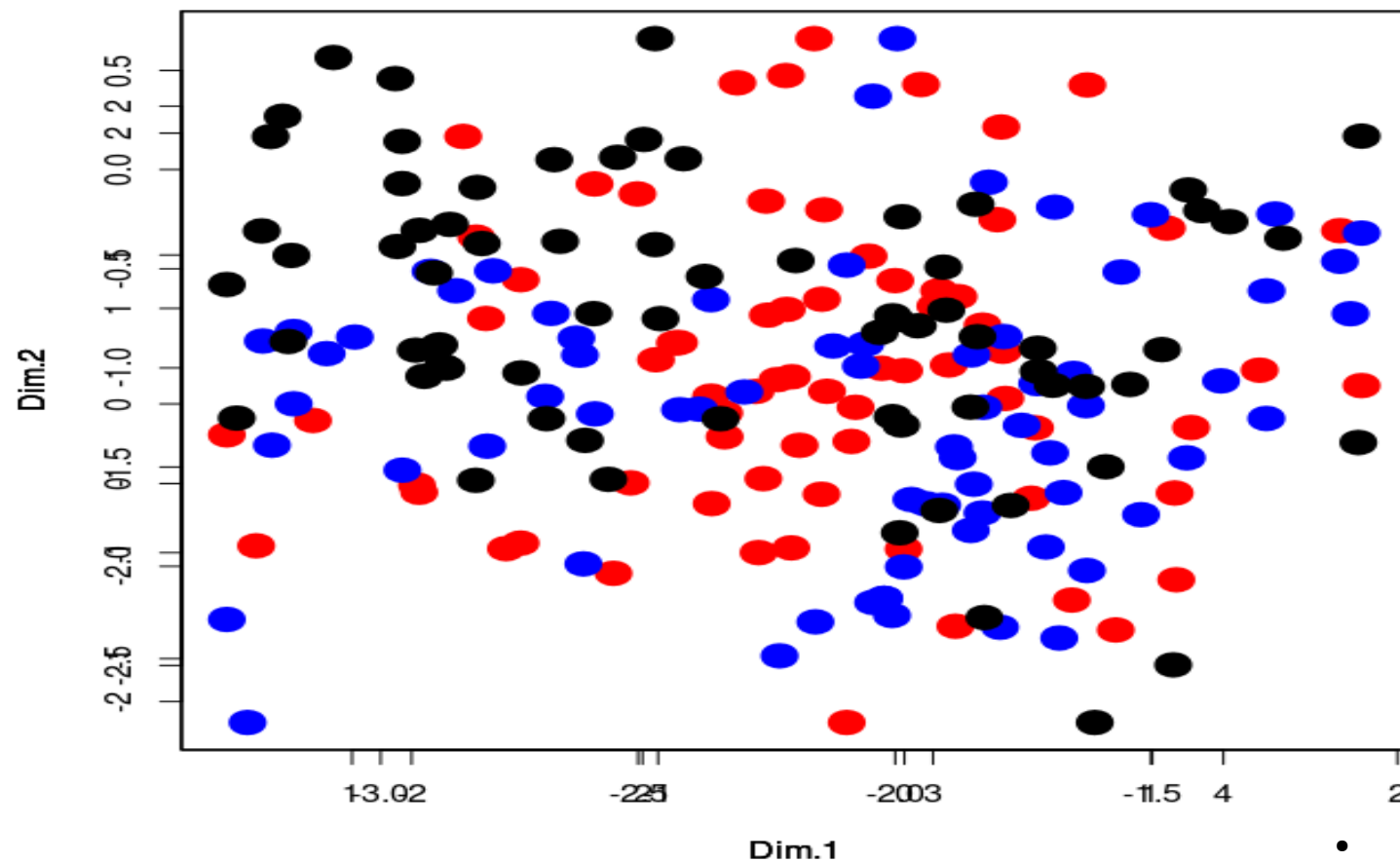
Conjuntos de dados Sementes

- Observações próximas no gráfico, tem comportamento semelhante.



Conjuntos de dados Sementes

R Graphics: Device 2 (ACTIVE)



- Vermelho: Kama
- Azul: Rosa
- Preto: Canadian

Referencia

- Gaston Sanchez, PCA.
- Santo, Rafael do E. - Principal Component Analysis applied to digital image compression
- Matone, Ricardo - Mercado brasileiro : aplicação de análise de componentes principais no cálculo de VAR para carteiras de renda fixa
- http://www.ime.unicamp.br/~wanderson/Aulas/M T803_Aula3_Reducacao_Sintetizacao_Dados.pdf
- <https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvec-tors-eigenvalues-and-dimension-reduction/>

Licença

- Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License
- <http://creativecommons.org/licenses/by-nc-sa/4.0/>
- <https://br.creativecommons.org/>

