Introdução à Análise Multivariada usando R

Marcio Nicolau

2016-10-24

Contents

Chapter 1

Pré-requisitos

Antes de iniciar o curso, faz-se nescessário instalar as ferramentas e pacotes para uso.

Será utilizado o sofware *R version 3.3.1 (2016-06-21)* disponível em CRAN e o *RStudio v1.0.44 Preview* disponível em RStudio.

A carga horária total será de 20 horas (3 dias) nos quais os seguintes tópicos de Análise Multivariada serão abordados.

- Análise de Agrupamentos (CA)
- Análise de Componentes Principais (PCA)
- Análise Fatorial (FA)

Alguns exemplos serão realizados com dados disponíveis na literatura científica e, em certos momentos, serão utilizados dados dos próprios participantes.

Espera-se que ao final do curso o aluno seja capaz de entender o uso de cada técnica e aplicá-la de forma correta em seu campo/área de pesquisa.

Cabe lembrar que este é um curso introdutório e que de forma alguma os conteúdos e aplicações serão apresentadas de forma exaustiva.

Marcio Nicolau

Estatístico / Embrapa Trigo

Palmas/TO, Outubro de 2016

Chapter 2

Introdução

As técnicas de Análise Multivariadas oferecem aplicações em diversas áreas do conhecimento no desenvolvimento científico.

Geralmente são utilizadas em fase exploratória de análise de dados, onde se busca entender melhor as relações entre as variáveis (medições físicas ou observações) de certo evento sob estudo ou de interesse científico.

Há também aplicações em conjunto com outros métodos da estatística onde é possível obter validações ou testes de carácter conclusivo.

Durante este curso e, certamente limitados pelo tempo, serão abordados somente as técnicas de caractere exploratório com a finalidade de melhor explicar as relações intrínsecas entre os dados, reduzir a dimensão, entender fontes de variabilidade, criar grupos homogêneos de indivíduos/espécies.

Pode-se dizer que a Análise Fatorial (FA), a Análise de Componentes Principais (PCA) e a Análise de Cluster (CA) são processo que tem por objetivo reduzir a complexidade dos dados observados, bem como entender o modelo estrutural presente nos dados.

No caso do FA, o objetivo é o de identificar construções poucos constructos para explicar os dados observados. No caso de PCA, pode não ser simples redução de dimensão, mas a interpretação dos componentes.

Por fim, a Análise de Cluster (CA) pode também ser usada para criar grupos de variáveis com interesse de reduzir a complexidade dos dados por meio da formação de grupos menores e homogêneos.

Tecnicamente, o problema de redução de dados pode ser resolvido como uma decomposição do valor singular (SVD) da matriz original, embora a solução mais típica seja o uso de PCA nas matrizes de covariância e/ou correlação.

Chapter 3

Análise de Agrupamentos

Nesta seção serão utilizados as seguinte bibliotecas do R.

```
libs <- c('cluster', 'psych')
sapply(libs, require, character.only = TRUE)
```

```
## Loading required package: cluster
```

```
## Loading required package: psych
```

```
## cluster   psych
##    TRUE    TRUE
```

```
knitr::opts_knit$set(fig.width=5, fig.height=5, fig.align='center')
```

3.1 Processo de Agrupamento

Um agrupamento pode ser construído de duas formas:

- hierarquia: funções *agnes*, *diana*, *mona* e *hclust*;
- particionamento: funções *pam*, *clara*, *fanny* e *kmeans*

3.2 Métodos de Agrupamento

Um agrupamento pode gerar os grupos utilizando algum dos métodos a seguir (mais comuns):

- average: *média* ou UPGMA (média dissimilaridade)
- single: *simple:* (vizinho mais próximo)
- complete: *completa* (vizinho mais distante)
- ward: *Ward* ou método da mínima variância.
- weighted/mcquitty: *média ponderada* ou WPGMA

3.3 Distâncias usadas para cálculo de agrupamentos

Para se calcular a distância entre os componentes, pode-se utilizar as funções a seguir (mais comuns):

- euclidian: *euclideana*, raiz da soma do quadrado das diferenças entre os pontos/observações (distância no plano cartesiano)
- mahalanobis: *Mahalanobis*, distância de cada valor em relação à média e covariância (também conhecida como distância estatística). *OBS* é capaz de trabalhar a distância para observações com repetições. (kernel da normal multivariada)
- manhattan: *Manhattan*, soma das diferenças média absoluta (L1 norm)
- maximum: *Máxima*, máxima distância entre dois componentes (supremum norm)
- canberra: *Canberra*, uso em valores não negativos (p.ex. contagem) $\sum(|x_i - y_i|/|x_i + y_i|)$
- binary: *Binária*, para valores do tipo "on"/"off" em que 0 representa desligado e números maiores que 0. A distância é a proporção de "on's".
- minkowski: *Minkowski*, P-Norm ou p-ésima raíz da soma de potência p das diferenças.

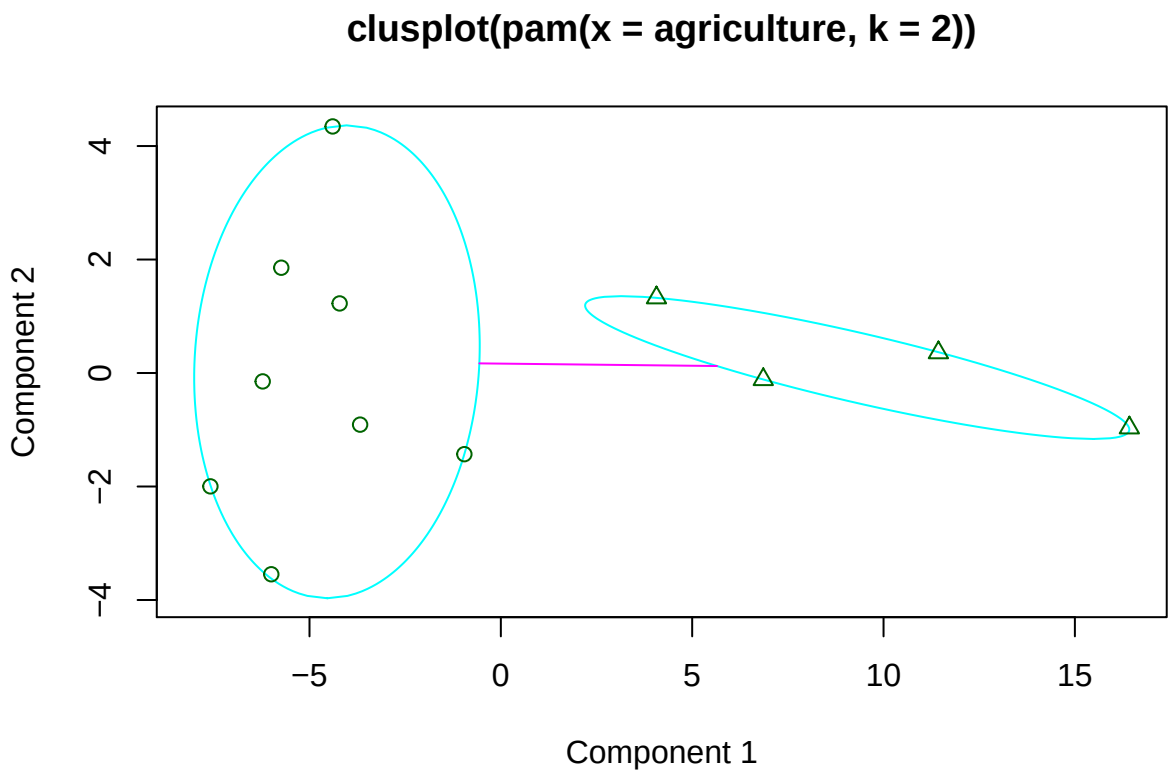
3.4 Exemplo: Força de trabalho agrícola na UE (1993)

Estes conjunto registra os dados da produção per capita e o percentual da população que trabalha na agricultura em cada país da UE em 1993.

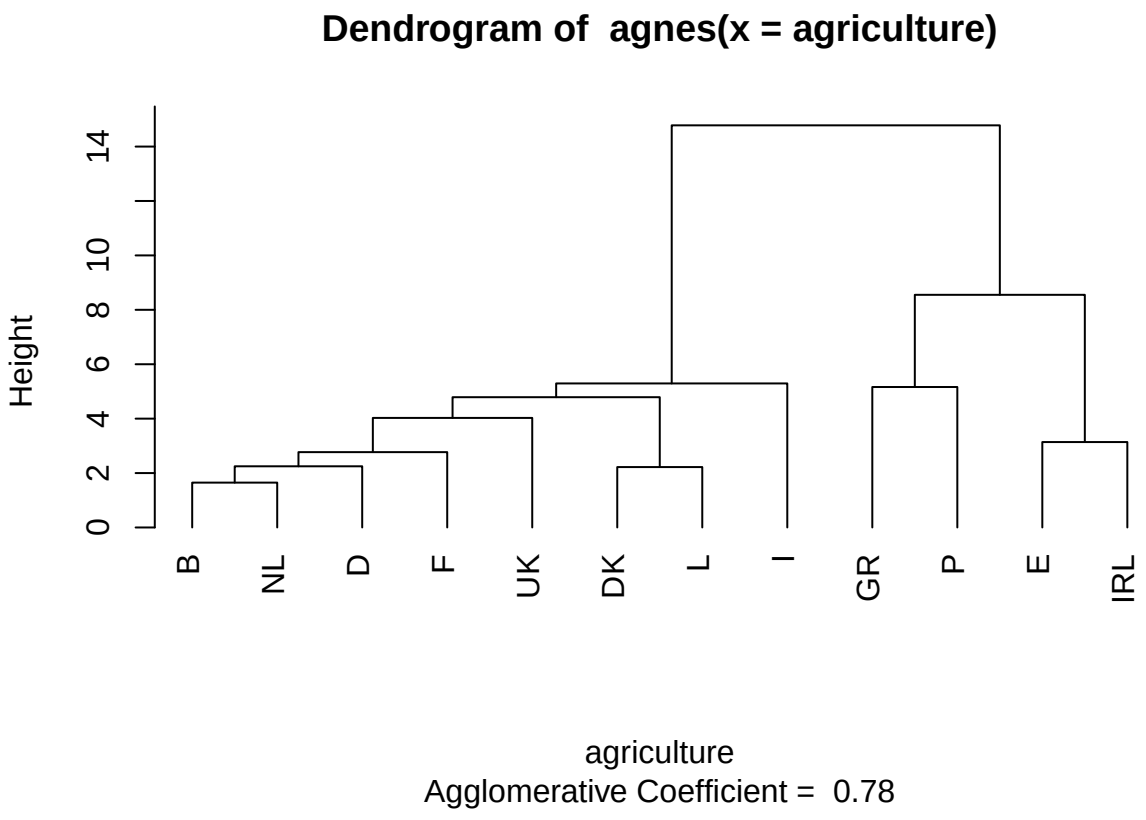
```
data(agriculture)

## Calcula matriz de dissimilaridade usando distância euclidiana
## e sem padronização das variáveis
print(daisy(agriculture, metric = "euclidean", stand = FALSE),
      digits = 2)
```

```
## Dissimilarities :
##        B   DK    D   GR    E    F  IRL    I    L   NL    P
## DK   5.4
## D    2.1  3.4
## GR  22.3 22.6 22.7
## E    9.8 11.2 10.4 12.6
## F    3.4  3.5  2.7 20.1  8.1
## IRL 12.7 13.3 13.1  9.6  3.1 10.6
## I    5.8  5.5  5.4 17.4  5.7  2.8  7.9
## L    4.3  2.2  2.3 24.0 12.1  4.1 14.6  6.7
## NL   1.6  5.1  2.4 20.8  8.3  2.2 11.2  4.2  4.7
## P   17.2 17.9 17.7  5.2  7.4 15.2  4.6 12.5 19.2 15.7
## UK   2.8  8.1  4.9 21.5  9.0  5.3 12.1  6.7  7.1  3.1 16.3
##
## Metric :  euclidean
## Number of objects : 12
```

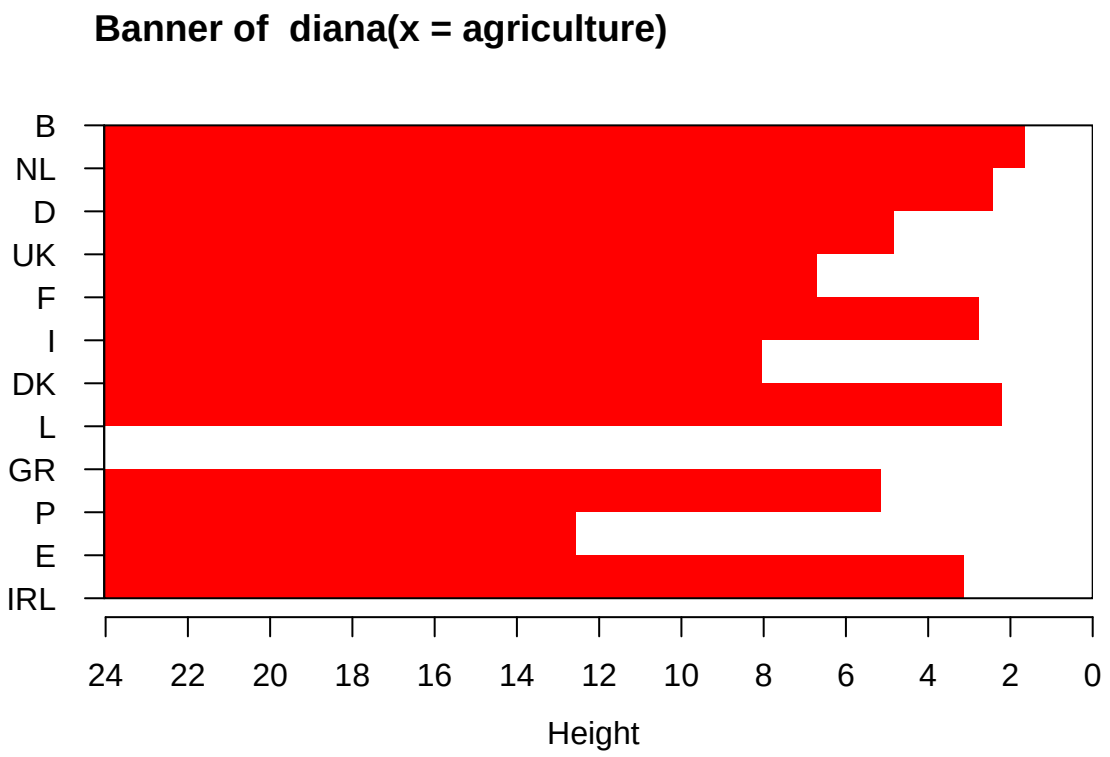
```
## Usa método de particionamento pelo meióide
## Partitioning Around Medoids (PAM)
plot(pam(agriculture, 2), which.plots = 1)
```



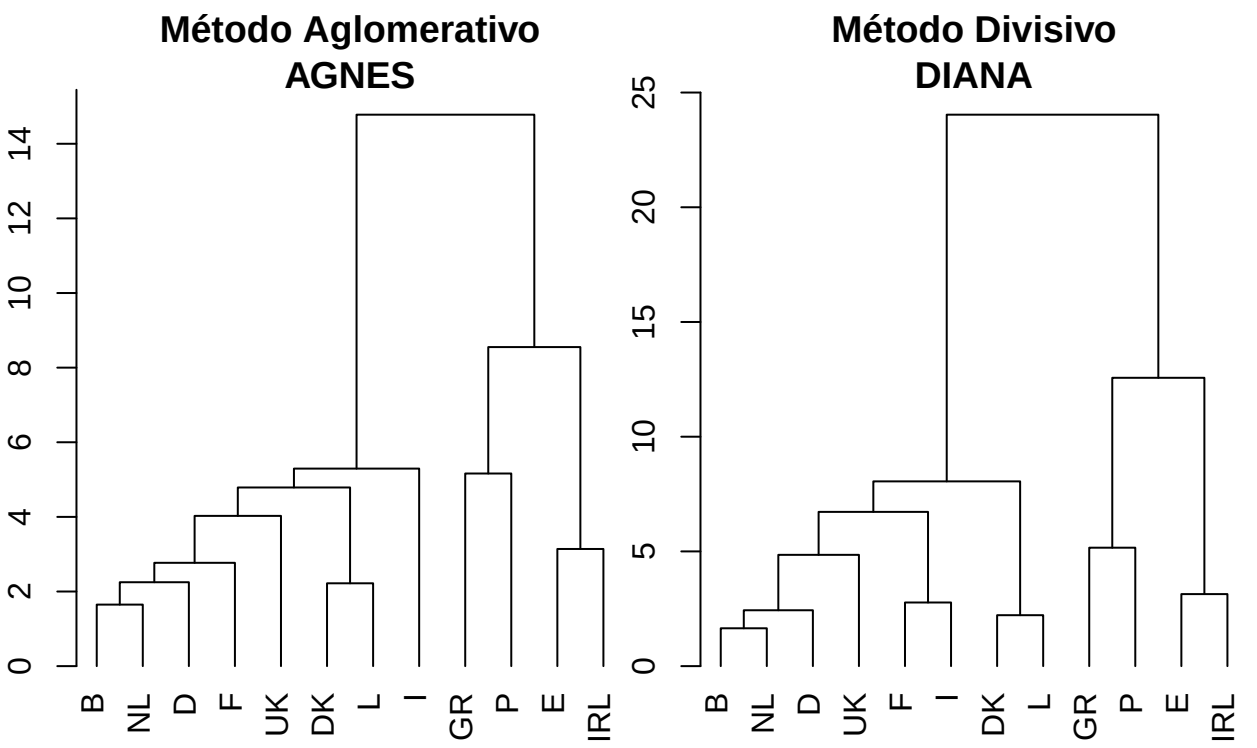

```
## Gráfico dendograma usando método aglomeração mais próximo
## agnes
plot(agnes(agriculture), which.plots = 2, hang = -1)
```




```
## Plot dissimilaridade usando método divisivo
## diana
plot(diana(agriculture), which.plots = 1)
```



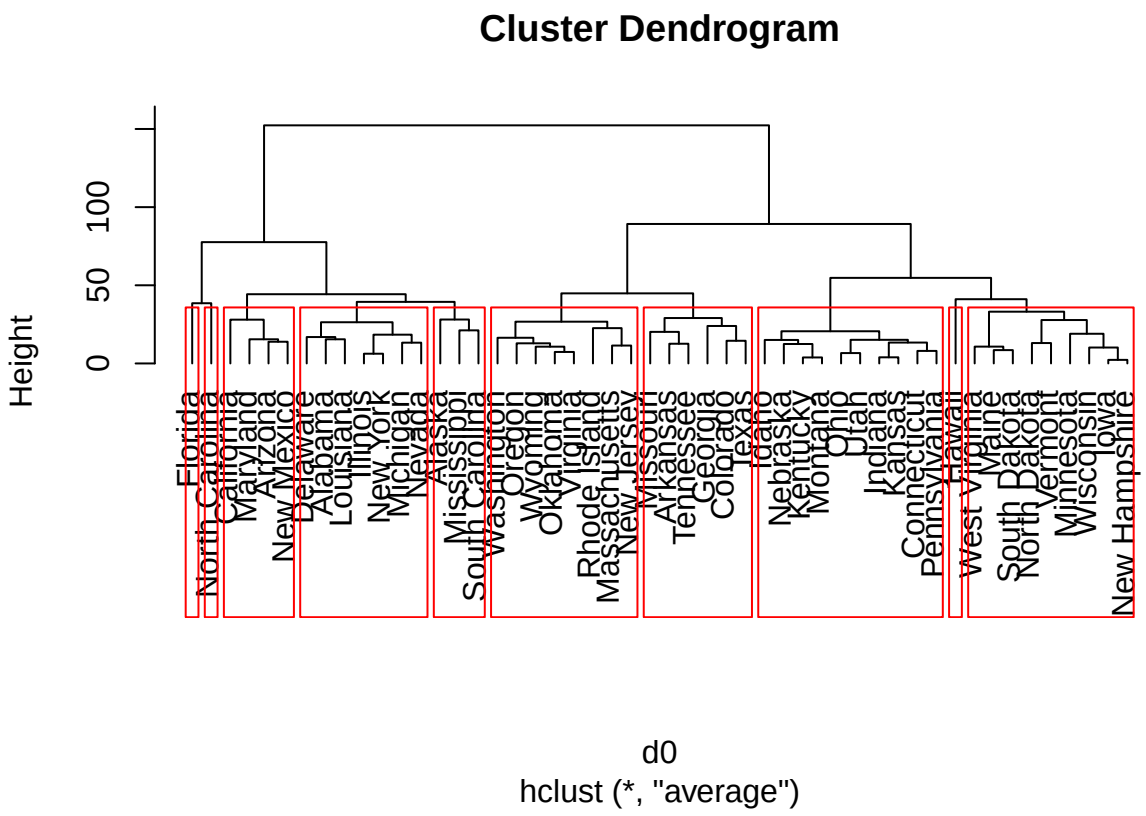

```
## Usando agnes e diana para conjunto agricultura
par(mfrow=c(1,2), mar=c(3,2,2,0))
plot(agnes(agriculture), which.plots = 2, hang = -1,
     main = "Método Aglomerativo\nAGNES")
plot(diana(agriculture), which.plots = 2, hang = -1,
     main = "Método Divisivo\nDIANA")
```



3.5 Exemplo cluster USArrest

Neste exemplo vamos utilizar o conjunto USArrest, disponível na instalação padrão do R.

```
data(USArrests)
#?USArrests

d0 = dist(USArrests) # euclidian
hc = hclust(d0, "average")
plot(hc, hang = -1)
# Criar 10 grupos
memb <- cutree(hc, k = 10)
# Anota no gráfico os 10 grupos
rect.hclust(hc, 10)
```




```
dFinal=data.frame(State=row.names(USArrests), grp = memb)
subset(dFinal, grp == 1)
```

```
##                State grp
## Alabama      Alabama   1
## Delaware   Delaware   1
## Illinois   Illinois   1
## Louisiana Louisiana   1
## Michigan   Michigan   1
## Nevada       Nevada   1
## New York   New York   1
```

Chapter 4

Análise de Componentes Principais

É uma alternativa à Análise Fatorial (FA), apesar dos objetivos serem semelhantes (PCA e FA), na PCA se busca obter o modelo descritivo dos dados enquanto na FA se busca o modelo estrutural.

Outro destaque importante é que a matriz/vetor de cargas *"loadings"* possuírem valores equivalentes, na FA estes são menores. Isto ocorre porque na PCA é ajustado um modelo para a variância completa da matriz de correlação das variáveis e na FA o processo é realizado somente para a variância comum.

Chapter 5

Análise Fatorial

Some *significant* applications are demonstrated in this chapter.

5.1 Example one

5.2 Example two

Bibliography