



Universidade Federal de Goiás
Escola de Veterinária e Zootecnia

R-Fácil

Análise Multivariada

Apostila destinada a usuários do R, com demonstrações de uso de funções em exemplos da área de Ciências Agrárias.

Goiânia
Mai de 2014

Apresentação

Atualmente encontram-se grande número de apostilas e livros abordando o uso do software R (R Core Team, 2014) para análises estatísticas. No entanto, são raros os materiais didáticos aplicados à área de ciências agrárias. Visando estimular o aprendizado com linguagem mais simples e aplicada, com exemplos práticos voltados, principalmente, as áreas de ciências agrárias e biológicas, o presente material foi desenvolvido para auxiliar a utilização do R na realização de análises multivariadas. O público alvo são alunos de graduação, pós-graduação, professores e pesquisadores. O objetivo desta apostila não é aprofundar em aspectos teóricos, mas apenas apresentar um tutorial de análise de dados utilizando pacotes e funções de forma bastante prática.

A apostila **“R-fácil: Análise Multivariada”** faz parte de uma série de quatro apostilas que contemplam parte do conteúdo do site <http://r-facil.webnode.com/>. Deve-se destacar que este material utiliza-se de funções de uso mais prático para a demonstração das análises e por isso o título R-fácil, com intuito de descomplicar um pouco a utilização do software.

No texto as discussões estão na fonte Times 12 e as análises realizadas no R em fonte Inconsolata em negrito, sendo a programação em azul e os resultados em preto. Boa leitura e não deixe de conferir os demais materiais da série R-fácil (download em <http://r-facil.webnode.com/>).

O autor

Índice

	página
Análise multivariada	1
Medidas de dissimilaridade	1
Medidas de dissimilaridade quantitativas	1
Medidas de dissimilaridade qualitativas (binárias)	3
Dendrogramas	4
Componentes principais (PCA)	11
Variáveis canônicas	18
Correlações canônicas	21
Análise discriminante	24
Referências	28

Análise multivariada

A análise multivariada consiste em técnicas estatísticas que permitem analisar várias variáveis de forma conjunta. Em alguns casos, além de uma análise variável por variável, existe interesse em uma comparação entre grupos ou elementos amostrais de forma mais abrangente, reunindo informação de um conjunto de variáveis consideradas simultaneamente. Assim, qualquer análise estatística onde são consideradas duas ou mais variáveis para obter uma informação conjunta das mesmas, pode-se classificar como análise multivariada. Neste material iremos abordar algumas destas técnicas estatísticas mais utilizadas, como as medidas de dissimilaridade, agrupamentos em forma de árvore (dendrogramas), análise de componentes principais, variáveis canônicas, correlações canônicas e análise discriminante. Iniciaremos a seguir com as medidas de dissimilaridade.

Medidas de dissimilaridade

A palavra dissimilaridade significa diferença. Então, uma medida de dissimilaridade significa uma medida de diferença. No caso de uma pesquisa, pode-se haver o interesse em obter uma medida da diferença (medida de dissimilaridade) entre elementos amostrais (animais, plantas, ecossistemas, etc) para um conjunto de variáveis. As variáveis consideradas podem ser quantitativas e qualitativas. A seguir abordaremos alguns métodos para obtenção de medidas multivariadas de dissimilaridade.

Medidas de dissimilaridade quantitativas

Será abordada aqui a medida de dissimilaridade mais comumente utilizada, a distância euclidiana. A distância euclidiana entre dois elementos amostrais (animais, plantas, etc) é obtida pela raiz quadrada do somatório da diferença ao quadrado, do valor de cada variável considerada na análise. Considerando dois indivíduos (elementos) P1 e P2 pode-se representar a distância euclidiana na Equação 1, onde são consideradas duas variáveis x e y.

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Equação 1: Representação matemática da distância euclidiana entre dois elementos amostrais (P1 e P2), considerando duas variáveis x e y

Para exemplificar o uso da distância euclidiana como medida de dissimilaridade quantitativa, vamos considerar as medidas de comprimento dos dígitos de 15 animais nelore, 15 curraleiros, 15 pantaneiros e 12 búfalos (dados parciais da dissertação de Silva 2012).

Para obter os dados baixe via web com o seguinte comando:

```
dados<-read.table("http://files.r-facil.webnode.com/200000016-b6ee0b765f/digitos.txt",header=TRUE, sep=" ", dec=".")
```

Para observar os dados:

dados

	A	B	C	D	E
Búfalo	0.34	0.72	0.61	0.63	5.39
Curraleiro	0.60	0.46	0.26	0.49	4.54
Nelore	0.86	0.64	0.33	0.53	4.95
Pantaneiro	0.59	0.52	0.30	0.53	4.56

Sugere-se padronizar os dados utilizando a função “scale()”, evitando problemas como diferentes unidades de medida ou, médias e variâncias diferentes. Com a padronização os dados têm média zero e variância igual a um.

```
dados=scale(dados)
```

dados

	A	B	C	D	E
Búfalo	-1.21263218	1.1533828	1.4754896	1.4232712	1.3230719
Curraleiro	0.01177313	-1.0679471	-0.7220481	-0.9209402	-0.7988359
Nelore	1.23617844	0.4698967	-0.2825406	-0.2511655	0.2246726
Pantaneiro	-0.03531938	-0.5553325	-0.4709009	-0.2511655	-0.7489086

Após a padronização dos dados realiza-se o cálculo das distâncias euclidianas utilizando a função “dist()”.

```
m=dist(dados, diag=TRUE, upper=TRUE)
```

```
m
```

	Búfalo	Curraleiro	Nelore	Pantaneiro
Búfalo	0.0000000	4.6109074	3.6830309	3.8975699
Curraleiro	4.6109074	0.0000000	2.3565794	0.8826988
Nelore	3.6830309	2.3565794	0.0000000	1.9107961
Pantaneiro	3.8975699	0.8826988	1.9107961	0.0000000

Valores maiores indicam pares mais dissimilares (diferentes) e valores menores pares menos dissimilares. O par mais dissimilar ocorre para pantaneiro e búfalo. O par menos dissimilar ocorre entre pantaneiro e curraleiro.

Medidas de dissimilaridade qualitativas (binárias)

Aqui será exemplificado como obter medidas de dissimilaridade através de variáveis qualitativas, utilizando medidas de dissimilaridade binárias.

Considere os dados simulados abaixo onde é observada a presença e ausência de cinco marcas moleculares em três genótipos (g1, g2 e g3).

```
g1=c(1,0,0,0,1)
```

```
g2=c(1,0,0,1,0)
```

```
g3=c(0,1,1,0,1)
```

```
dados=data.frame(g1,g2,g3)
```

```
dados
```

```
g1 g2 g3
```

```
1 1 1 0
```

```
2 0 0 1
```

```
3 0 0 1
```

```
4 0 1 0
```

```
5 1 0 1
```

Para estimar medidas binárias utilizaremos a função “vegdist()” do pacote “vegan()”, que deve ser previamente instalado antes de ser carregado.

```
require(vegan)
```

```
vegdist(t(dados), method="jaccard", diag=TRUE,  
upper=TRUE)
```

```
      g1      g2      g3  
g1 0.0000000 0.6666667 0.7500000  
g2 0.6666667 0.0000000 1.0000000  
g3 0.7500000 1.0000000 0.0000000
```

Na função o argumento “method” define o método a ser utilizado entre os vários disponíveis na função (ver documentação da função para mais métodos). O método aqui utilizado é um dos mais comuns (jaccard). O resultado obtido indica maior dissimilaridade (diferença) entre genótipos quanto maior o valor e, menor dissimilaridade quanto menor o valor. Genótipos onde a dissimilaridade é igual a “1” são diferentes em todas as variáveis (marcas moleculares) analisadas. O genótipo 3 (g3) é totalmente diferente do genótipo 2 (g2) para as marcas analisadas. Valores iguais a “0” indicam que aquele par tem as mesmas marcas moleculares. O par de genótipo mais similar (menos dissimilar) foi g1 e g2.

Dendrogramas

A análise de agrupamento multivariada por esquema de “árvore” ou dendrograma, será realizada através de métodos chamados hierárquicos. Para exemplificar será utilizada a distância euclidiana como medida de dissimilaridade (é necessária uma matriz de dissimilaridade para construção dos dendrogramas).

Vamos considerar as medidas de comprimento dos dígitos de 15 animais nelore, 15 curraleiros, 15 pantaneiros e 12 búfalos (dados parciais da dissertação de Silva 2012). Estes dados estão no formato adequado para análise e podem ser baixados via web com o comando a seguir:

```
dados<-read.table("http://files.r-facil.webnode.com/200
000016-b6ee0b765f/digitos.txt",header=TRUE,      sep="",
dec=".")
```

dados

	A	B	C	D	E
Búfalo	0.34	0.72	0.61	0.63	5.39
Curraleiro	0.60	0.46	0.26	0.49	4.54
Nelore	0.86	0.64	0.33	0.53	4.95
Pantaneiro	0.59	0.52	0.30	0.53	4.56

Os dados devem ser padronizados (para tal propósito utilizaremos a função “scale()”) para eliminar a interferência da escala de medida dos dados.

```
dados=scale(dados)
```

dados

	A	B	C	D	E
Búfalo	-1.21263218	1.1533828	1.4754896	1.4232712	1.3230719
Curraleiro	0.01177313	-1.0679471	-0.7220481	-0.9209402	-0.7988359
Nelore	1.23617844	0.4698967	-0.2825406	-0.2511655	0.2246726
Pantaneiro	-0.03531938	-0.5553325	-0.4709009	-0.2511655	-0.7489086

Após a padronização obtém-se a distância euclidiana (padronizada), salvando a matriz resultante em um objeto denominado "m".

```
m=dist(dados, diag=TRUE, upper=TRUE)
```

m

Búfalo	Curraleiro	Nelore	Pantaneiro
--------	------------	--------	------------

Búfalo	0.0000000	4.6109074	3.6830309	3.8975699
Curraleiro	4.6109074	0.0000000	2.3565794	0.8826988
Nelore	3.6830309	2.3565794	0.0000000	1.9107961
Pantaneiro	3.8975699	0.8826988	1.9107961	0.0000000

O par mais similar ocorre entre pantaneiro e curraleiro e o menos similar entre curraleiro e búfalo. Note também que búfalos têm menos similaridade em relação aos demais.

Na sequência será realizada a construção de dendrogramas por quatro métodos hierárquicos, utilizando a matriz “m” acima.

Método UPGMA (Unweighted Pair Group Method with Arithmetic Mean).

```
cluster1=hclust(m, method="average")
```

Método de Ward.

```
cluster2=hclust(m, method="ward")
```

Método do vizinho mais próximo (ligação simples)

```
cluster3=hclust(m, method ="single")
```

Método do vizinho mais distante (ligação completa)

```
cluster4=hclust(m, method="complete")
```

Para melhor apresentação dos dendrogramas vamos transformar os objetos criados, como na programação que segue.

```
c1=as.dendrogram(cluster1)
```

```
c2=as.dendrogram(cluster2)
```

```
c3=as.dendrogram(cluster3)
```

```
c4=as.dendrogram(cluster4)
```

E enfim, utilizando a função “plot()”, obtém-se a construção dos dendrogramas em cada método (Figuras 1, 2, 3 e 4). Observe que os búfalos formam um grupo isolado e curraleiro e pantaneiro estão no mesmo grupo. Resultado anteriormente observado na matriz

de distância euclidiana, mas aqui, representado na forma de grupos. Repare também que todos os métodos formam os mesmos grupos.

```
plot(c1, ylim=c(-0.5,6), main="upgma")  
plot(c2, ylim=c(-0.5,8), main="ward")  
plot(c3, ylim=c(-0.5,6), main="single")  
plot(c4, ylim=c(-0.5,8), main="complete")
```

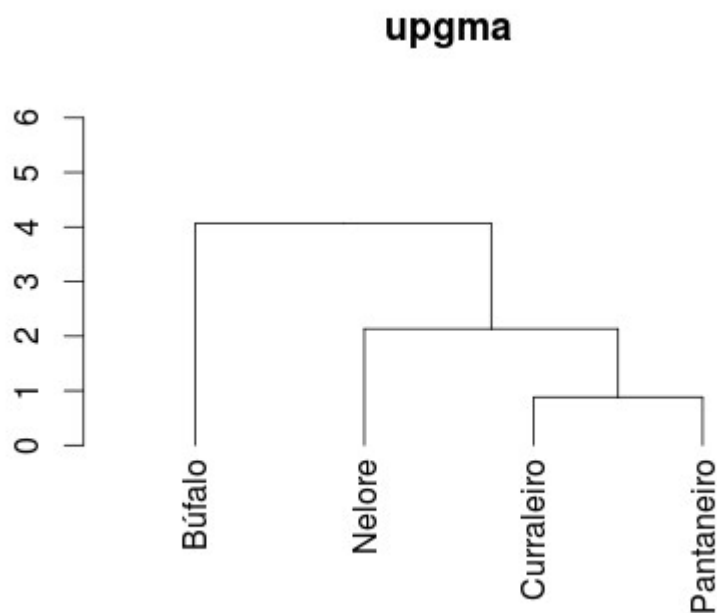


Figura 1. Dendrograma pelo método UPGMA, construído a partir da distância euclidiana de medidas de comprimento de dígitos de búfalos, nelores, curraleiros e pantaneiros.

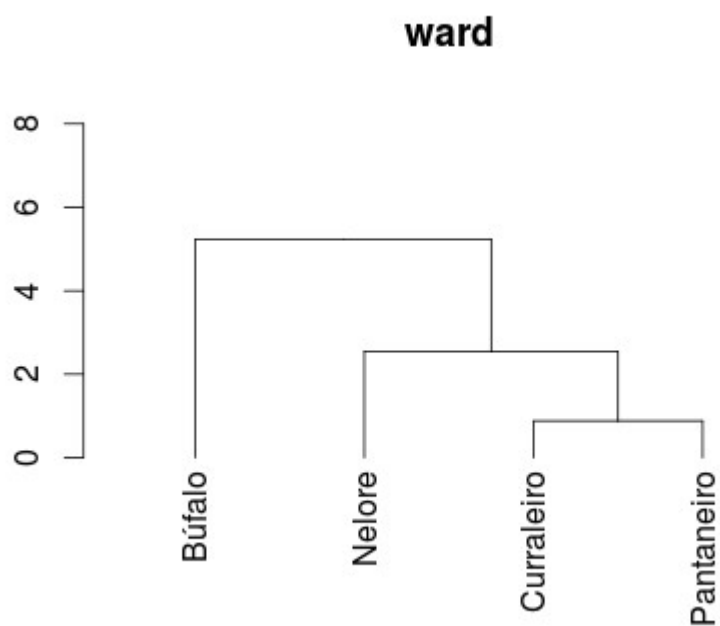


Figura 2. Dendrograma pelo método Ward, construído a partir da distância euclidiana de medidas de comprimento de dígitos de búfalos, nelores, curraleiros e pantaneiros.

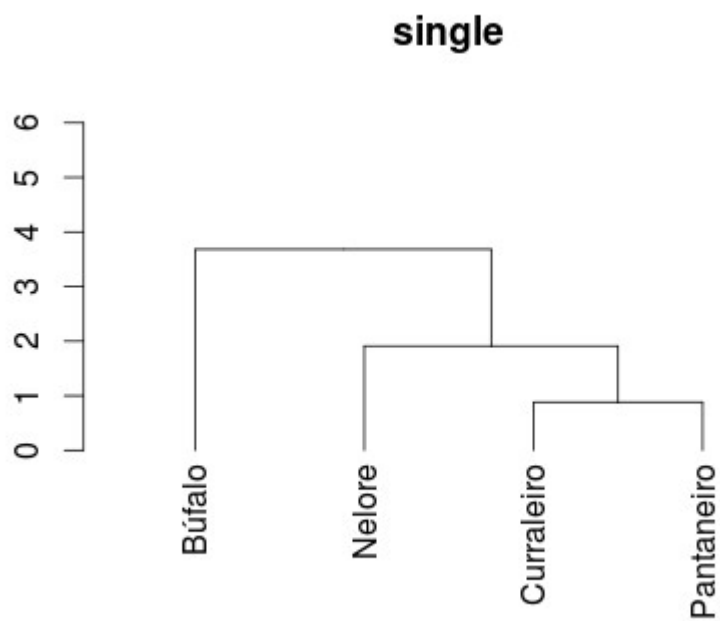


Figura 3. Dendrograma pelo método do vizinho mais próximo (ligação simples), construído a partir da distância euclidiana de medidas de comprimento de dígitos de búfalos, nelores, curraleiros e pantaneiros.

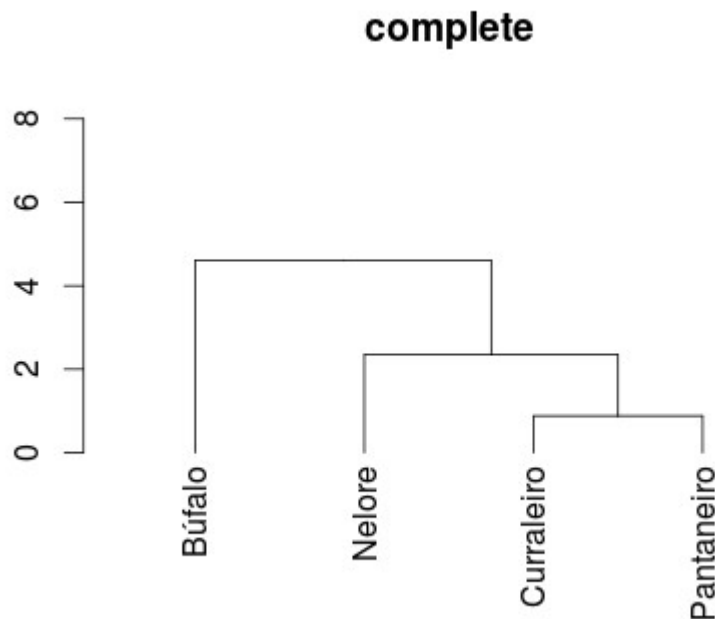


Figura 4. Dendrograma pelo método do vizinho mais distante (ligação completa), construído a partir da distância euclidiana de medidas de comprimento de dígitos de búfalos, nelores, curraleiros e pantaneiros.

Para criar dendogramas horizontais (Figura 5), basta usar o argumento `horiz=TRUE`, como programado abaixo.

```
plot(c1, xlim=c(6,-1.6), horiz=TRUE)
```

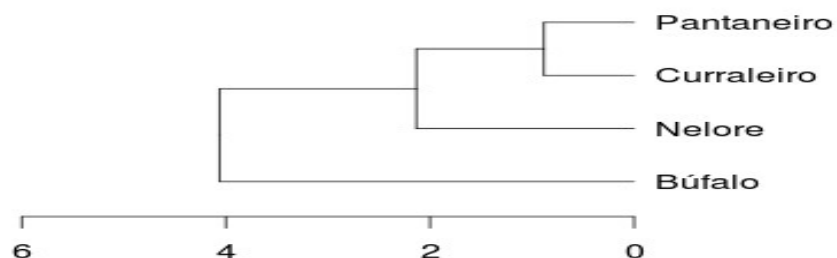


Figura 5. Dendrograma pelo método UPGMA, construído a partir da distância euclidiana de medidas de comprimento de dígitos de búfalos, nelores, curraleiros e pantaneiros.

Qual o melhor método de construção de dendrograma? O melhor método é aquele que possui maior correlação da matriz cofenética (matriz de distância obtida através do método de agrupamento) com a matriz de dissimilaridade.

Abaixo segue a forma de obtenção destas correlações.

Matrizes cofenéticas:

```
cof1=cophenetic(cluster1)
cof2=cophenetic(cluster2)
cof3=cophenetic(cluster3)
cof4=cophenetic(cluster4)
```

A seguir as correlações entre matrizes cofenéticas e a matriz de distância euclidiana com dados padronizados (m).

```
cor(cof1,m) # UPGMA
[1] 0.970652
```

```
cor(cof2,m) # Ward
[1] 0.9705927
```

```
cor(cof3,m) # Ligação simples
[1] 0.9702712
```

```
cor(cof4,m) # Ligação completa
[1] 0.9706495
```

Pelo critério exposto recomenda-se utilizar o método UPGMA (maior correlação).

Componentes principais (PCA)

A análise de componentes principais, no inglês “principal component analysis” (PCA) é um método estatístico que transforma um conjunto de observações de variáveis em um conjunto de valores de variáveis chamadas componentes principais. Com esta análise pode-se:

- 1) Analisar graficamente no espaço bidimensional (ou tridimensional) a dissimilaridade entre os elementos amostrais (animais, plantas, ecossistemas, etc).
- 2) Analisar no espaço bidimensional (ou tridimensional) as correlações existentes entre as variáveis quantitativas utilizadas para obtenção dos componentes principais.
- 3) Analisar a importância das variáveis quantitativas para a variação encontrada entre os elementos amostrais (animais, plantas, ecossistemas, etc). Ou seja, quais variáveis têm maior importância para diferenciar os elementos amostrais.

Para exemplificar o uso desta técnica vamos considerar as medidas de comprimento dos dígitos de 15 animais nelore, 15 curraleiros, 15 pantaneiros e 12 búfalos, (dados parciais da dissertação de Silva 2012). Estes dados podem ser baixados via web com o comando a seguir:

```
dados<-read.table("http://files.r-facil.webnode.com/200000016-b6ee0b765f/digitos.txt",header=TRUE, sep=" ", dec=".")
```

dados

	A	B	C	D	E
Búfalo	0.34	0.72	0.61	0.63	5.39
Curraleiro	0.60	0.46	0.26	0.49	4.54
Nelore	0.86	0.64	0.33	0.53	4.95
Pantaneiro	0.59	0.52	0.30	0.53	4.56

Após entrar com a tabela acima em um objeto denominado "dados", obtemos os componentes principais utilizando a função "prcomp()".

```
cp=prcomp(~A+B+C+D+E, scale = TRUE, data=dados)
```

A análise foi gravada em um objeto aqui denominado "cp". Na função “prcomp()” deve-se acrescentar após o “~” o nome das variáveis (A+B+C+D+E). O argumento “scale=TRUE” (opcional, mas recomendado) refere-se a padronização dos dados e, o argumento “data” indica o nome do objeto onde estão contidas as variáveis.

A importância de cada variável para diferenciar os animais (grupos de animais) é tanto maior quanto maior for o valor (em módulo) associado à variável. Observe abaixo que para o primeiro componente todas as variáveis (medidas de dígitos) tem importância semelhante no estudo (apesar de pequena superioridade da variável C). Para o segundo componente a variável de maior importância é a A. Esta análise deve ser feita para os dois primeiros componentes, com maior importância para o primeiro.

```
cp$rotation
```

	PC1	PC2	PC3	PC4
A	-0.3204996	-0.83524822	0.1099610	0.38065822
B	0.4454724	-0.46343394	0.2486452	-0.37843193
C	0.4899380	0.08733382	-0.1748664	0.80383126
D	0.4865881	0.05482483	0.6689900	0.03392931
E	0.4711975	-0.27741022	-0.6692969	-0.25415095

Pode-se também observar a importância das variáveis pela correlação entre os componentes de cada variável e a matriz de correlação dos dados originais (abaixo). Maiores correlações indicam maior importância da variável. Nesta análise deve-se considerar principalmente o componente 1 (PC1) e também, com menor importância, o componente 2 (PC2).

```
cor(cor(dados), cp$rotation)
```

	PC1	PC2	PC3	PC4
A	-0.9777456	-0.9264544	0.03186677	0.0943770
B	0.9885728	0.7328249	-0.10800235	-0.3879260
C	0.9997947	0.8378016	-0.08120713	-0.2516533
D	0.9996691	0.8367275	-0.05927786	-0.2586718
E	0.9961929	0.7762294	-0.12236727	-0.3297475

Outra informação importante é quanta variância foi acumulada com os dois primeiros componentes. Esta informação é obtida pela função “summary()”. Considera-se que uma análise de componentes principais é adequada quando a variância acumulada com os dois primeiros componentes for superior a 70% (0,70). Veja a seguir, neste exemplo, que os dois primeiros componentes acumularam 0,99 da variância.

```
summary(cp)
```

Importance of components:

	PC1	PC2	PC3	PC4
Standard deviation	2.0333	0.9077	0.20388	5.646e-16
Proportion of Variance	0.8269	0.1648	0.00831	0.000e+00
Cumulative Proportion	0.8269	0.9917	1.00000	1.000e+00

Com a dispersão gráfica dos dois primeiros componentes (Figura 6), utilizando a função “biplot()”, pode-se avaliar a similaridade entre os grupos de animais e as correlações entre variáveis. Animais mais próximos no gráfico são mais similares considerando o conjunto de variáveis analisadas. Variáveis mais próximas, com ângulo entre as setas se fechando, têm correlação elevada e positiva. Variáveis opostas no gráfico, tendendo a formar um ângulo de 180°, tem correlação elevada e negativa. E variáveis onde o ângulo entre as setas tende a 90° tem correlação aproximadamente nula.

`biplot(cp)`

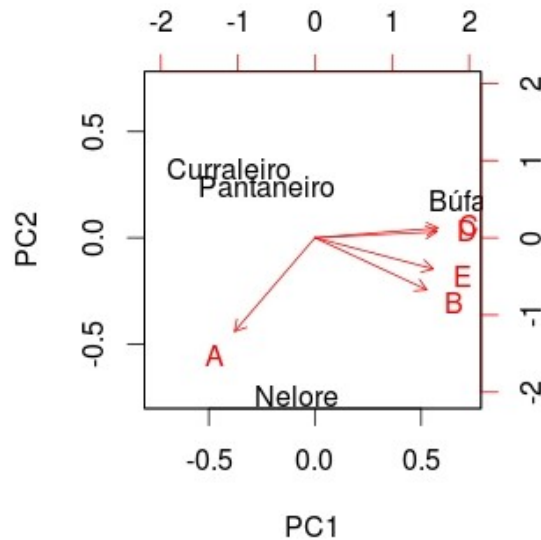


Figura 6. Dispersão gráfica dos dois primeiros componentes principais, obtidos de cinco medidas de comprimento de dígitos (medidas A, B, C e D) de búfalos, nelores, curraleiros e pantaneiros.

A seguir temos uma programação para fazer este mesmo gráfico com alguns ajustes (Figura 7).

```
biplot(cp,      xlim=c(-0.9,0.9),      ylim=c(-0.8,0.8),  
ylab="Componente 2 (16%)", xlab="Componente 1 (83%)",  
col=1)  
abline(h=0, v=0, lty=2, col=2)  
points(x=0,y=0,pch=1,cex=30, col="dark blue")
```

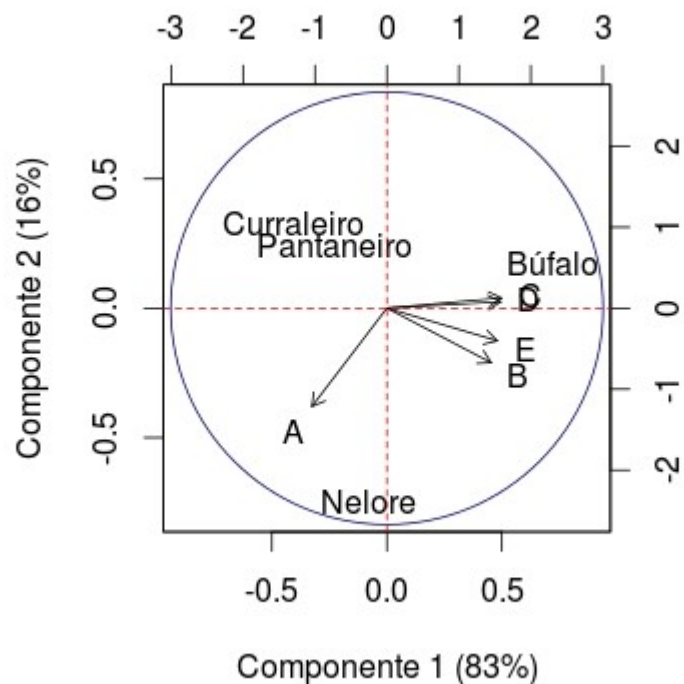


Figura 7. Dispersão gráfica dos dois primeiros componentes principais, obtidos de cinco medidas de comprimento de dígitos (medidas A, B, C e D) de búfalos, nelores, curraleiros e pantaneiros.

Neste gráfico observa-se que os curraleiros e os pantaneiros são mais similares. As variáveis C e D são correlacionadas positivamente (com o ângulo aproximando de zero a correlação se aproxima de +1). As variáveis A e B tem correlação próxima a zero (com ângulo aproximando a 45 graus a correlação aproxima a 0). E as variáveis A e C têm correlação negativa (ângulo aproximando a 180 graus a correlação aproxima de -1).

Por fim, pode-se obter gráficos com maior controle para observar a dissimilaridade entre grupos (Figura 8) e as correlações entre variáveis (Figura 9).

Gráfico para observar a dissimilaridade entre grupos:

```
c=cp$x
c=as.data.frame(c)
attach(c)
x1=c$PC1# primeiro componente
x2=c$PC2 # segundo componente
```

```

plot(x2~x1, ylab= "Componente 2 (16%) ", xlab=
"Componente 1 (83%) ", cex=0, xlim=c(-5,5),
ylim=c(-3,3), bty="n", axes=FALSE)
axis(1,c(-4,-3,-2,-1,0,1,2,3,4))
axis(2,c(-4,-3,-2,-1,0,1,2,3,4))
text(x1,x2, label=rownames(dados), cex=1.2,
col=c(rep(2),rep(3), rep(4),rep(5)))
abline(v=0,h=0, lty=2, col="dark blue")
rect(4,0.6,2,0, lty=2)
rect(-1.8,-1.8,1.1,-1, lty=1)
rect(-3,1,0.7,0, lty=3)

```

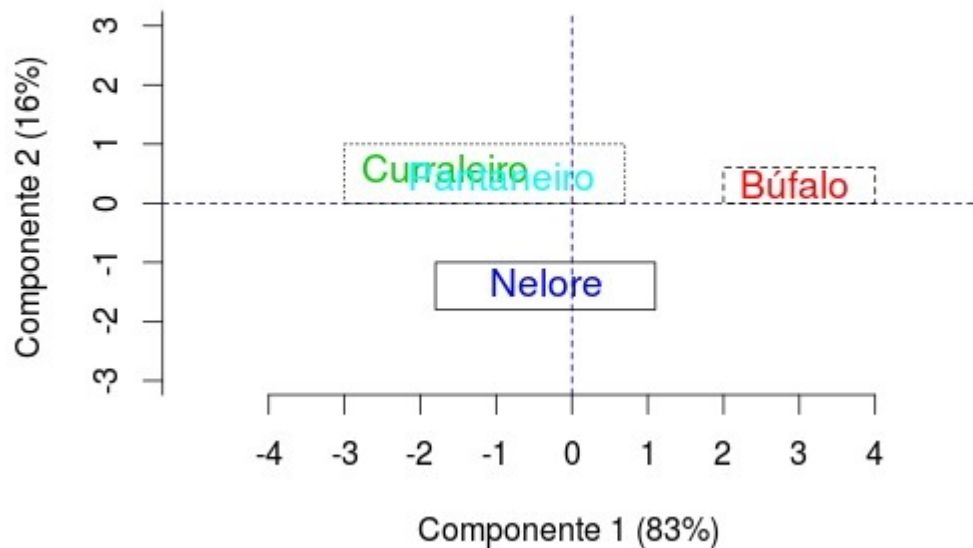


Figura 8. Dispersão gráfica dos dois primeiros componentes principais, obtidos de cinco medidas de comprimento de dígitos (medidas A, B, C e D) de búfalos, nelores, curraleiros e pantaneiros.

Gráfico para observar as correlações:

```
cpr=as.data.frame(cp$rotation)
y1=cpr$PC1
y2=cpr$PC2
plot(y1,y2, cex=0, xlim=c(-0.4,0.6), xlab="PC1",
ylab="PC2")
text(y1,y2, label=rownames(cpr), pos=4)
arrows(0,0,y1,y2, lty=2, col="dark blue")
```

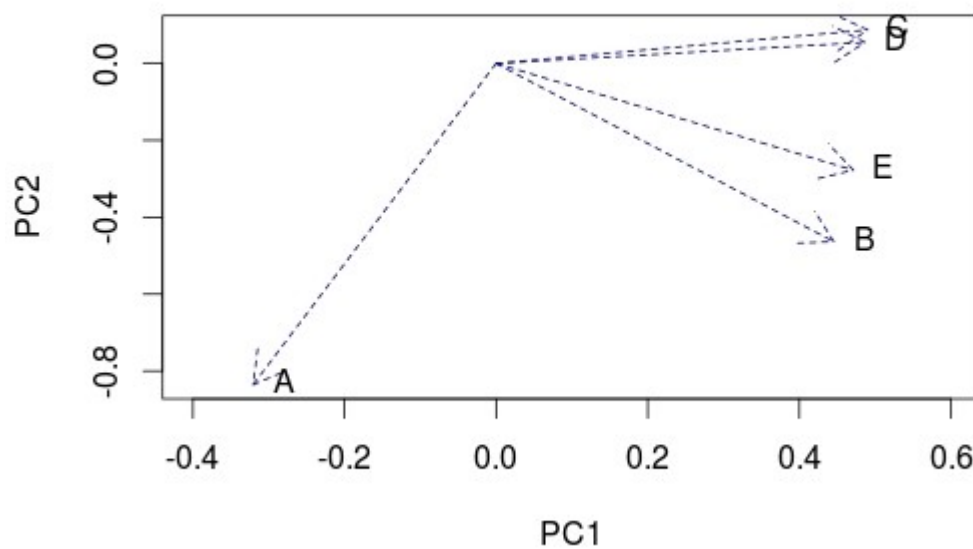


Figura 9. Dispersão gráfica dos dois primeiros componentes principais, obtidos de cinco medidas de comprimento de dígitos (medidas A, B, C e D) de búfalos, nelores, curraleiros e pantaneiros.

Variáveis canônicas

A análise de variáveis canônicas é uma técnica estatística multivariada, similar a técnica de componentes principais. Com a análise de variáveis canônicas é possível avaliar em um gráfico (bidimensional ou tridimensional), a dissimilaridade entre elementos amostrais (animais, plantas, raças, ecossistemas, etc). Também é possível avaliar a correlação entre as variáveis quantitativas obtidas dos elementos amostrais. Esta técnica exige informações entre grupos (elementos amostrais) e dentro de grupos, havendo necessidade de repetições para realizar uma análise de variáveis canônicas. No caso dos componentes principais, não há necessidade de repetições. Caso haja repetições e a intenção é realizar uma análise de componentes principais, deve-se utilizar a média das repetições.

Para exemplificar o uso das variáveis canônicas, vamos considerar as medidas de comprimento dos dígitos de 15 animais nelore, 15 curraleiros, 15 pantaneiros e 12 búfalos (dados parciais da dissertação de Silva 2012). Estes podem ser baixados via web com o comando a seguir:

```
dados<-read.table("http://files.r-facil.webnode.com/200  
000017-de9e4df161/dig.txt",header=TRUE,      sep="",  
dec=".")
```

Vamos requerer o pacote necessário denominado "candisc".

```
require(candisc)
```

Na sequência faremos uma análise de variância multivariada, considerando o delineamento em questão (inteiramente ao acaso). Observe que existe diferença estatística entre os grupos de animais.

```
dados2=dados[,-1]  
dados2=as.matrix(dados2)  
m=lm(dados2~dados$RAÇA)
```

```
anova(m)
```

Analysis of Variance Table

	Df	Pillai	approx	F	num	Df	den	Df	Pr(>F)
(Intercept)	1	0.9996		8739.2	12		42		< 2.2e-16 ***
dados\$RAÇA	3	1.7212		4.9	36		132		7.151e-12 ***
Residuals	53								

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Na sequência pode-se observar a dissimilaridade entre grupos de animais e a correlação entre variáveis (Figura 10).

```
c=candisc(m, term="dados$RAÇA")
plot(c, scale=6, type="n ")
```

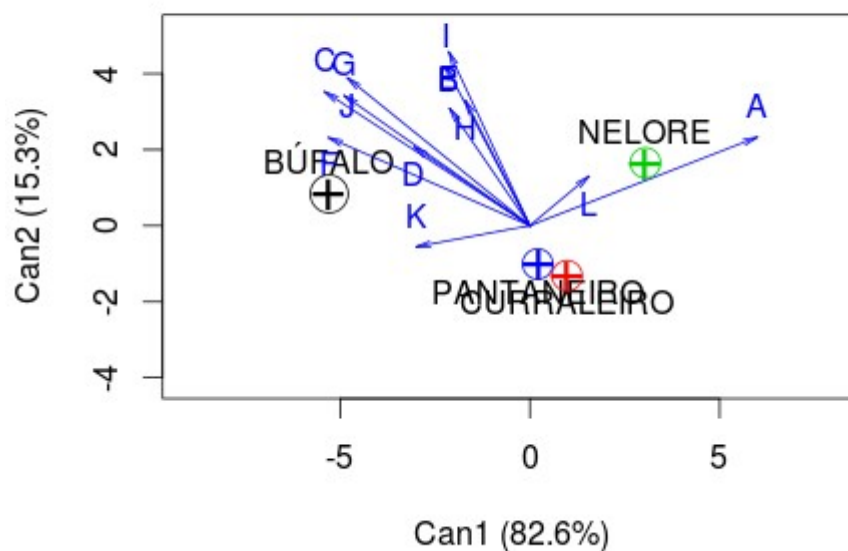


Figura 10. Dispersão gráfica das duas primeiras variáveis canônicas, obtidas de medidas de comprimento de dígitos de búfalos, nelores, curraleiros e pantaneiros.

Com este gráfico (Figura 10), pode-se avaliar a similaridade entre os grupos de animais e as correlações entre variáveis. Animais mais próximos no gráfico são mais similares considerando o conjunto de variáveis analisadas. Variáveis mais próximas, com ângulo entre as setas se fechando, têm correlação elevada e positiva. Variáveis opostas no gráfico, tendendo a formar um ângulo de 180°, tem correlação elevada e negativa. E variáveis onde o ângulo entre as setas tende a 90° tem correlação aproximadamente nula.

É importante que as duas primeiras variáveis canônicas tenham uma variância acumulada acima de 70% (0,70). No caso (veja na Figura 10) a soma da variância das duas primeiras variáveis canônicas foi de 97,9% (0,979), sendo que só a primeira variável canônica contém 82,6% da variância.

Na sequência uma programação para obter um gráfico bastante interessante para avaliar a dissimilaridade entre os grupos de animais (Figura 11).

```
names(c)
can=c$"scores"
attach(can)
plot(Can2~Can1, ylab= "Can 2 (15,28% )", xlab= "Can 1
(82,65%)", col.main= "dark blue", cex=0,
xlim=c(-8,8), ylim=c(-8,8), axes=FALSE)
axis(1,c(-7,-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6,7),
col="dark blue", lty=2)
axis(2,c(-5,-4,-3,-2,-1,0,1,2,3,4,5), col="dark blue",
lty=2)
abline(h=0,v=0, lty=2, col="dark blue")
text(Can1,Can2, label=dados$RAÇA, cex=1, col=c(rep(2,
15),rep(3, 15), rep(4, 15), rep(5,12)))
```

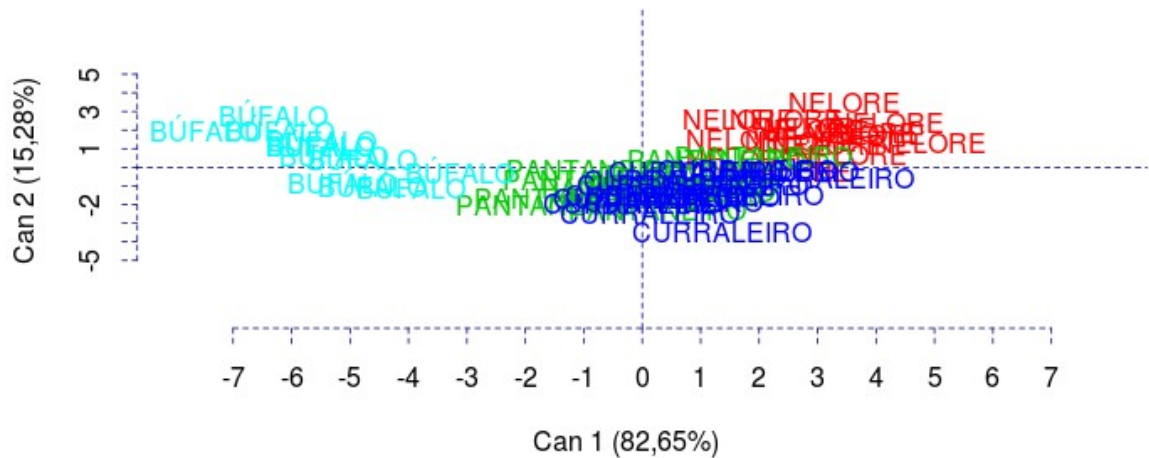


Figura 11. Dispersão gráfica das duas primeiras variáveis canônicas, obtidas de medidas de comprimento de dígitos de búfalos, nelores, curraleiros e pantaneiros.

Correlações canônicas

A análise de correlações canônicas é uma técnica estatística multivariada utilizada para avaliar correlações entre dois grupos de variáveis. Pode-se utilizar esta técnica para correlacionar variáveis da parte aérea de plantas com variáveis do sistema radicular ou, correlacionar variáveis relacionadas a produção com variáveis relacionadas a qualidade (ex.: produção de leite vs qualidade do leite).

Para exemplificar vamos considerar os dados simulados a seguir, que podem ser baixados via web.

```
dados<-read.table("http://files.r-facil.webnode.com/20000047-24ffa26f24/correlações.txt",header=TRUE, sep=" ", dec=".")
```

dados

	V1	V2	V3	L1	L2	L3	L4
1	45	500	25	12	5	18	89
2	48	460	29	16	9	25	80
3	55	420	56	19	8	69	55
4	62	390	48	19	4	32	23
5	63	375	75	22	3	14	12
6	72	350	96	23	2	5	10

Os dados se referem a dois grupos de variáveis (Grupo V com três variáveis e Grupo L com quatro). Vamos separar os grupos em dois arquivos.

```
V=dados[,c(1,2,3)]
```

```
L=dados[,c(4,5,6,7)]
```

V

	V1	V2	V3
1	45	500	25
2	48	460	29
3	55	420	56
4	62	390	48
5	63	375	75
6	72	350	96

L

	L1	L2	L3	L4
1	12	5	18	89
2	16	9	25	80
3	19	8	69	55
4	19	4	32	23
5	22	3	14	12
6	23	2	5	10

Agora vamos requerer o pacote "yacca" que deve estar previamente instalado.

require(yacca)

Na sequência utilizaremos a função "cca()"

cca(V,L)

A seguir pode-se utilizar a função "summary()" para um resultado mais completo.

summary(cca(V,L))

Do resultado gerado pela função devemos ficar atento no seguinte:

Valores das correlações canônicas (abaixo) que devem ser significativos (pelo menos a primeira correlação CV1):

Canonical Correlations:

	CV 1	CV 2	CV 3
	1.0000000	1.0000000	0.7363036

Bartlett's Chi-Squared Test:

	rho^2	Chisq	df	Pr(>X)
CV 1	1.00000	67.47488	12	9.476e-10 ***
CV 2	1.00000	32.12437	6	1.545e-05 ***
CV 3	0.54214	0.78120	2	0.6767

Na sequência devemos observar as correlações que permitem o estudo da associação entre as variáveis dos dois grupos. Observando o primeiro par canônico abaixo (CV 1), verifica-se que V1 e V3 se correlacionam no mesmo sentido com L1 e de forma oposta a L4. Também se verifica que a variável V2 correlaciona de forma oposta a L1 e no mesmo sentido

de L4. E as variáveis L2 e L3 não são correlacionadas com variáveis do grupo V. A variável L2 ainda tem uma correlação fraca (0,57) que pode ser considerada. O segundo par canônico (CV 2) não fornece informações relevantes. O terceiro par canônico nem deve ser considerado, pois não é significativo.

Structural Correlations (Loadings):

X Vars:

	CV 1	CV 2	CV 3
V1	-0.9477672	0.2499445	-0.31089437
V2	0.9857650	-0.1219877	0.15446669
V3	-0.9427324	0.5003291	-0.06434098

Y Vars:

	CV 1	CV 2	CV 3
L1	-0.9974319	0.1680747	0.05495388
L2	0.5743614	-0.4074123	0.76009809
L3	0.1405778	-0.3569417	0.62288443
L4	0.9473280	-0.0914460	0.30499271

Análise discriminante

A análise discriminante é realizada com objetivo de classificar elementos amostrais (indivíduos, animais, plantas, raças, cultivares, ecossistemas, etc) em grupos pré definidos, a partir de variáveis (variáveis quantitativas) obtidas destes elementos.

Para exemplificar a utilização da análise discriminante, vamos considerar os valores das medidas de comprimento dos dígitos de 15 animais nelore, 15 curraleiros, 15 pantaneiros e 12 búfalos, contidos no quadro abaixo (dados parciais da dissertação de Silva 2012) e que podem ser baixados via web com o comando a seguir.

```
dados<-read.table("http://files.r-facil.webnode.com/200
000017-de9e4df161/dig.txt",header=TRUE,
sep=" ",
dec=".")
```

dados

RAÇA	A	B	C	D	E	F	G	H	I	J	K	L	
1	NELORE	0.920	0.557	0.192	0.628	4.944	4.621	3.445	7.673	9.218	3.420	53.375	46.481
2	NELORE	0.931	0.551	0.322	0.497	4.688	4.174	3.251	7.204	8.134	3.366	53.024	51.105
3	NELORE	0.859	0.520	0.349	0.457	4.866	4.179	3.204	6.588	8.675	3.252	54.087	50.789
4	NELORE	0.807	0.427	0.222	0.542	4.953	4.471	3.281	6.596	8.204	3.142	53.459	47.441
5	NELORE	0.734	0.880	0.321	0.434	5.171	4.361	3.401	6.769	9.552	3.831	54.505	50.377
6	NELORE	0.653	0.651	0.292	0.499	5.051	4.098	3.365	7.018	9.347	3.834	52.216	52.380
7	NELORE	0.897	0.726	0.272	0.639	5.231	4.361	3.465	8.117	9.825	3.554	46.522	46.882
8	NELORE	0.793	0.753	0.396	0.527	5.445	4.216	3.313	8.534	10.186	4.153	47.479	47.552
9	NELORE	1.014	0.700	0.464	0.335	4.763	4.324	3.273	7.070	9.726	3.682	48.430	50.385
10	NELORE	0.817	0.817	0.426	0.523	4.916	4.129	3.251	6.362	8.668	3.762	55.546	52.500
11	NELORE	0.971	0.569	0.356	0.450	4.261	3.443	2.659	6.480	8.830	3.616	50.634	51.172
12	NELORE	0.962	0.753	0.356	0.654	5.211	4.132	3.229	7.417	10.230	3.567	50.989	51.660
13	NELORE	0.937	0.682	0.255	0.614	5.485	4.412	3.290	7.442	10.455	3.368	44.898	48.589
14	NELORE	0.956	0.401	0.318	0.600	4.611	4.050	3.180	7.395	9.814	3.268	48.730	51.583
15	NELORE	0.707	0.619	0.338	0.513	4.706	4.397	3.413	7.266	9.397	3.010	47.943	50.452
16	PANTANEIRO	0.651	0.311	0.180	0.531	4.842	4.540	3.231	6.860	9.477	3.900	52.912	49.703
17	PANTANEIRO	0.806	0.726	0.190	0.581	0.522	4.643	3.319	7.664	9.822	3.970	54.636	47.893
18	PANTANEIRO	0.722	0.608	0.288	0.467	4.494	3.786	2.969	7.264	9.172	3.598	50.452	50.387
19	PANTANEIRO	0.831	0.529	0.366	0.598	4.765	4.735	3.385	6.356	8.575	3.668	59.785	45.970
20	PANTANEIRO	0.466	0.572	0.394	0.581	5.215	4.545	3.419	6.036	8.871	3.488	57.026	49.114
21	PANTANEIRO	0.498	0.784	0.218	0.554	5.257	4.529	3.457	10.099	8.456	3.654	46.983	45.139
22	PANTANEIRO	0.490	0.469	0.348	0.428	4.732	3.760	2.864	6.180	9.295	3.152	45.868	46.622
23	PANTANEIRO	0.595	0.552	0.342	0.655	4.902	4.102	3.138	6.641	8.208	3.085	46.785	48.984
24	PANTANEIRO	0.509	0.675	0.240	0.636	5.581	4.645	3.585	7.410	9.975	3.663	48.876	47.938
25	PANTANEIRO	0.509	0.511	0.268	0.550	4.814	4.133	3.305	6.387	8.052	2.918	52.381	50.816
26	PANTANEIRO	0.492	0.369	0.222	0.563	4.472	3.980	3.099	8.259	6.342	2.623	51.228	53.126
27	PANTANEIRO	0.618	0.428	0.333	0.404	4.374	3.880	2.810	6.571	9.013	3.220	48.175	47.443
28	PANTANEIRO	0.532	0.512	0.394	0.460	4.870	4.158	3.030	6.449	8.697	3.195	47.556	47.370
29	PANTANEIRO	0.585	0.273	0.402	0.488	4.593	4.870	3.290	6.515	8.586	3.103	51.220	45.937
30	PANTANEIRO	0.584	0.493	0.360	0.433	5.018	4.414	3.012	6.424	8.617	3.137	51.982	43.736
31	CURRALEIRO	0.461	0.380	0.233	0.345	4.943	4.346	3.202	6.575	9.403	2.845	51.091	51.845
32	CURRALEIRO	0.854	0.456	0.204	0.544	4.597	4.009	2.928	7.813	6.342	2.925	51.881	48.195
33	CURRALEIRO	0.628	0.628	0.194	0.582	4.865	4.668	3.599	6.280	8.371	3.446	57.825	50.866
34	CURRALEIRO	0.570	0.402	0.302	0.416	4.394	4.128	3.181	5.695	7.633	2.977	59.306	49.670
35	CURRALEIRO	0.794	0.441	0.434	0.601	4.532	4.257	3.130	6.733	9.184	3.029	52.248	48.427
36	CURRALEIRO	0.652	0.180	0.146	0.650	3.902	3.672	2.605	5.312	7.380	3.214	57.009	46.499
37	CURRALEIRO	0.527	0.612	0.254	0.439	4.963	4.322	3.057	7.032	9.088	3.335	51.573	46.250
38	CURRALEIRO	0.440	0.555	0.231	0.633	5.084	4.587	3.426	6.354	8.069	3.141	57.941	49.334
39	CURRALEIRO	0.383	0.575	0.164	0.562	4.496	4.013	3.001	6.257	8.237	3.517	53.211	50.621
40	CURRALEIRO	0.468	0.555	0.256	0.474	5.048	4.117	3.001	5.916	7.458	3.383	59.720	48.694
41	CURRALEIRO	0.612	0.497	0.243	0.424	3.761	4.009	3.154	6.323	8.237	2.966	50.214	52.852
42	CURRALEIRO	0.693	0.431	0.305	0.351	4.135	3.798	2.693	5.921	7.274	2.832	53.057	46.422
43	CURRALEIRO	0.743	0.406	0.191	0.495	4.814	4.131	3.208	6.923	8.537	2.428	50.309	52.906
44	CURRALEIRO	0.643	0.472	0.389	0.419	4.584	4.307	3.140	6.669	9.324	3.090	48.775	47.885
45	CURRALEIRO	0.558	0.378	0.365	0.424	3.976	3.928	2.971	5.492	7.700	2.705	54.233	50.396
46	BÚFALO	0.496	0.565	0.788	0.763	5.336	5.133	3.970	7.562	11.153	5.481	56.888	51.064
47	BÚFALO	0.180	0.692	0.514	0.495	5.261	4.444	3.323	6.455	8.337	4.078	58.691	48.547
48	BÚFALO	0.328	0.534	0.588	0.612	5.368	4.794	3.466	7.223	8.835	3.966	56.125	46.916
49	BÚFALO	0.334	0.906	0.597	0.769	5.487	6.349	4.715	8.156	11.845	5.834	57.940	47.409
50	BÚFALO	0.336	0.806	0.658	0.658	5.686	5.720	4.194	7.502	11.264	5.199	54.378	44.460
51	BÚFALO	0.283	0.768	0.645	0.687	5.691	5.024	3.802	7.477	9.865	4.328	55.121	51.683
52	BÚFALO	0.430	0.625	0.552	0.698	5.400	5.611	4.078	8.296	9.956	4.179	49.848	52.110
53	BÚFALO	0.290	0.581	0.431	0.405	5.023	5.002	3.624	6.945	7.980	3.655	55.510	48.155
54	BÚFALO	0.336	0.806	0.658	0.658	5.686	5.720	4.194	7.502	11.264	5.199	54.378	44.460
55	BÚFALO	0.364	0.545	0.428	0.487	4.626	4.438	3.548	6.965	8.057	3.599	56.963	51.065
56	BÚFALO	0.300	0.898	0.795	0.768	5.788	5.931	4.320	7.639	10.882	5.524	56.122	48.142
57	BÚFALO	0.387	0.886	0.653	0.610	5.372	5.533	4.019	8.292	9.910	4.794	55.014	46.958

Vamos requerer o pacote necessário ("MASS").

```
require(MASS)
```

Depois vamos criar um arquivo de dados sem a primeira coluna que define as raças

```
dados2=dados[,-1]
```

E agora procederemos com a função discriminante linear “lda()” com o resultado em uma tabela (abaixo).

```
y1=lda(dados2,dados$RAÇA, CV=T)  
table(dados$RAÇA, y1$class)
```

	BÚFALO	CURRALEIRO	NELORE	PANTANEIRO
BÚFALO	11	0	0	1
CURRALEIRO	0	7	2	6
NELORE	0	1	13	1
PANTANEIRO	0	6	2	7

No resultado acima temos na diagonal os animais que foram alocados de forma correta e fora da diagonal os animais que estão classificados fora do grupo de origem. Búfalos e nelores foram bem alocados, mas pantaneiros e curraleiros não.

Com a programação a seguir podemos perceber a acurácia da análise discriminante.

```
ct <- table(dados$RAÇA, y1$class)  
diag(prop.table(ct, 1))
```

BÚFALO	CURRALEIRO	NELORE	PANTANEIRO
0.9166667	0.4666667	0.8666667	0.4666667

Percebe-se que a acurácia para búfalos e nelores é maior, enquanto que para curraleiros e pantaneiros tem-se uma menor probabilidade de acerto (menor acurácia).

Abaixo podemos verificar qual animal foi classificado fora do grupo de origem e para qual grupo foi classificado

```
y1$class
```

```
[1] NELORE      NELORE      NELORE      CURRALEIRO NELORE      NELORE      NELORE      NELORE
[9] NELORE      NELORE      NELORE      NELORE      NELORE      NELORE      PANTANEIRO CURRALEIRO
[17] CURRALEIRO NELORE      NELORE      CURRALEIRO PANTANEIRO PANTANEIRO PANTANEIRO PANTANEIRO
[25] CURRALEIRO CURRALEIRO PANTANEIRO PANTANEIRO PANTANEIRO CURRALEIRO CURRALEIRO NELORE
[33] CURRALEIRO CURRALEIRO PANTANEIRO PANTANEIRO PANTANEIRO CURRALEIRO PANTANEIRO CURRALEIRO
[41] PANTANEIRO CURRALEIRO NELORE      PANTANEIRO CURRALEIRO BÚFALO      BÚFALO      BÚFALO
[49] BÚFALO      BÚFALO      BÚFALO      BÚFALO      BÚFALO      BÚFALO      PANTANEIRO BÚFALO
[57] BÚFALO
Levels: BÚFALO CURRALEIRO NELORE PANTANEIRO
```

Percentual de acertos (total de acertos).

```
sum(diag(prop.table(ct)))
```

```
[1] 0.6666667
```

No geral, considerando os quatro grupos, a probabilidade de acerto foi de 0,67. No entanto, alguns grupos tiveram uma probabilidade de acerto bem maior, como os búfalos, com 0,92.

Também se pode utilizar a função discriminante quadrática "qda()" para realizar uma análise discriminante.

Referências

KAPS, M. and LAMBERSON, W. R. **Biostatistics for Animal Science: an introductory text**. 2nd Edition. CABI Publishing, Wallingford, Oxfordshire, UK, 2009. 504p.

SAMPAIO, I. B. M. **Estatística aplicada à experimentação animal**. 3ª Edição. Belo Horizonte: Editora FEPMVZ, Fundação de Ensino e Pesquisa em Medicina Veterinária e Zootecnia, 2010. 264p.

Silva, L.H. Morfometria radiográfica e tomográfica de dígitos de bovinos e bubalinos. **Dissertação**. Universidade Federal de Goiás. Curso de Pós-Graduação em Ciência Animal. 63p.

R-Fácil (2014)

URL <http://r-facil.webnode.com/>

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/> .