Escola Superior de Agricultura Luiz de Queiroz - USP
LCE5859 - Métodos Estatísticos Multivariados
Eduardo Elias Ribeiro Junior (edujrrib@gmail.com)
24 de fevereiro de 2017

Lista de Exercícios

Resolução dos exercícios propostos durante a disciplina LCE5859 - Métodos Estatísticos Multivariados, ministrada pelo professor Carlos Tadeu dos Santos Dias pelo Programa de Pós-Graduação em Estatística e Experimentação Agronômica da ESALQ-USP como curso de verão. Os exercícios são descritos no livro *Métodos Estatísticos Multivariados: uma introdução* (MANLY, 2008), que também é adotado como livro-texto na disciplina. Todas as análises presentes nesse documento são realizadas com o software R (R CORE TEAM, 2016) e estão disponíveis (texto e códigos) no endereço `https://github.com/jreduardo/lce5859-mem`. Os dados utilizados foram carregados do pacote `labestData` (PET ESTATÍSTICA UFPR, 2016), que mantém digitado e documentado todos os dados presentes no livro-texto.

Sumário

1 Testes de Significância com Dados Multivariados **2**

2 Medindo e Testando Distâncias Multivariadas **5**

3 Análise de Componentes Principais **7**

4 Análise de Fatores **8**

5 Análise de Função Discriminante **11**
5.1 Regressão Logística 12

6 Análise de Agrupamentos **14**

7 Análise de Correlação Canônica **15**

8 Escalonamento Multidimensional **17**

Referências **19**

1 Testes de Significância com Dados Multivariados

Esse exercício refere-se à aplicação dos métodos apresentados no capítulo 4 do livro-texto. O conjunto de dados disponibilizado diz respeito à comparação entre cães pré-históricos da Tailândia e outros quatro grupos de animais (cães modernos da Tailândia, chacais dourados, cuons e lobos indianos) em termos de nove medidas de mandíbula:

- X_1 Comprimento da mandíbula (mm).
- X_2 Largura da mandíbula, abaixo do primeiro molar (mm).
- X_3 Largura do côndilo auricular (mm).
- X_4 Altura da mandíbula, abaixo do primeiro molar (mm).
- X_5 Comprimento do primeiro molar (mm).
- X_6 Largura do primeiro molar (mm).
- X_7 Comprimento do primeiro ao terceiro molar (mm).
- X_8 Comprimento do primeiro ao quarto pré-molar (mm).
- X_9 Largura do canino inferior (mm).

Por simplicidade iremos considere a abreviação Pré-históricos, Modernos, Chacais, Cuons e Indianos para os grupos cães pré-históricos da Tailândia, cães modernos da Tailândia, chacais dourados, cuons e lobos indianos respectivamente.

Nesse conjunto de dados há também a informação sobre o sexo de todos os cães, exceto os pré-históricos. Foram avaliados 77 cães, cujo 10 pertenciam ao grupo dos cães Pré-históricos, 16 ao grupo dos cães Modernos, 20 ao grupo dos cães Chacais, 17 ao grupo dos cães Cuons e 14 ao grupo cães Indianos. Na Figura 1 são apresentados o comportamento das 9 medidas de mandíbulas para cada grupo canino (à esquerda) e a matriz de distâncias multivariadas (à direita). Observe que a distribuição marginal empírica das variáveis X_i, $i = 1,2,\ldots,9$ difere em cada grupo canino tanto em posição (medianas diferentes) quanto em dispersão (amplitudes diferentes). No gráfico à direita são exibidas as distâncias entre os vetores médios dos grupos caninos, calculadas como

$$d(i,j) = \sqrt{\sum_{k=1}^{9} (\bar{x}_{ki} - \bar{x}_{kj})^2} \tag{1}$$

sendo $\bar{x}_{ki}$ a média da k-ésima medida para o i-ésimo grupo, i e j variam de 1 a 5 conforme os cinco grupos caninos. Portanto $d(i,j)$ representa a distância multivariada entre os grupos caninos i e j em mm. Observa-se que a maior diferença entre os vetores médios se dá entre os grupos Chacais e Indianos, enquanto que a menor se dá entre Pré-históricos e Modernos.

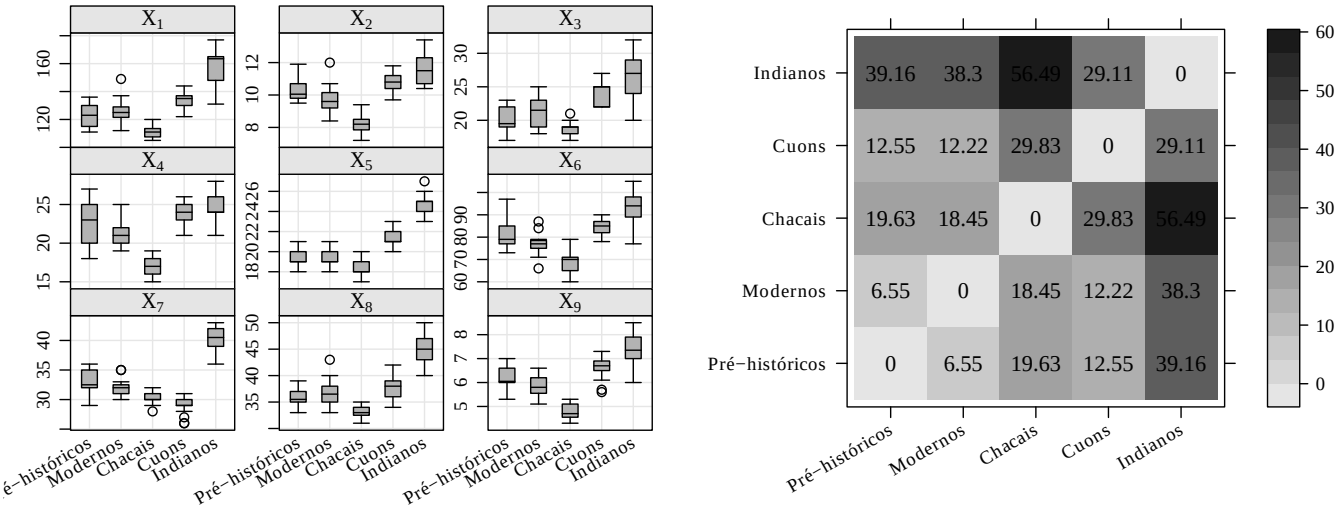


Figura 1: Box-plots das nove medidas de mandíbula para cada grupo canino (esquerda) e matriz de distâncias multivariadas entre os grupos caninos (direita).

Um dos interesses levantados sobre esse estudo é em testar as diferenças entre os vetores médios dos grupos caninos a fim de identificar, principalmente se há diferenças entre os cães Pré-históricos e os demais. Uma representação do perfil médio de cada grupo canino é apresentada na Figura 2 em forma de gráficos de radar onde a magnitude da média de cada medida é apresentada de forma conjunta para cada grupo, essa é uma forma alternativa aos gráficos-estrela descritos em MANLY (2008, cap. 3). A leitura dessa figura é análoga a leitura da representação da matriz de distâncias na Figura 1.

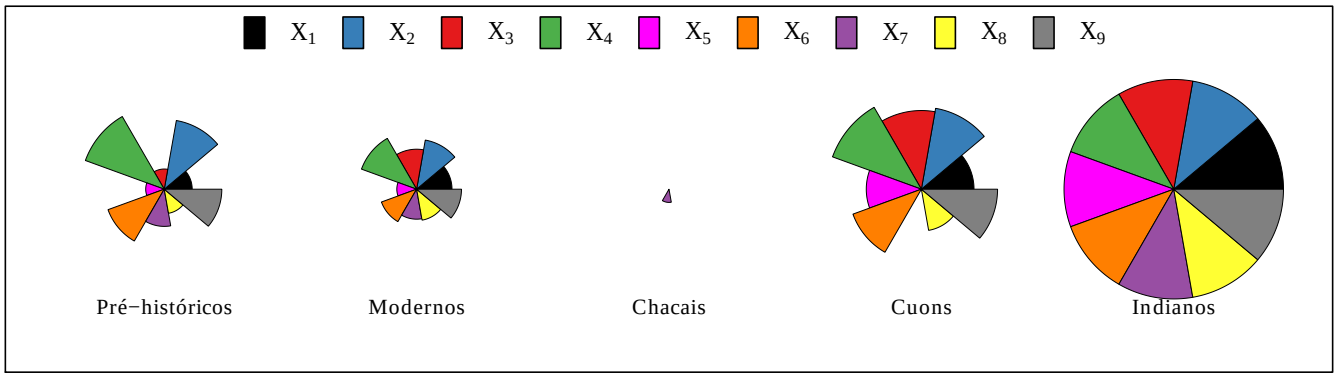


Figura 2: Gráficos-radar representando as médias de cada medida da mandíbula para cada grupo canino.

O método para comparação entre vetores médios quando se tem mais de duas amostras é denominado MANOVA (*Multivariate Analysis of Variance*) e o teste de significância para H_0: $\underline{\mu_i} = \underline{\mu_j}\ \forall i \neq j$ pode ser realizado via diferentes estatísticas . Nesse trabalho consideramos as estatísticas Traço de Pillai-Bartlett, Traço de Hotelling-Lawley, Lambda de Wilks e Raiz máxima de Roy tanto para MANOVA como para contrastes multivariados, conforme descrito em JOHN FOX (2011).

Na análise de variância multivariada para comparação de médias para várias amostas, o teste pressupõe a normalidade multivariada dentro dos grupos e homogeneidade nas matrizes de variâncias e covariâncias. Para verificar a normalidade multivariada procedeu-se com o teste de Mardia com correção para pequenas amostras (**???**). Os resultados são apresentados na Tabela 1. Note que não há sérias evidências de assimetria para todos os grupos, pois a hipótese testada é a de não assimetria da distribuição, assim para todos os grupos obteve-se níveis descritivos superiores a 0.20. Quando consideramos a curtose da distribuição, os dados para os grupos *Pré-históricos* e *Indianos* apresentaram níveis descritivos do teste que sustentam a rejeição da hipótese nula, que neste caso é de que a curtose da distribuição é coerente com a distribuição Normal multivariada.

Tabela 1: Teste de Mardia para normalidade multivariada. Estatísticas de teste e respectivos níveis descritivos entre parênteses.

	Assimetria	Curtose
Pré-históricos	165.000 (0.4854)	-2.023 (0.0431)
Modernos	178.214 (0.2280)	-1.315 (0.1885)
Chacais	163.021 (0.5290)	-1.464 (0.1431)
Cuons	174.550 (0.2903)	-1.585 (0.1130)
Indianos	164.131 (0.5045)	-1.824 (0.0682)

A segunda suposição, de matrizes de variâncias e covariâncias iguais para todos os grupos, foi avaliada pelo teste de M de Box (MANLY, 2008) que resultou na estatística 263.213 com um *p-valor* de 5.168×10^{-5} o que evidência a rejeição da hipótese nula de matrizes de variâncias e covariâncias homogêneas.

Claramente nota-se que algumas suposições não foram atendidas. Aqui dar-se-á continuidade ao estudo deixando de lado as suposições. Algumas estatísticas são mais robustas a falta de normalidade e a heterogeneidade das matrizes de variâncias e covariâncias, como o traço de Pillai e portanto devem ser preferíveis, porém todos serão prejudicadas.

Tabela 2: Tabela de análise de variância multivariada global (MANOVA)

	Df	test stat	approx F	num Df	den Df	Pr(>F)
Pillai	4	2.5892	13.6622	36	268.00	1.812E-42
Wilks	4	0.0022	27.6656	36	241.57	1.204E-66
Hotelling-Lawley	4	25.1290	43.6268	36	250.00	5.501E-88
Roy	4	16.3476	121.6990	9	67.00	5.864E-38

Na Tabela 3 são apresentados os resultados da análise de variância multivariada (MANOVA) com as diferentes estatísticas citadas anteriormente. Note que todas as estatísticas indicam que há um pelo menos um par de vetores de médias diferentes dentro os 10 pares possíveis (5 grupos). Esse resultado já era esperado, pois como apresentado nas análises descritivas, há grupos bastante distintos.

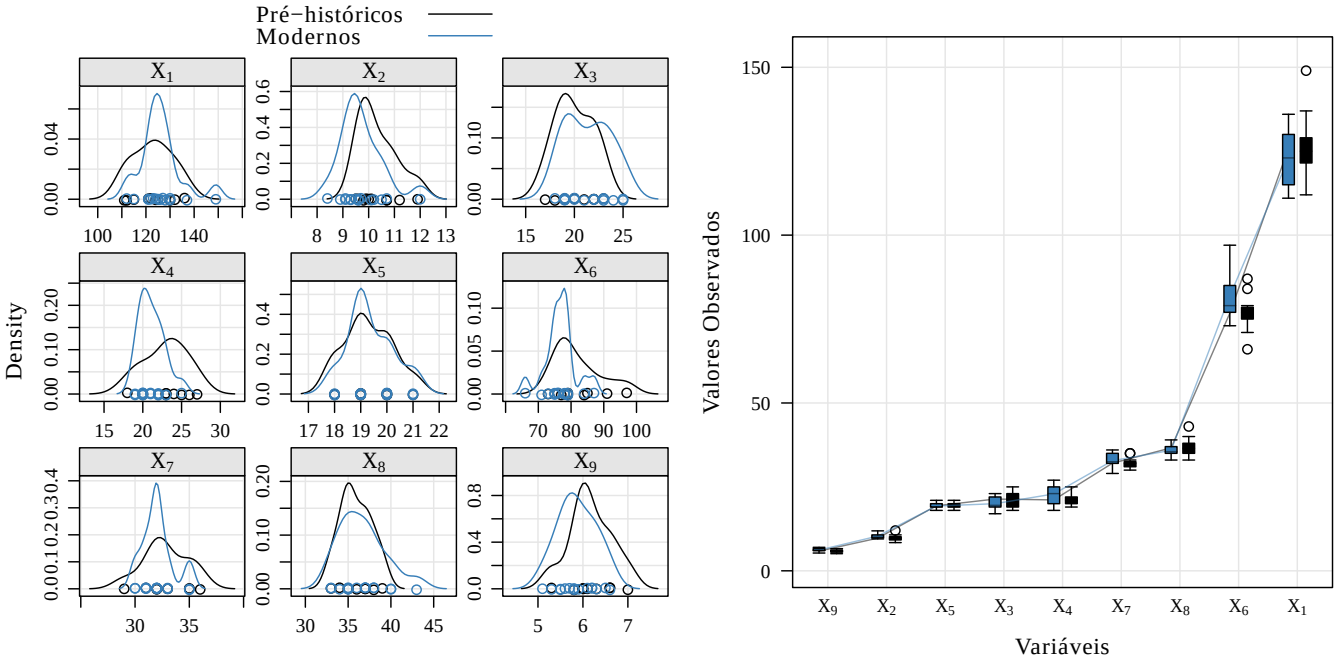
Na Tabela 3 são apresentados os resultados dos testes para contrastes multivariados de médias do grupo *Pré-histórico* contra os demais dois a dois. Note que os testes acusaram diferenças em todos os contrastes, indicando que o vetor das médias das medidas para os cães *Pré-históricos* é diferente de todos os demais, porém com um nível

descritivo demasiadamente menor quando contrastado com as médias calculadas para os cães modernos. Todavia vale ressaltar que esses resultados foram influenciados pela não verificação dos pressupostos.

Tabela 3: Tabela de contrastes multivariados

Contrast	Statistic	Df	test stat	approx F	num Df	den Df	Pr(>F)
Pré-Históricos = Modernos	Pillai	1	0.3741	4.2505	9	64.00	2.388E-04
	Wilks	1	0.6259	4.2505	9	64.00	2.388E-04
	Hotelling-Lawley	1	0.5977	4.2505	9	64.00	2.388E-04
	Roy	1	0.5977	4.2505	9	64.00	2.388E-04
Pré-Históricos = Chacais	Pillai	1	0.7271	18.9480	9	64.00	6.157E-15
	Wilks	1	0.2729	18.9480	9	64.00	6.157E-15
	Hotelling-Lawley	1	2.6646	18.9480	9	64.00	6.157E-15
	Roy	1	2.6646	18.9480	9	64.00	6.157E-15
Pré-Históricos = Cuons	Pillai	1	0.8705	47.7834	9	64.00	4.990E-25
	Wilks	1	0.1295	47.7834	9	64.00	4.990E-25
	Hotelling-Lawley	1	6.7195	47.7834	9	64.00	4.990E-25
	Roy	1	6.7195	47.7834	9	64.00	4.990E-25
Pré-Históricos = Indianos	Pillai	1	0.8112	30.5621	9	64.00	6.711E-20
	Wilks	1	0.1888	30.5621	9	64.00	6.711E-20
	Hotelling-Lawley	1	4.2978	30.5621	9	64.00	6.711E-20
	Roy	1	4.2978	30.5621	9	64.00	6.711E-20

A fim de avaliar o quão próximos foram os dados mensurados para os cães *Pré-históricos* e *Modernos* a **??** é apresentada. Nessa figura a distribuição de densidade empírica das nova variáveis para os dois grupos é exibida à esquerda. Note que há grande similaridade das distribuições, com algumas exceções como para X_1 (comprimento da mandíbula) que tem variabilidade maior para os *Pré-históricos*. O mesmo ocorre para X_4, X_6 e X_7. À direita da figura têm-se um gráfico de perfil, onde box-plots das nove medidas são apresentados para cada grupo, unindo as médias por linhas. Analogamente ao gráfico anterior nota-se para X_6 (largura do primeiro molar) a maior diferença entre as medidas. Nesse gráfico é nítido que as médias calculadas nos dois grupos são bastante próximas o que evidência que os resultados apontados pelos testes de contrastes não é confiável devido a não verificação dos pressupostos.



2 Medindo e Testando Distâncias Multivariadas

Esse exercício é descrito ao final do capítulo 5 do livro-texto, onde propõe-se a análise de um conjunto de dados que traz o registro de 4 variáveis ambientais e de proporções gênicas de *Fósforo Glucose-Isomerase* (Pgi) para 6 diferentes tipos genéticos de Pgi. Os registrados foram feitos em 16 colônias de borboletas *Euphydryas editha* na Califórnia e Oregon. Uma breve descrição das variáveis é realizada abaixo:

- Variáveis ambientais:
  - X_{A1}: Altitude (pés);
  - X_{A2}: Precipitação anual (polegadas);
  - X_{A3}: Temperatura máxima (°F);
  - X_{A4}: Temperatura mínima (°F).
- Proporções de mobilidade gênica Pgi:
  - X_{G1}: Para o tipo genético de Pgi 0.40;
  - X_{G2}: Para o tipo genético de Pgi 0.60;
  - X_{G3}: Para o tipo genético de Pgi 0.80;
  - X_{G4}: Para o tipo genético de Pgi 1.00;
  - X_{G5}: Para o tipo genético de Pgi 1.16;
  - X_{G6}: Para o tipo genético de Pgi 1.30;

Para as proporções de mobilidade gênica Pgi $\sum X_{Gi} = 1$. Na Figura 3 são apresentados dois gráficos que descrevem o comportamento das variáveis ambientais e genéticas. No gráfico da esquerda as variáveis ambientais são dispostas em gráficos de dispersão aos pares. Observa-se que as colônias GH e GL são as que mais diferem das demais colônias, sendo essas colônias de alta altitude e baixas temperaturas. No gráfico a direita são apresentadas as proporções acumuladas das frequências gênicas registradas em cada colônia. Destes perfis gênicos as colônias que mais se destacam são a GH e LO, que acumulam uma proporção maior de mobilidade gênica nos primeiros tipos de Pgi (0.4, 0.6 e 1.0).

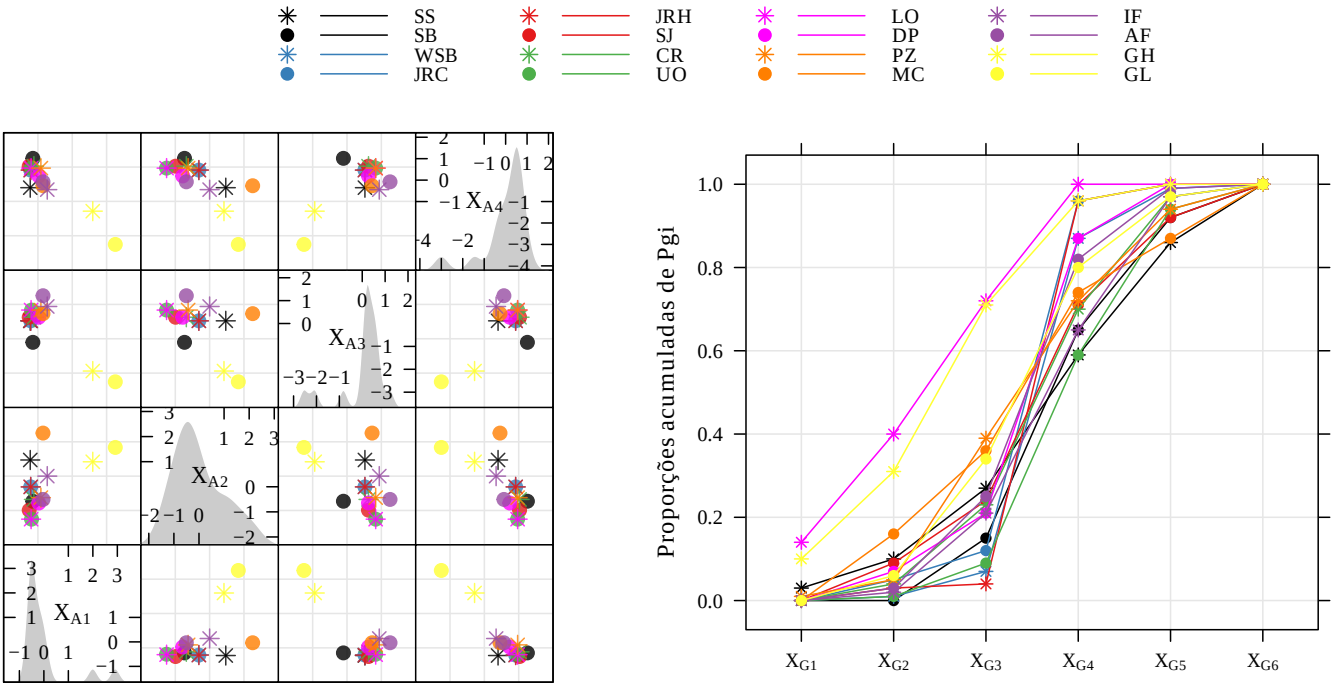


Figura 3: Gráfico de dispersão por pares entre as variáveis ambientais (esquerda). Proporções acumuladas de mobilidade gênica dos cinco diferentes tipos genéticos de Pgi.

O interesse nesse estudo é avaliar o relacionamento das variáveis ambientais e genéticas. Para tal trabalhar-se-á com as matrizes de distâncias para os dois conjuntos de variáveis. Para as variáveis ambientais as distâncias serão calculadas conforme Equação 1. Já para as variáveis genéticas, que tem a restrição de soma 1, será utilizado a seguinte métrica de distância

$$d(i,j) = \frac{1}{2}\sum_{k=1}^{6} |p_{ki} - p_{kj}| \tag{2}$$

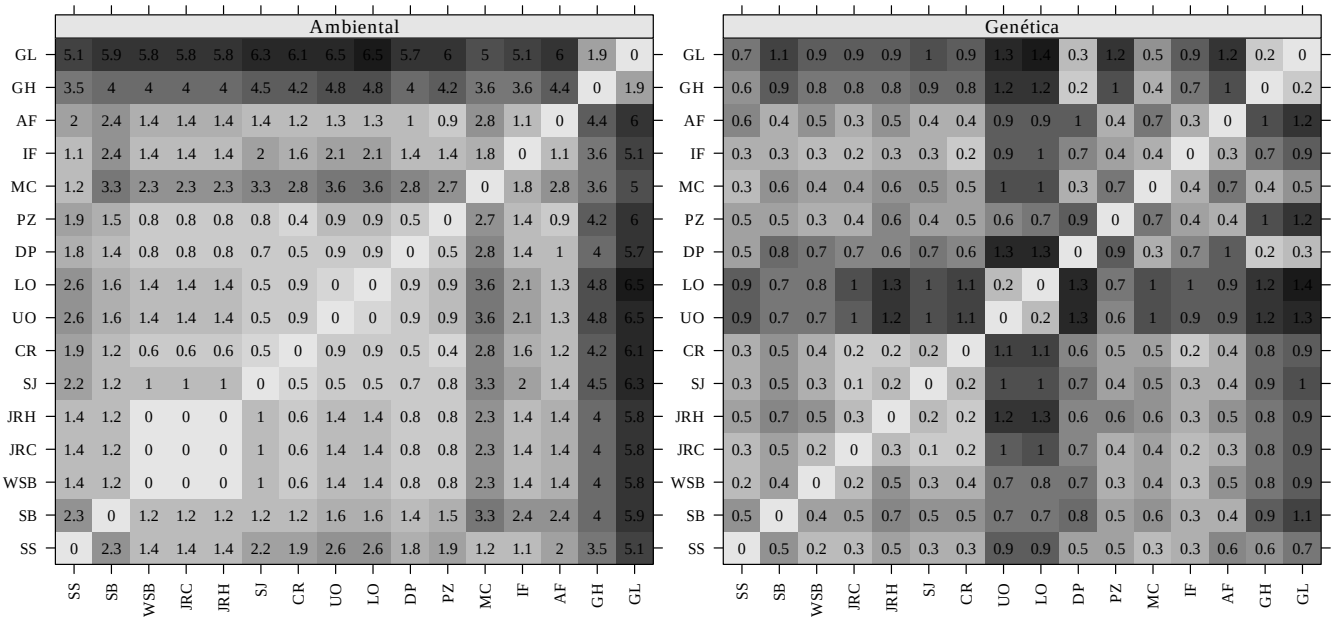


Figura 4: Matrizes de distâncias considerando as variáveis ambientais (esquerda) e genéticas (direita).

As matrizes de distância entre as colônias considerando as variáveis ambientais e genéticas são exibidas na Figura 4. Note que, assim como já observado na Figura 3, as colônias GH e GL obtiveram as maiores distâncias. Agora considerando a matriz de distâncias genéticas, temos as colônias LO e UO como as mais distantes.

Em posse das matrizes de distâncias entre as colônias para o conjunto de variáveis ambientais e genéticas, procedeu-se com a realização do teste de aleatorização matricial de Mantel a fim de avaliar se há correlação positiva entre as distâncias ambientais e genéticas. O procedimento do teste é basicamente i) aleatorizar os valores das matrizes e ii) calcular o coeficiente de correlação entre as distâncias aleatorizadas. Os passos i) e ii) são repetidos N vezes para se obter uma distribuição empírica das correlações sob a hipótese de correlação nula e por fim a correlação calculada com base nas distâncias originais é confrontada com essa distribuição.

A correlação entre as matrizes de distâncias ambientais e genéticas foi de 0.435. Foram realizadas 1000 permutações das matrizes e calculadas as correlações, cujo o quantil de 95% foi de 0.318. Como $0.435 > 0.318$ há fortes evidências de que as distâncias ambientais estejam positivamente correlacionadas com as distâncias genéticas.

Conforme análises apresentadas anteriormente mostra-se que há correlação positiva à 5% entre as variáveis ambientais e genéticas. Porém nessa análise foi negligenciada a posição espacial de cada colônia. Essa informação pode estar associada aos resultados obtidos, uma vez que as variáveis genéticas podem ser hereditárias e o processo migratório das borboletas pode ter ocorrido entre colônias.

3 Análise de Componentes Principais

Esse exercício faz referência as técnicas apresentadas no capítulo 6 do livro-texto. Os dados descrito no exercício é referente a um estudo sobre taças de cerâmica escavadas de lugares pré-históricos na Tailândia. Foram 6 medidas feitas em cada taça. Ao todo foram mensuradas as medidas em 25 taças. A natureza das medidas pode ser vista em MANLY (2008 pág.102, Figura 6.3).

O objetivo nesse estudo é caracterizar as taças identificando grupos com características similares e identificar, se houverem, taças incomuns. O método apresentado nesse capítulo para atingir os objetivos do estudo é via análise de componentes principais.

A construção dos componentes principais foi realizada via matriz de correlação para que a amplitude de variação das variação não influencie fortemente na análise. Os resultados da análise são apresentados na Figura 5.

O primeiro gráfico, superior à esquerda, da Figura 5 auxilia a escolha que quantas componentes serão necessárias para explicar os dados. Note que com duas componentes já explica-se 89% da variância dos dados, o que já é bastante satisfatório. Outro critério usualmente adotado é escolher tantas componentes quanto fores os autovalores maiores que 1, nesse caso esse critério também leva a escolha de apenas duas componentes (autovalores 4.272, 1.092).

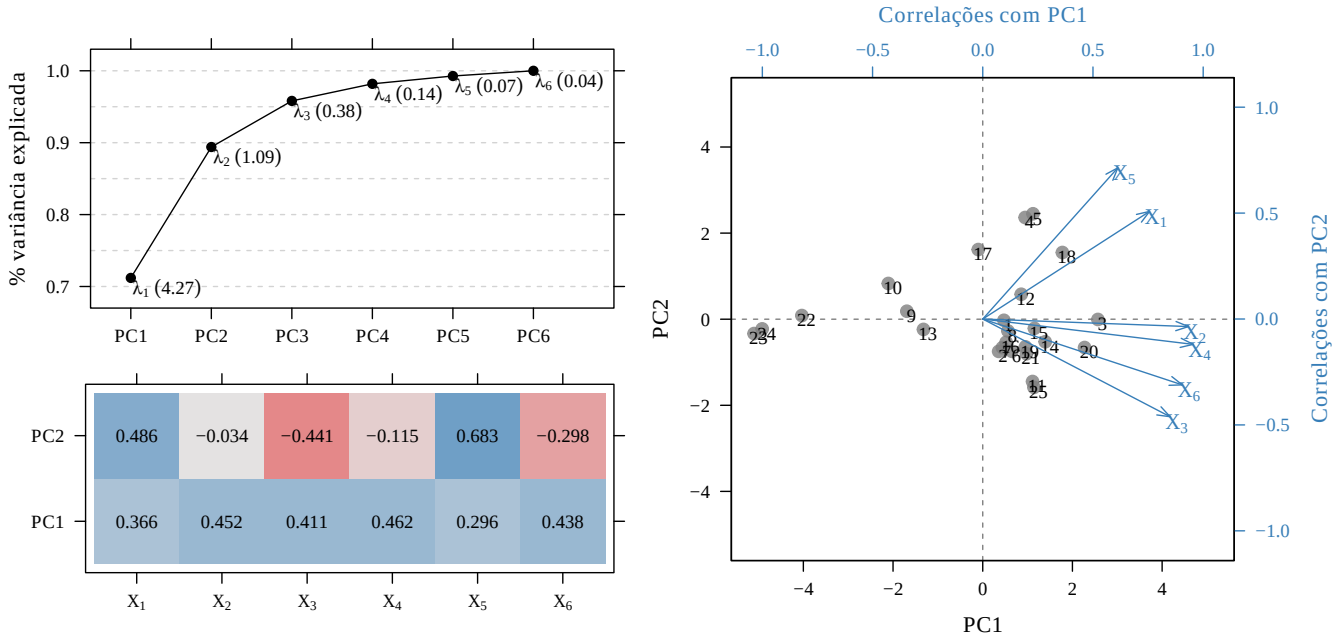


Figura 5: Proporção acumulada da variância por cada componente com apresentação dos autovalores (superior à esquerda). Matriz de carregamentos, auto vetores, associados as 2 primeiras componentes (inferior à esquerda). Biplot (à direita), dispersão dos escores calculados com base nas 2 primeiras componentes e correlação das variáveis originais com as componentes.

Já no segundo gráfico, inferior à esquerda da Figura 5, têm-se os coeficientes da combinação linear das variáveis originais que compõem as primeira e segunda componentes. Note que a primeira componente é basicamente a soma de todas as variáveis originais, a partir da definição das variáveis em MANLY (2008 pág.102, Figura 6.3), pode-se interpretar essa variável como o *tamanho* da taça, uma vez que valores altos definem uma taça grande e baixo uma taça menor. Para a segunda componente temos basicamente um contraste de $X_1 + X_5$ contra $X_3 + X_6$. X_1 e X_5 são medidas horizontais que definem a largura da taça, enquanto que X_3 e X_6 são medidas verticais que definem altura. Assim essa componente pode ser interpretada como *forma*, onde taças mais "gordinhas" recebem valores altos e as mais "magrinhas" valores baixos.

A última visualização apresentada na Figura 5 é o gráfico *biplot* (HOLLAND, 2008). Nesse gráfico os escores de cada taça, calculados a partir das componentes principais, são apresentados em um gráfico de dispersão conjuntamente com a correlações entre as componentes e as variáveis originais, que são representadas pelos vetores com valores em um eixo adicional. Observe que as taças 23, 24 e 22 se destacam pelos valores baixos na primeira componente, isso as caracteriza como taças baixas e com relação a segunda componente têm-se um valor mediano para as três, ou seja, são taças pequenas e com formato mediano. De forma geral nota-se que os escores da primeira componente são bem mais variáveis que o da segunda, isso mostra que o tamanho das taças é mais variado do que a forma. Para as correlações entre as variáveis originais têm-se a mesma interpretação realizada a partir da matriz de carregamentos. Valores altos da primeira componente estão relacionados a valores altos de X_1, X_2, X_3, X_4, X_5 e X_6 (correlações 0.76, 0.93, 0.85, 0.95, 0.61, 0.91). A segunda componente apresenta correlação negativa com X_2, X_3, X_4 e X_6 e positivas com X_1 e X_5.

4 Análise de Fatores

Nesta seção será abordado o método de análise fatorial ou análise de fatores, descrito no capítulo 7 do livro-texto. Os dados apresentados para análise trazem estimativas do consumo médio de proteínas de 9 diferentes fontes de alimentos para os habitantes de 25 países europeus. As variáveis foram codificadas da forma

- X_1: Consumo de proteínas provenientes de carne vermelha;
- X_2: Consumo de proteínas provenientes de carne branca;
- X_3: Consumo de proteínas provenientes de ovos;
- X_4: Consumo de proteínas provenientes de leite;
- X_5: Consumo de proteínas provenientes de peixe;
- X_6: Consumo de proteínas provenientes de cereais;
- X_7: Consumo de proteínas provenientes de carboidratos;
- X_8: Consumo de proteínas provenientes de grãos nozes e sementes oleaginosas; e
- X_9: Consumo de proteínas provenientes de frutas e vegetais.

Uma análise descritiva dos dados é fornecida na Figura 6. No gráfico à esquerda são apresentadas as correlações entre as nove variáveis, essa é a matriz de correlações amostrais a qual será denotada por R nessa seção. Note que, em geral, as variáveis apresentam fortes correlações com exceção de X_9 (frutas e vegetais) cujo a correlação mais expressiva foi de apenas -0.333 ocorrida com X_4 (leite). Para as demais variáveis a correlação maior correlação positiva foi entre X_6 e X_8 (0.636) o que indica altos consumos de proteínas de cereais é acompanhada de consumos altos de grãos, nozes e sementes oleaginosas. E a maior correlação negativa ocorre entre X_3 e X_6 indicando que valores de consumo altos de proteínas provenientes de ovos estão relacionados a valores baixo de consumo de proteínas provenientes de cerais e vice-e-versa.

A direita da Figura 6 são apresentadas as densidades empíricas estimadas para cada variável. Note que todas as densidades podem ser razoavelmente ajustadas por uma distribuição Normal, ou seja, seus valores estão razoavelmente dispostos em acima e abaixo da média. Para algumas variáveis nota-se uma bimodalidade mais acentuada, por exemplo X_2 (carne branca) e X_9 (frutas e vegetais), e outras apresentam certo grau de assimetria, por exemplo X_5 (peixe), X_6 (cereais) e X_7 (carboidratos).

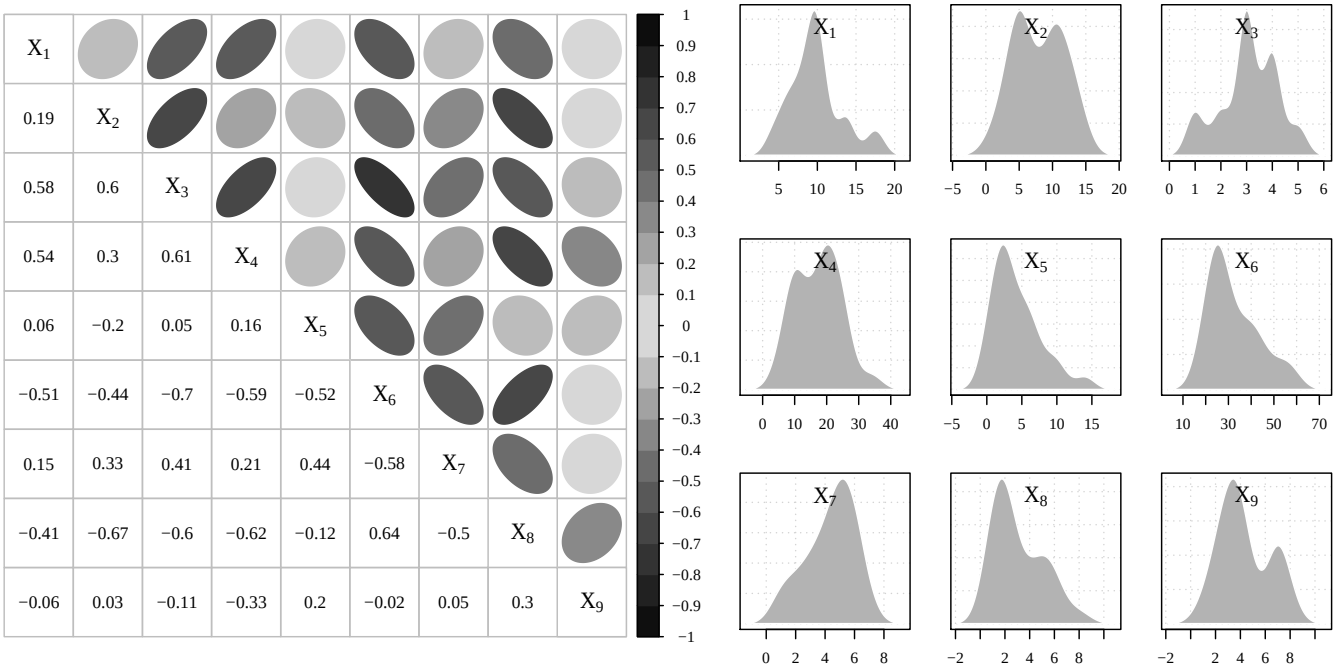


Figura 6: Representação da matriz de correlação dos dados (esquerda) e distribuições marginais empíricas das variáveis originais (direita).

A proposta deste exercício é a realização de uma análise fatorial a fim de identificar fatores importantes que descrevam as variáveis observadas além de avaliar o relacionamento entre os países.

O modelo fatorial foi ajustado via método método das componentes principais, cujo abordagem é via aproximação da matriz de correlações R por $LL^t + \Psi$, em que L é a matriz de carregamentos fatoriais e Ψ a matriz diagonais das variâncias específicas (JOHNSON; WICHERN, 2007). Dos nove fatores possíveis, permaneceu-se apenas com 3, pois não há um considerável aumento no percentual de variação explicada em cada variável com o acréscimo de um quarto fator. Ainda, para melhor interpretação do relacionamento entre as variáveis e os fatores, os fatores foram rotacionados. O modelo ajustado é apresentado na Equação 3.

$$
\begin{aligned}
\text{(Carne vermelha)} \quad & X_1 = +0.912\,F_1 + 0.067\,F_2 + 0.008\,F_3 + 0.037\,F_4 + \epsilon_1 \\
\text{(Carne branca)} \quad & X_2 = +0.150\,F_1 + 0.944\,F_2 - 0.079\,F_3 + 0.061\,F_4 + \epsilon_2 \\
\text{(Ovos)} \quad & X_3 = +0.662\,F_1 + 0.589\,F_2 + 0.126\,F_3 - 0.035\,F_4 + \epsilon_3 \\
\text{(Leite)} \quad & X_4 = +0.727\,F_1 + 0.220\,F_2 + 0.187\,F_3 - 0.421\,F_4 + \epsilon_4 \\
\text{(Peixes)} \quad & X_5 = +0.107\,F_1 - 0.226\,F_2 + 0.915\,F_3 + 0.098\,F_4 + \epsilon_5 \\
\text{(Cereais)} \quad & X_6 = -0.577\,F_1 - 0.412\,F_2 - 0.608\,F_3 - 0.005\,F_4 + \epsilon_6 \\
\text{(Carboidratos)} \quad & X_7 = +0.007\,F_1 + 0.501\,F_2 + 0.724\,F_3 - 0.005\,F_4 + \epsilon_7 \\
\text{(Grãos nozes e sementes)} \quad & X_8 = -0.368\,F_1 - 0.701\,F_2 - 0.273\,F_3 + 0.376\,F_4 + \epsilon_8 \\
\text{(Frutas e Vegetais)} \quad & X_9 = -0.057\,F_1 - 0.006\,F_2 + 0.118\,F_3 + 0.961\,F_4 + \epsilon_9
\end{aligned}
\tag{3}
$$

As comunalidades para X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9 são 0.838, 0.923, 0.802, 0.788, 0.909, 0.872, 0.776, 0.842, 0.94 respectivamente. Note que as comunalidades são todas altas indicando que muito da variação de X_i pôde ser explicada pelos 4 fatores comuns, para $i = 1,2,\ldots,9$. Uma avaliação da qualidade do modelo pode ser realizado através dos gráficos à esquerda da Figura 7. Note que os resíduos variam de -0.10 a 0.10 centrados e pouco dispersos em torno de 0 (boxplot superior) o que indica um bom ajuste . Na matriz de resíduos (heatmap inferior) pode-se verificar os resíduos para cada par de variáveis. Na diagonal da matriz são apresentadas as variâncias específicas (complementar das comunalidades) e da mesma forma indicam um bom ajuste.

Avaliando os coeficientes dos fatores no modelo expresso em (3) nota-se que nas equações para X_1, X_2, X_5 e X_9 há cargas bastante altas para somente um fator, sendo F_1 para X1; F_2 para X_2; F_3 para X_5; e F_4 para X_9 isso indica que a variação dessas variáveis é praticamente explicada por esse fator de maior carga. Isso também evidencia a necessidade de 4 fatores, uma vez que todos contribuem significativamente para explicação de alguma variável. Nas equação que descrevem as demais variáveis há uma composição de pelo menos 2 dos 4 fatores.

Ainda avaliando a Equação 3 é conveniente rotular os fatores, avaliando as cargas estimados de um fator para todas as variáveis. Por exemplo F_1 tem as maiores cargas para X_1 (carne vermelha), $X_3(ovos)$, X_4 (leite) e X_6 (cereais), coeficientes 0.912, 0.15, 0.727, -0.577 respectivamente. Obviamente que "dar nomes" aos fatores requer conhecimento prático dos dados analisados, pesquisadores não envolvidos com a pesquisa tem dificuldade nessa interpretação. Portanto, nesse trabalho não se rotulará os fatores.

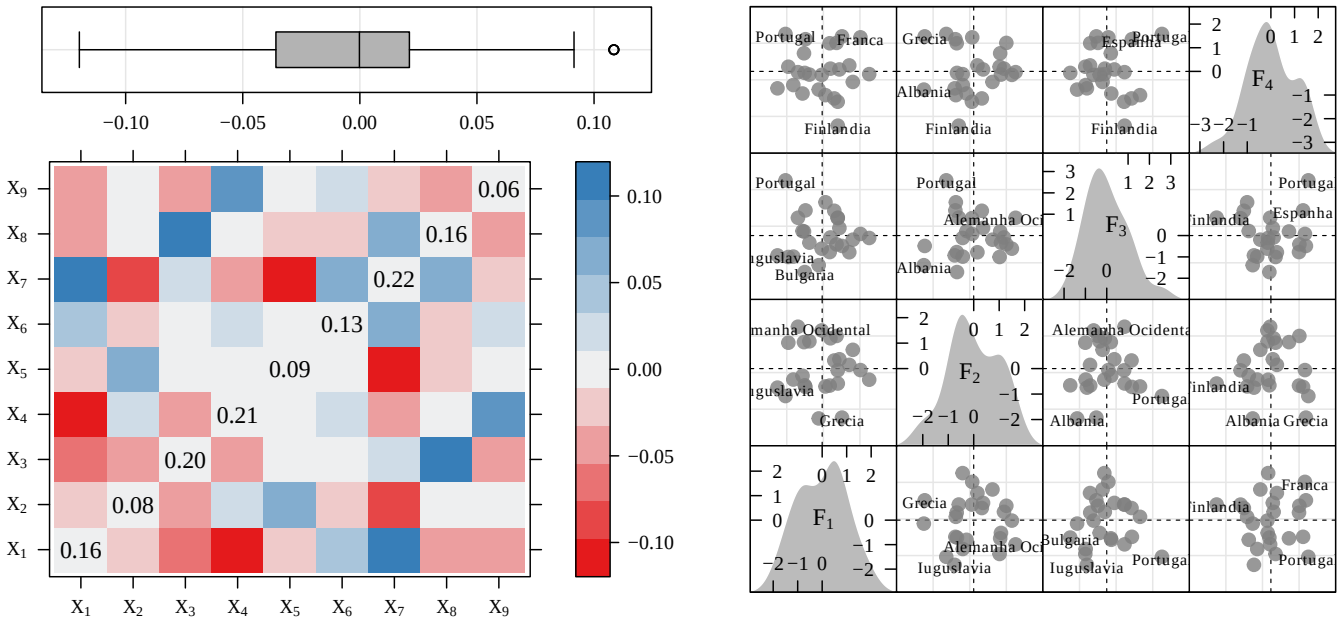


Figura 7: Avaliação da qualidade de ajuste do modelo. Distribuição dos resíduos calculados pela diferença entre R e $LL^t + \Psi$ (esquerda superior) e representação da matriz de resíduos (esquerda superior) e escores dos países com base nos quatro fatores obtidos (direita).

Na Tabela 4 são apresentados os escores calculados para cada um dos 25 países europeus. Com os valores disposto nessa tabela pode-se caracterizar a "habilidade", em termos de consumo de proteínas, de cada país através de seus escores obtidos. Uma visualização dos escores dispostos na Tabela 4 é realizada no gráfico à direita na Figura 7, em que os escores dos fatores são apresentados dois a dois em gráficos de dispersão. Linhas que delimitam os quadrantes são úteis para interpretação. Para todos os gráficos de dispersão os 3 países mais distantes dos demais tem seus nomes apresentados. Note que, quanto ao consumo de proteínas, não há nenhum país fortemente

discrepante da maioria, isso pode ser observado pelas densidades empíricas apresentadas para os escores de cada fator.

Tabela 4: Escores dos fatores rotacionados para 23 países europeus.

	Fator 1	Fator 2	Fator 3	Fator 4
Albania	-0.1409	-1.9569	-1.3883	-0.7692
Austria	-0.0191	1.4986	-0.6139	-0.1489
Belgica	0.7035	0.3665	0.3553	0.0822
Bulgaria	-0.7111	-0.6564	-1.7097	-0.0637
Tchecoslovaquia	-0.5213	1.0713	-0.3318	-0.1351
Dinamarca	0.4936	0.3187	1.1617	-1.1435
Alemanha Ocidental	-0.9892	1.6441	0.8334	-0.0320
Finlandia	0.6380	-0.5857	0.8393	-2.2952
Franca	1.5625	-0.0579	0.0764	1.4395
Grecia	0.8108	-1.9235	-0.4867	1.4805
Hungria	-1.3813	1.0161	-1.0040	0.1967
Irlanda	1.2518	0.7366	-0.1932	-0.4466
Italia	0.3076	-0.6666	-0.7738	1.1891
Paises Baixos	0.5869	1.2812	-0.4117	1.2038
Noruega	0.1348	-0.7008	1.5600	-1.0048
Polonia	-0.7418	1.0393	0.1919	0.7649
Portugal	-1.4972	-1.0775	2.5875	1.5833
Romenia	-1.1788	-0.4300	-0.9854	-0.5781
Espanha	-0.6810	-0.7196	1.1794	1.3608
Suecia	0.6349	-0.0693	0.8147	-1.2836
Suica	1.1066	0.1460	-0.7824	0.2526
Reinou Unido	1.9201	-0.4360	-0.1135	-0.1162
USSR	-0.8024	-0.2800	0.2127	-0.9356
Alemanha Oriental	0.3377	1.1886	-0.0891	0.1159
Iuguslavia	-1.8247	-0.7468	-0.9287	-0.7169

5 Análise de Função Discriminante

Essa seção é dedicada a realização do exercício proposto ao final do capítulo 8 do livro-texto. O exercício propõe a realização de uma análise discriminante a fim de separar grupos caninos com base em nove medidas de mandíbula. Os dados são os mesmos utilizados na Seção 1. São 77 observações, cujo 10 pertenciam ao grupo dos cães Pré-históricos, 16 ao grupo dos cães Modernos, 20 ao grupo dos cães Chacais, 17 ao grupo dos cães Cuons e 14 ao grupo cães Indianos. As nove variáveis mensuradas são:

- X_1 Comprimento da mandíbula (mm).
- X_2 Largura da mandíbula, abaixo do primeiro molar (mm).
- X_3 Largura do côndilo auricular (mm).
- X_4 Altura da mandíbula, abaixo do primeiro molar (mm).
- X_5 Comprimento do primeiro molar (mm).
- X_6 Largura do primeiro molar (mm).
- X_7 Comprimento do primeiro ao terceiro molar (mm).
- X_8 Comprimento do primeiro ao quarto pré-molar (mm).
- X_9 Largura do canino inferior (mm).

A Figura 8 apresenta os resultados da análise discriminante. Como são 5 grupos e o número de variáveis são nove têm-se 4 variáveis canônicas obtidas (LD1, LD2, LD3, LD4). A densidade de probabilidade empírica estimada para cada variável canônica estratificando por grupo canino é apresentada à esquerda da Figura 8. Note que nesse gráfico é claro a atuação de cada variável canônica na tarefa de discriminação. A LD1 discrimina muito bem o grupo dos *Cuons* dos demais; LD2 discrimina o grupo dos *Indianos*; LD3 discrimina o grupo dos *Pré-históricos* do grupo dos *Chacais*; e LD4 tem pouco poder discriminativo. Complementar a visualização das densidades empíricas, um gráfico de dispersão entre a primeira e segunda variável canônica. Nota-se claramente a discriminação dos grupos *Indianos* e *Cuons* na dimensão dessas duas variáveis. Nesse também pode-se observar que as únicas classificações incorretas da análise discriminante ocorreram para observações dos grupos *Pré-históricos* e *Modernos*, cujo nessa dimensão estão bastante sobrepostos.

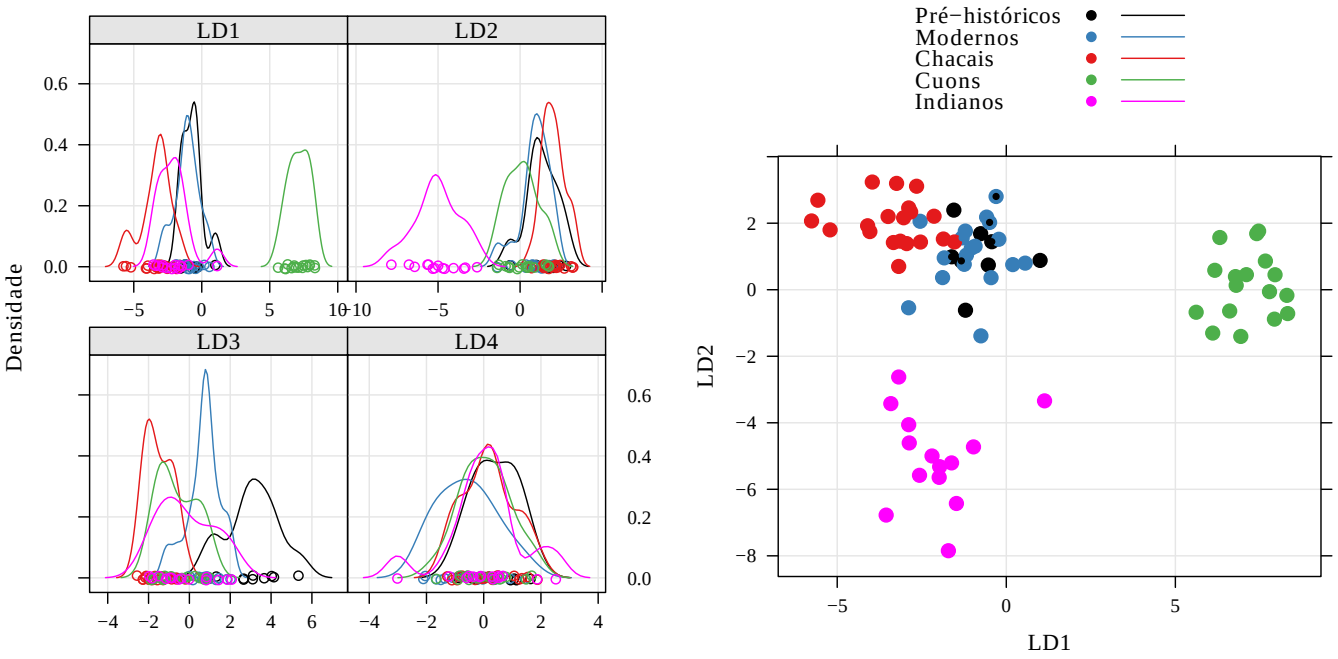


Figura 8: Densidades empíricas estimadas para as quatro variáveis canônicas estratificando pelos 5 grupos caninos (esquerda) e gráfico de dispersão das duas primeiras componentes (direita). No gráfico de dispersão o preenchimento dos pontos representa o grupo real e o contorno o grupo predito.

Para possibilitar uma interpretação das variáveis canônicas, além da discriminação dos grupos a Tabela 4 apresenta as correlações das variáveis originais com as variáveis canônicas. Observa-se que a variável LD1 tem correlações positivas, mas não fortes, com todas as variáveis originais exceto X_7 (comprimento do primeiro ao terceiro molar) então pode-se interpretá-la como tamanho da arcada dentária em contraste com o comprimento do 1º ao 3º molar; LD2 tem todas as correlações fortes e negativas o que pode ser interpretado como o inverso do tamanho da arcada dentária. Interpretações similares podem ser realizadas com as demais variáveis canônicas e novamente um conhecimento do estudo é imprescindível nessa etapa. Com as interpretações das variáveis canônicas LD1 e LD2 pode-se voltar ao gráfico de dispersão, à direita na **??** e interpretar a posição dos grupo. Por exemplo o

grupo dos *Cuons* apresentou valores altos de LD1 o que indica que são cães com arcada dentária grande quando contrastada com o comprimento do 1º ao 3º molar, já quando considerado a tamanho da arcada dentária como um todo nota-se que o grupo tem arcada dentária de um tamanho mediano (LD2 próximo de 0). Outro exemplo, no grupo dos *Indianos* é nítido que sua arcada dentária é maior que a dos outros grupos, uma vez que os valores de LD2 neste grupo são negativos e bem distantes de 0.

Tabela 5: Correlações entre as medidas de mandíbula originais e as quatro variáveis canônicas.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
LD1	0.196	0.409	0.366	0.475	0.246	0.351	-0.412	0.093	0.387
LD2	-0.926	-0.743	-0.741	-0.660	-0.962	-0.798	-0.874	-0.925	-0.815
LD3	0.135	0.409	0.011	0.481	0.003	0.327	0.233	0.121	0.334
LD4	-0.256	-0.056	-0.335	-0.089	0.092	0.144	0.022	-0.211	-0.083

Na Tabela 6 são apresentadas as classificações provenientes da análise discriminante, que são realizadas pela probabilidade a posteriori da observação pertencer a cada grupo calculada conforme Teorema de Bayes. Nessa tabela têm-se a predição na própria base de treino, o que potencialmente subestima o erro de classificação, e a predição quando considerada a abordagem *leave-one-out*, em que ajusta-se o modelo de classificação sem retira a i-ésima observação e posteriormente a classifica. Note que considerando ambas as matrizes cruzadas o modelo de classificação baseado nas funções lineares discriminantes de Fisher foram bastante satisfatórias.

Tabela 6: Matrizes de confusão (grupos reais vs grupos preditos). Preditos considerando a base de treino (esquerda) e considerando a abordagem *leave-one-out* (direita).

	Predito base					Predito *leave-one-out*				
Grupo real	Pré-históricos	Modernos	Chacais	Cuons	Indianos	Pré-históricos	Modernos	Chacais	Cuons	Indianos
Pré-históricos	8	2	0	0	0	7	3	0	0	0
Modernos	0	16	0	0	0	3	12	1	0	0
Chacais	0	0	20	0	0	0	0	20	0	0
Cuons	0	0	0	17	0	0	0	0	17	0
Indianos	0	0	0	0	14	1	0	0	0	13

5.1 Regressão Logística

Essa subseção será bastante breve e compreende a segunda proposta de exercício do capítulo 8 do livro-texto. Esse exercício propõe o ajuste de uma regressão logística para discriminar o sexo dos cães em função das medidas da mandíbula em cada grupo canino. Como são 9 medidas e poucas observações em cada grupo canino abordagens para seleção de variáveis deverão ser adotadas.

O modelo denominado regressão logístico é um modelo da classe dos modelos lineares generalizados, cujo a distribuição considerada para a relação condicional $Y \mid X$ é Binomial(m, π) e função de ligação logito (que dá nome ao modelo). Assim o modelo pode ser escrito da seguinte forma:

$$Y_i \mid \underline{x}_i \sim \text{Binomial}(m_i, \pi_i)$$
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \underline{x}_i^t \beta$$

em que Y_i é a variável aleatória dependente; $\underline{x}_i$ o vetor de covariáveis do i-ésimo indivíduo; m_i o número de ensaios de Bernoulli realizados na i-ésima unidade amostral; π_i a probabilidade de sucesso; e β os parâmetros do modelo a ser ajustado.

No exemplo dessa seção a dimensão de $\underline{x}_i$ é 9×1, todavia como temos poucas observações em cada grupo canino procedeu-se com o algoritmo *stepwise* com critério AIC para seleção de variáveis. Nesse algoritmo define-se o maior e menor modelo e o algoritmo irá, a partir do maior, retirar e recolar variáveis conforme menor AIC. O algoritmo foi executado definindo o menor modelo como o modelo sem covariáveis e o maior modelo aquele contendo as nove de forma aditiva.

Os resultados dos modelos ajustados são exibidos na Tabela 7, em que β_j representa o efeito estimado da variável X_j. Assim no grupo dos cães *Modernos*, permaneceram no modelo apenas os efeitos das variáveis X_7 e X_1, para os *Chacais* apenas os efeitos das variáveis X_6, X_9 e X_1 e assim por diante. Note que as estimativas dos efeitos para

os modelos dos grupos *Chacais* e *Indianos* são demasiadamente altas, isso ocorreu devido a probabilidades ajustadas como 1 e/ou 0, ou seja, nesse conjunto de variáveis há um hiperplano que separa perfeitamente machos de fêmeas. Isso é um problema estimação na borda do espaço paramétrico que causa não convergência do algoritmo de estimação. O único grupo em que não se pôde discriminar machos e fêmeas através da regressão logística foi o grupo dos cães *Cuons*, observe que nenhuma variável foi selecionada para esse grupo, ou seja, a probabilidade de cães fêmeas nesse grupo é 0.471 para qualquer cão independente de suas medidas da mandíbula.

Tabela 7: Estimativas dos parâmetros estimados dos modelos para cada grupo e seus respectivos quadros de análise de deviance sequencial.

Grupo	Parameter	Estimativa	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
Modernos	β_0	118.415			15	22.181	
	β_7	-2.548	1	7.279	14	14.902	0.0070
	β_1	-0.294	1	3.484	13	11.418	0.0620
Chacais	β_0	64723.457			19	27.726	
	β_6	-413.059	1	16.670	18	11.055	0.0000
	β_9	-1894.501	1	3.186	17	7.870	0.0743
	β_1	-240.995	1	7.870	16	0.000	0.0050
Cuons	β_0	-0.118			16	23.508	
Indianos	β_0	3884.440			13	19.121	
	β_2	-210.937	1	11.076	12	8.045	0.0009
	β_9	-211.338	1	8.045	11	0.000	0.0046

6 Análise de Agrupamentos

Essa seção refere-se ao capítulo 9 do livro-texto. O exercício propõe a realização de uma análise de agrupamentos. Os dados apresentados no exercício mostram as quantidades das 25 espécies de plantas mais abundantes em 17 lotes de um prado de pastagem em uma reserva natural na Suécia. Ao todo são 25×17 observações, ou seja, uma quantidade mensurada para cada combinação de espécie e lote.

Um descrição dos dados é apresentada na Figura 9 onde box-plots das medidas de abundância são construídos para cada espécie (à esquerda) e para cada lote (à direita). Note que em geral as distribuições das medidas são bastante assimétricas à direita, com muitos valores baixos e poucos altos, as exceções ficam por conta das espécies *Anemone nemorosa*, *Rumex acetosa* e *Veronica Chamaedrys* e do *Lote 6* que apresentam distribuições mais simétricas.

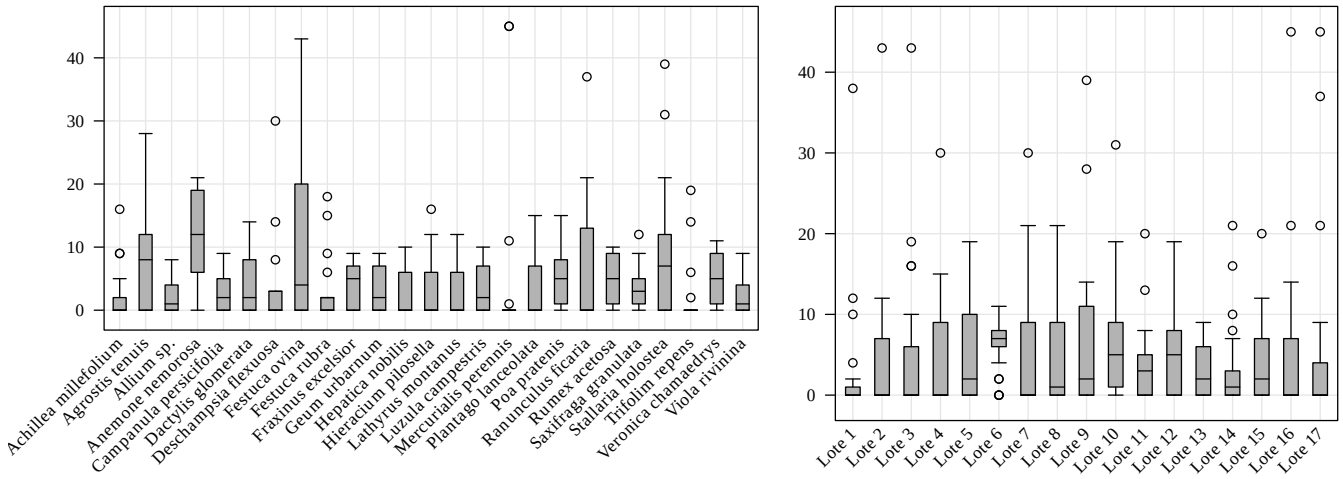


Figura 9: Box-plots das medidas de abundância para as 25 espécies (esquerda) e para os 17 lotes (direita).

A análise de agrupamento desses dados será realizada sob duas perspectivas. Na primeira agruparemos as espécies, avaliando as similaridades das medidas de abundâncias distribuídas nos 17 diferentes lotes. Sob a segunda perspectiva serão agrupados os lotes, a fim de identificar grupos de lotes que tem abundância de espécies similares.

Nessa análise realizou-se um agrupamento hierárquico pelo método de *Ward* (MURTAGH; LEGENDRE, 2014) a partir das distâncias euclidianas entre os vetores $\underline{x}_i$ ($i = 1,2,\ldots,25$ vetores de tamanho 17 para o agrupamento de espécies e $i = 1,2,\ldots,17$ vetores de tamanho 25 para o agrupamento de lotes).

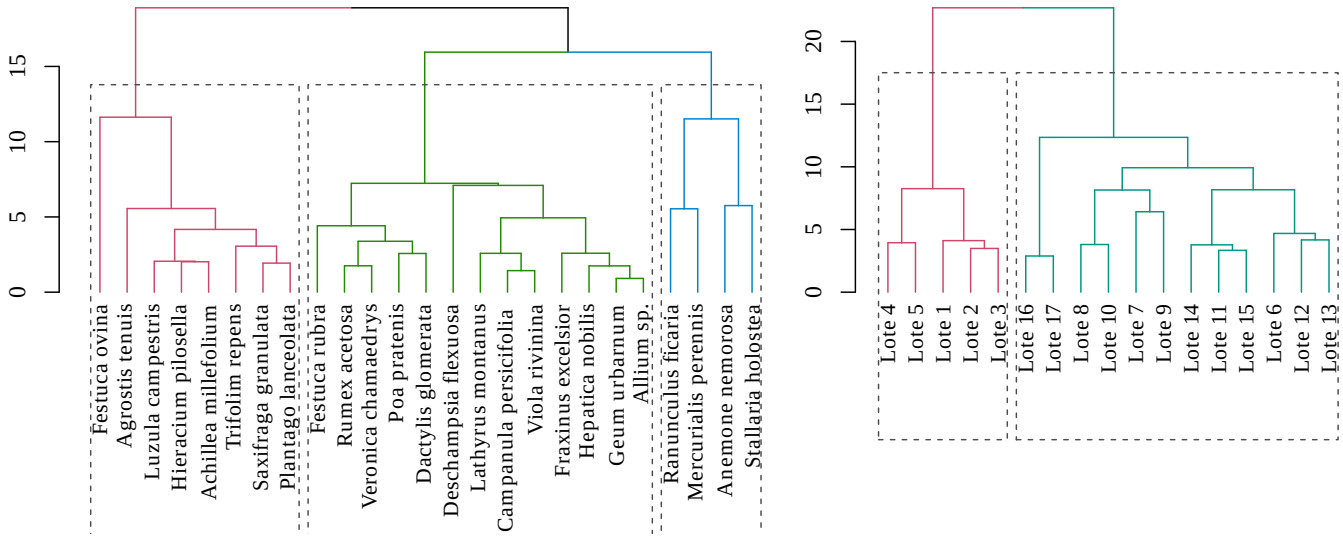


Figura 10: Dendrograma obtidos da análise de agrupamento hierárquico pelo método de Ward. Agrupamento de espécies (esquerda) e de lotes (direita).

Na Figura 10 são apresentados os resultados da análise de agrupamentos para as espécies (à direita) e para os lotes (à esquerda). Sem um apelo aplicado, neste trabalho optou-se pela adoção de um algoritmo automático para a definição do número de grupos. A definição do número de grupos deu-se pelos grupos formados na maior distância de ligação. Para espécies foram tomados três grupos e para os lotes dois.

7 Análise de Correlação Canônica

Esta seção apresenta a análise dos dados apresentados no exercício proposto ao final do 10º capítulo do livro-texto. Os dados reference a complementação daqueles analisados na Seção 4, relembrando são dados com estimativas do consumo médio de proteínas de 9 diferentes fontes de alimentos para os habitantes de países europeus. Para 22 desses países europeus também foram mensuradas as porcentagens de força de trabalho empregadas em diferentes 9 diferentes grupos de indústria.

Denotaremos o conjunto de variáveis de consumo de proteínas por X_i, em que $i = 1$ representa o consumo de proteínas provenientes de carne vermelha; $i = 2$ de carne branca; $i = 3$ de ovos; $i = 4$ de leite; $i = 5$ de peixe; $i = 6$ de cereais; $i = 7$ de carboidratos; $i = 8$ de grãos, nozes e óleo de linhaça; e $i = 9$ de frutas e verduras. E o conjunto de variáveis sobre a força de trabalho será Y_j sendo em $j = 1$ a porcentagem de força de trabalho empregada em agricultara ; $j = 2$ em mineração e exploração de pedreiras; $j = 3$ em fabricação; $j = 4$ em fornecimento de energia e água; $j = 5$ em construção; $j = 6$ em serviços; $j = 7$ em finanças; $j = 8$ em serviço social e pessoal; e $j = 9$ em transporte e comunicações.

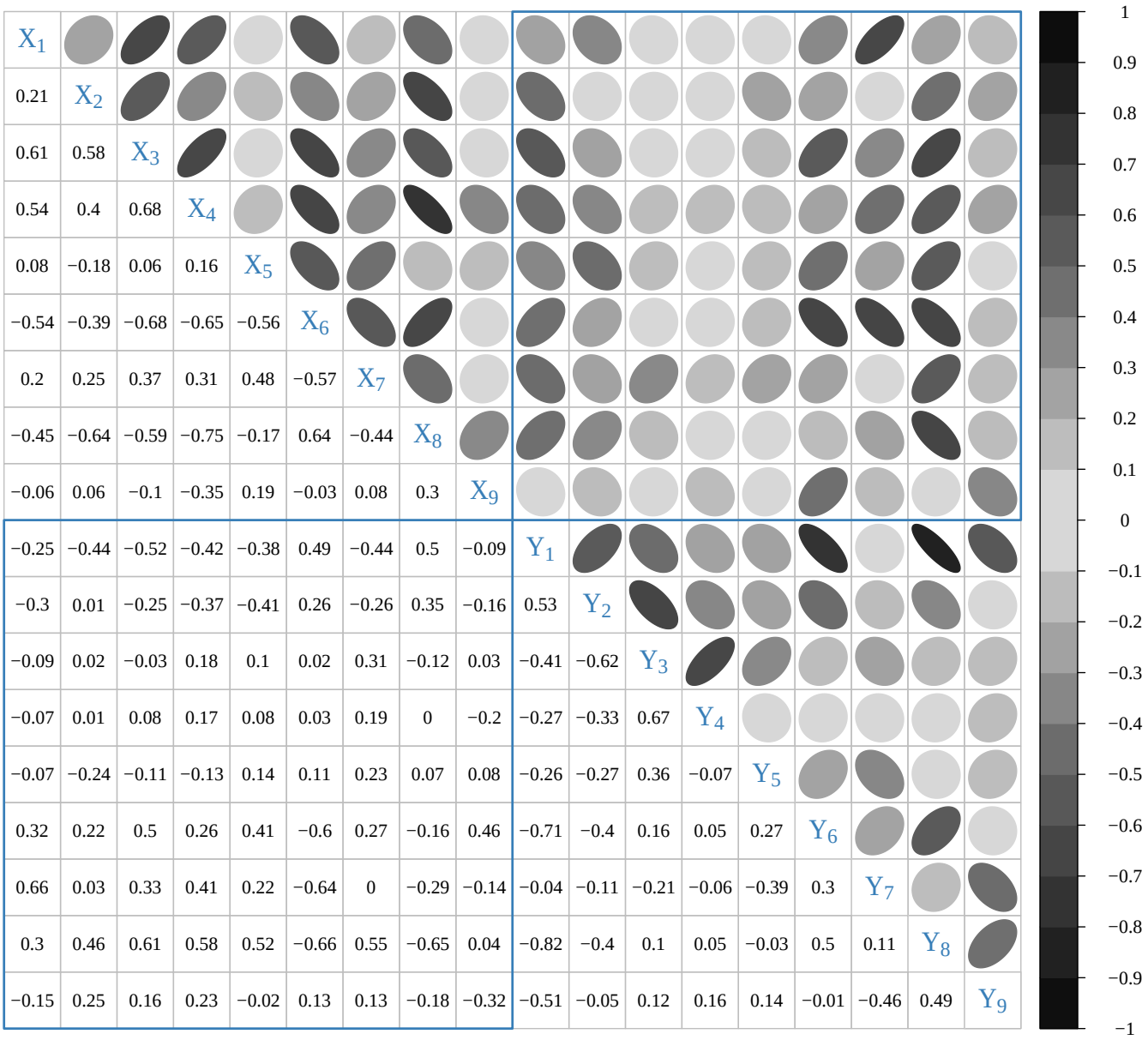


Figura 11: Matriz de correlações amostrais entre todas as 18 variáveis mensuradas para 22 países europeus.

As correlações dentre as variáveis X_i e Y_j, i,j $= 1,2,\ldots,9$ é apresentada na Figura 11. A diagonal secundária da matriz de correlações é destaca na figura para enfatizar que em análise de correlação canônica estamos resumindo todas essas correlações. Em geral, nota-se que há mais correlações fortes entre as variáveis de um mesmo grupo do que entre variáveis de grupos diferentes, embora que nesse segundo caso também se encontra correlações fortes, por exemplo X_1 (consumo de carne vermelha) e Y_7 (força empregada no setor de finanças), com correlação 0.663 e X_6 (consumo de cereais) e Y_8 (serviço social e pessoal), com -0.662.

Como são nove variáveis em cada um dos grupos, foram obtidas 9 correlações canônicas 0.977, 0.975, 0.931, 0.873, 0.716, 0.663, 0.541, 0.167, 0.027 que são respectivas aos nove pares de variáveis canônicas fornecidas pelo método. Note que são correlações bastante fortes o que indica que há correlação entre o grupo de variáveis X_i, sobre consumo de proteínas de diferentes fontes, e o grupo Y_j, sobre as porcentagens de esforço de trabalho empregadas em diferentes grupos de indústrias. O teste de Barlett aplicado as correlações obtidas resultou em uma estatística F de 129.175 que comparada com a distribuição F de Snedecor com 64 graus de liberdade, está bastante distante da região de alta probabilidade o que sugere fortemente a rejeição da hipótese nula (*p-valor* de 5.367×10^{-4}) de que todas as correlações não são diferentes de 0. Isso evidência que pelo menos uma das correlações canônicas é significante.

Uma das possibilidades interessantes da análise de correlação canônica é a interpretação das variáveis canônicas geradas. Para auxiliar nessa interpretação são apresentadas na Tabela 8 as correlações das variáveis canônicas com as variáveis originais. Como já enfatizado nas seções anteriores para uma correta e completa interpretação se faz necessário a avaliação de um pesquisador com conhecimentos sobre os dados. Aqui far-se-á apenas uma breve interpretação para o primeiro par de variáveis canônicas, U_1 e V_1. Note para U_1 que as correlações são todas positivas com exceção de X_6 (consumo de proteína de cereais) e X_8 (consumo de proteína de grão, nozes e óleo de linhaça), então pode-se interpretar essa variável como um contraste do consumo total de proteínas contra o consumo devido a proteínas provenientes de cereais, grãos, nozes e linhaça. A variável canônica V_1 é basicamente expressa pelo contraste de Y_1 (porcentagem empregada a agricultura) e Y_2 (porcentagem empregada com mineração) versus Y_6 (porcentagem empregada com serviços) e Y_8 (porcentagem empregada com serviço social e pessoal). Como U_1 e V_1 são positivamente correlacionadas pode-se dizer que os contraste definidos anteriormente estão relacionados.

Tabela 8: Correlações entre as variáveis canônicas e as variáveis originais

Variável	U_1	U_2	U_3	Variável	V_1	V_2	V_3
X_1	0.287	-0.612	0.226	Y_1	-0.589	-0.010	-0.135
X_2	0.170	0.024	-0.146	Y_2	-0.652	-0.093	-0.297
X_3	0.508	-0.306	0.032	Y_3	0.186	0.230	0.232
X_4	0.415	-0.436	0.405	Y_4	0.005	0.121	0.287
X_5	0.682	-0.200	0.204	Y_5	0.176	0.255	0.186
X_6	-0.648	0.655	-0.025	Y_6	0.632	-0.299	-0.394
X_7	0.555	-0.049	0.289	Y_7	0.104	-0.930	0.006
X_8	-0.452	0.212	-0.431	Y_8	0.736	-0.075	0.232
X_9	0.425	0.204	-0.735	Y_9	-0.012	0.363	0.412

Na Figura 12 são apresentados gráficos que auxiliam a avaliação e interpretação dos três primeiros e mais correlacionados pares de variáveis canônicas. No primeiro gráfico à esquerda exibi-se a dispersão dos valores das variáveis canônicas U_1 e V_1. Nesse gráfico são destacados os países Albânia, Iugoslávia, Hungria e Romênia com valores baixos de U_1 e V_1.

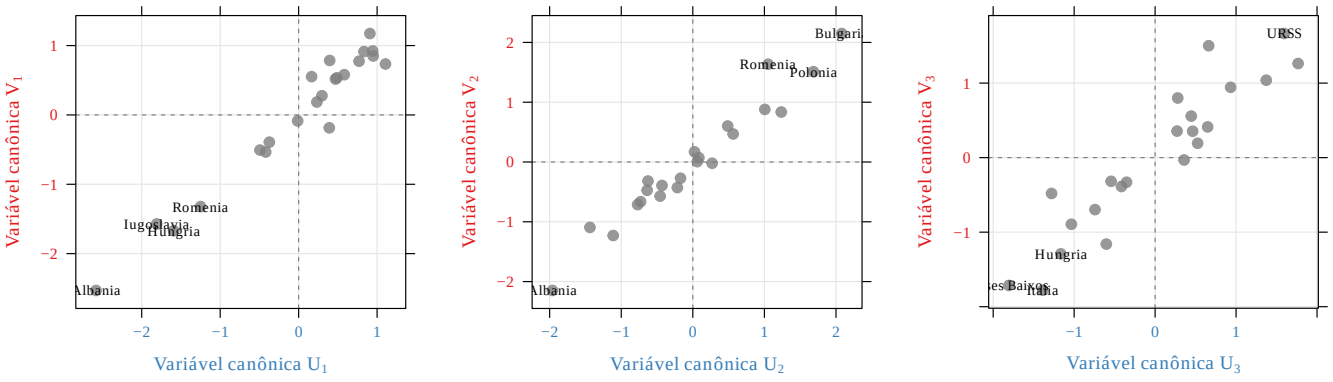


Figura 12: Gráficos de dispersão entre os três primeiros pares de variáveis canônicas.

8 Escalonamento Multidimensional

Essa seção destina-se à aplicação da análise de escalonamento multidimensional, descrita no capítulo 11 do livro-texto. Essa abordagem é ideal para casos em que se mensura distâncias ou medidas de dissimilaridade entre objetos. O objetivo da análise multidimensional é reproduzir dimensões que reflitam a matriz de medidas de dissimilaridade obtida.

A aplicação proposta nessa seção refere-se aos dados já apresentados anteriormente na Seção 7, onde foram apresentados os dados sobre empregos em países europeus. O estudo mensurou a porcentagem da força de trabalho de empregados em nove diferentes grupos de indústria (agricultura, floresta e pesca; mineração e exploração de pedreiras; fabricação; fornecimento de energia e água; construção; serviços; finanças; serviços sociais e pessoais; e transportes e comunicações). Como aqui só considera-se os dados sobre empregos, há dados sobre 30 países.

Vale ressaltar que em análises não didáticas têm-se apenas essa matriz de dados observados, é incomum e não recomendável trabalhar somente com as distâncias quando se tem os dados originais.

Na Figura 13 são apresentadas as distâncias calculadas entre cada par de países. As siglas ao lado do nome de cada país indicam o grupo político (UE, União Européia; AELC, área européia livre comércio, países do leste europeu e outros países). No gráfico é nítido a dissimilaridade da Albânia com relação a praticamente todos os países, razoavelmente similar somente com a Turquia. Outra característica marcante é a similaridade entre países de mesmo grupo político, sobretudo destaca-se as semelhanças entre países da União Européia (UE) e da área européia de livre comércio (AELC).

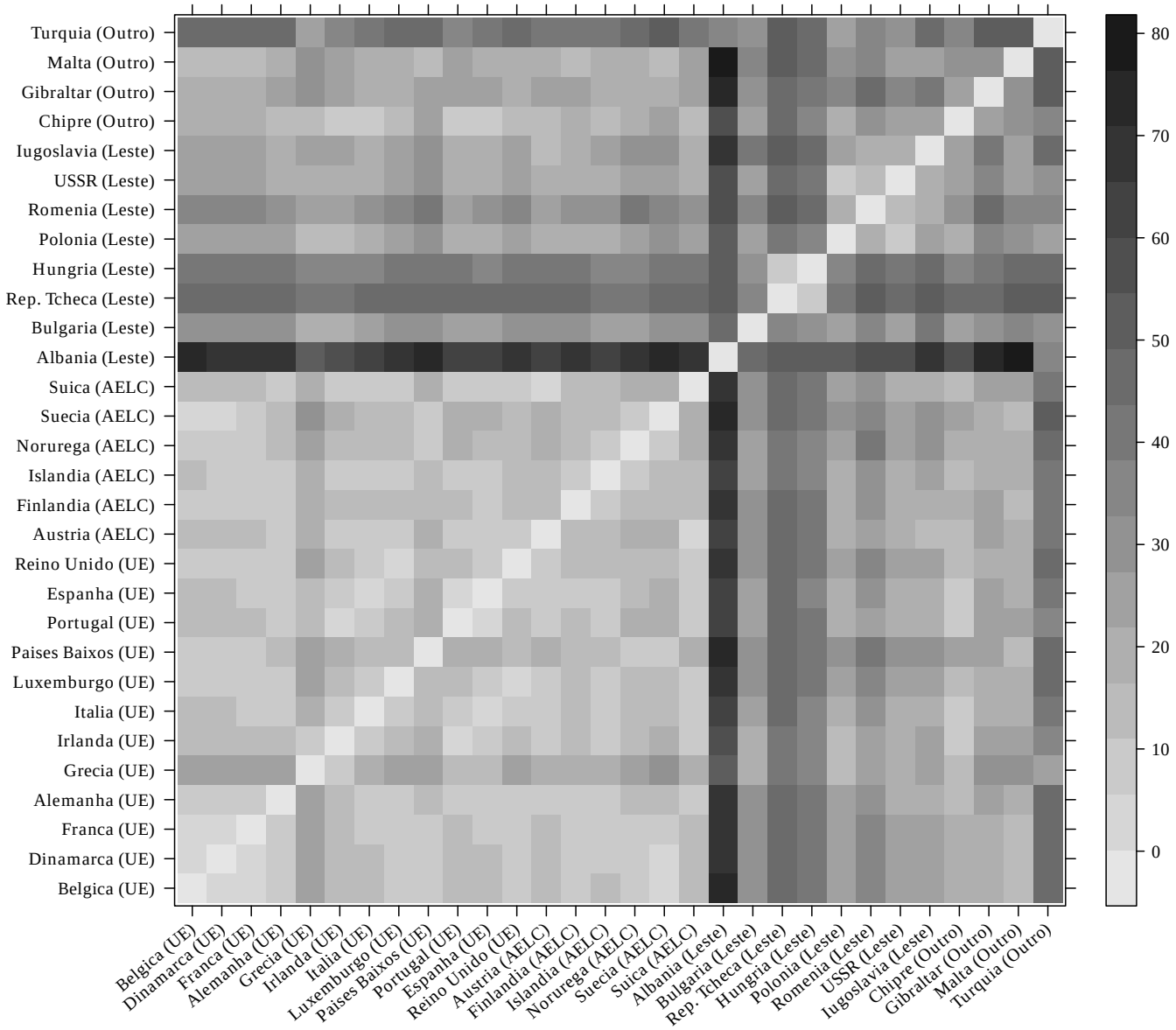


Figura 13: Matriz de distância entre os países europeus considerando as porcentagens de trabalho empregada em cada setor industrial.

A partir da matriz de distâncias procedeu-se com a análise de escalonamento multidimensional. Considerou-se apenas o escalonamento não-métrico de Kruskal. A definição do número de dimensões necessárias para que

se reflita adequadamente a matriz de distâncias foi realizada por meio da avaliação do STRESS considerando diferentes dimensões.

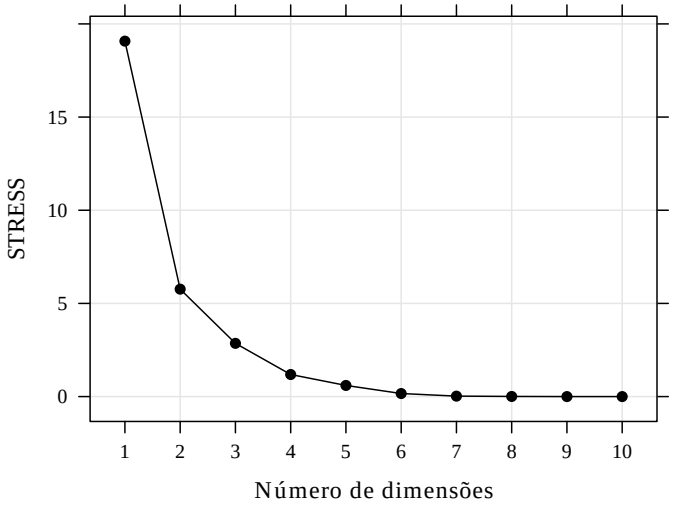


Figura 14: Valor de STRESS para diferentes números de dimensões no escalonamento multidimensional não-métrico de Kruskal.

Na Figura 14 são exibidos os valores de STRESS calculados para dimensões de 1 a 10. Note que claramente uma ou duas dimensões são insatisfatórias, pois os valores de STRESS são superiores a 5 e o decréscimo quando consideradas mais dimensões é expressivo. O valor de STRESS para 3 dimensões foi de 2.857 e para 4 dimensões resultou em 1.185 um decréscimo de 1.672. Para 5 dimensões o houve uma diminuição de apenas 0.587. Portanto foram escolhidas 4 dimensões para representar a matriz de distâncias apresentada na Figura 13.

A representação das observações nas dimensões obtidas é apresenta na Figura 15 (à esquerda). Esse gráfica apresenta as observações nas coordenadas de cada dimensões. Nota-se que algumas dimensões podem ser interpretadas, por exemplo a dimensão 1 D_1, separa bem os países do Leste europeu dos demais. Ao investigar a tabela de dados observa-se que países do leste investem mais força de trabalho em agricultura, florestal e pesca e em mineração e exploração de pedreira e tem porcentagens baixas para serviços.

O maior valor obtido nessa dimensão ocorreu para Albânia com 59.7 u.m.. A segunda dimensão tem dois países destacados na extremidade inferior, esses são Hungria e Rep. Tcheca e da mesma forma poderíamos elencar suas características para buscar interpretação dessa dimensão.

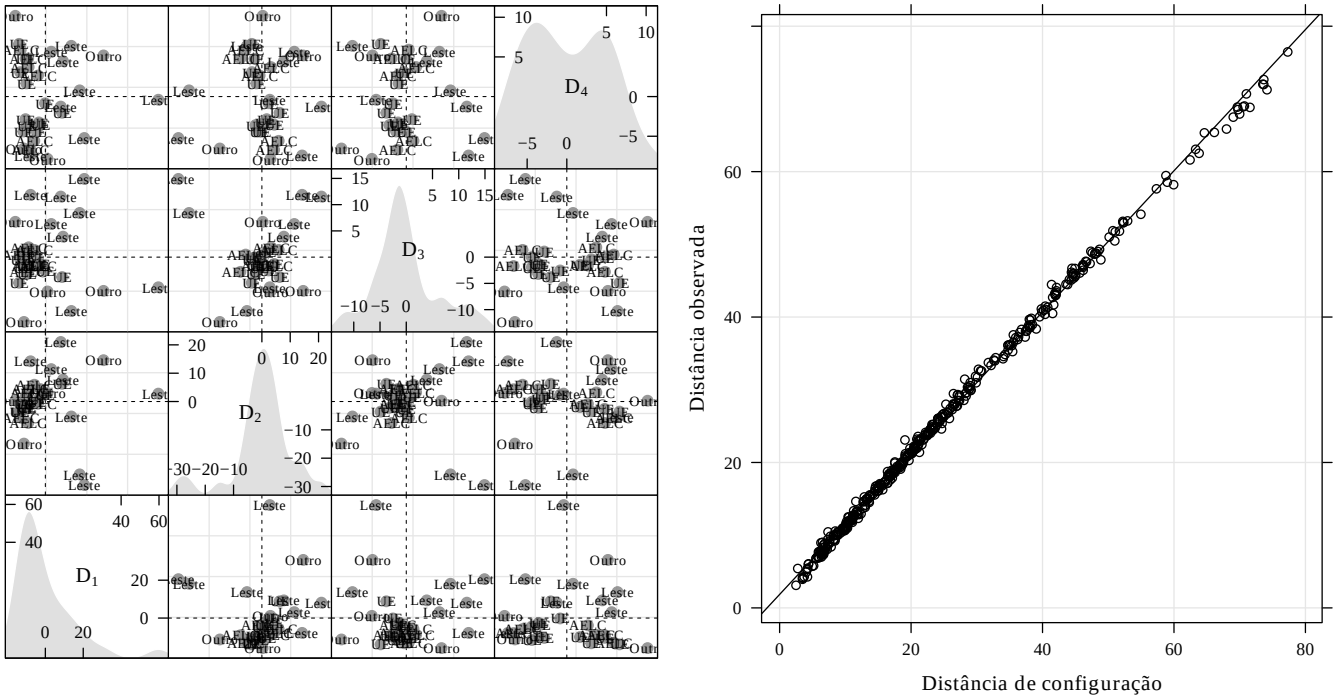


Figura 15: Representação dos países europeus nas 4 dimensões obtidas pela análise multidimensional não métrica de Krukal (esquerda) e distâncias observadas e distâncias obtidas das dimensões conforme configuração da análise.

A direita da Figura 15 têm-se uma avaliação da qualidade do modelo 4-dimensional não-métrico. São apresentadas as distâncias originais e as distâncias calculadas das quatro dimensões obtidas. O esperado é que as dimensões recuperem as distâncias que foram utilizadas na sua definição e como pode ser observado essa representação das distâncias pelas 4 dimensões se mostra bastante satisfatória.

Referências

HOLLAND, S. M. **Principal Components Analysis (PCA)**. University of Georgia, 2008.
JOHN FOX, S. W. **"Multivariate linear models in R." An R companion to applied regression**. 2nd. ed.
JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 6th. ed. Prentice hall Upper Saddle River, NJ, 2007.
MANLY, B. F. J. **Métodos Estatísticos Multivariados: uma introdução**. 3rd. ed.
MURTAGH, F.; LEGENDRE, P. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? **Journal of Classification**, v. 31, n. 3, p. 274–295, 2014.
PET ESTATÍSTICA UFPR. **labestData: Biblioteca de Dados para Aprendizado de Estatística**.
R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2016.