Universidade Estadual Do Oeste do Paraná
PGEAGRI
Programa de Pós-Graduação em Engenharia Agrícola

ANÁLISE MULTIVARIADA
Prof.: Dra. Luciana Pagliosa

Resolução da 1ª Parte da Prova de Análise Multivariada 2018

Vanessa Mendes Pientosa
Willyan Goergen de Souza

Cascavel - PR
Novembro de 2018

Questão 1: Descreva as diferenças conceituais entre as análises fatorial, de agrupamento e de componentes principais.

Resposta:

- **Análises Fatorial:** A ideia de analise fatorial foi introduzida por Charles Edward Spearman (1863 -1945), como sendo medidas obtidas em teste de habilidades mental (medidas de habilidade matemática, vocabulário, habilidade verbal, raciocínio lógico, habilidades artísticas)podem todas serem explicadas por um fator determinado de "inteligência geral", uma idéia essa interessante porém errônea.

Os objetivos da análise fatorial são: redução dos dados originais; caracterizar os elementos amostrais (com um índice geral ou um diagnostico único), considerando o conjunto das *P* variáveis pesquisadas e evidenciar as relações entre as variáveis, formando grupos (fatores).

Ou seja, é um conjunto de técnicas estatísticas que nos permite representar um numero de variáveis iniciais a partir de um menor número de variáveis hipotéticas, a partir da estrutura de dependência das variáveis iniciais.

- Análise de agrupamento ou clustering: Tem como proposito separar objetos em grupos se baseando nas características que estes objetos possuem, ou seja, a ideia básica é colocar em um mesmo grupo objetos que sejam similares de acordo com algum critério pré-determinado. O critério se baseia em uma função de dissimilaridade, função esta que recebe dois objetos e retoma a distância entre eles, isso quer dizer que os elementos de um determinado conjunto devem ser mutualmente similares e de preferência muito diferentes dos elementos de outros conjuntos.

Estas analise é uma ferramenta muito útil para a analise de dados em muitas situações diferentes, como por exemplo, usada para reduzir a dimensão de um conjunto de dados, reduzindo uma ampla gama de objetos a informação do centro do seu conjunto. Também pode servir para extrair características escondidas dos dados e desenvolver as hipóteses a respeito da sua natureza (LINDEN, 2009).

- Análise de componentes principais: A Análise de Componentes Principais (ACP) envolve um procedimento matemático que transforma um número de variáveis possivelmente correlacionadas em um número menor de variáveis não correlacionadas denominadas componentes principais. O primeiro componente principal contém a maior variabilidade possível da variabilidade total dos dados, e cada componente principal sucessiva contém parte da variabilidade restante, tão grande quanto possível. Para tal, efetua-se a decomposição da matriz de covariância dos dados, usualmente normalizados, em seus autovalores e autovetores correspondentes. A ACP é utilizada principalmente como um instrumento de análise exploratória de dados e para construir modelos preditivos.

<u>Questão 2</u>: Os dados no quadro abaixo representam uma pesquisa sobre o consumo diário de proteína, por pessoa (em gramas), de origem animal e vegetal, em 25 países da Europa (FONTE: MANLY, F.J. Bryan (1986). Nessa pesquisa foi observado o consumo diário de proteína proveniente de: Carne de gado (x1), carne de frango (X2), ovos (X3), leite (X4), peixe (X5), cereais (X6), alimentos enriquecidos (X7), amêndoas e sementes oleaginosas (X8) e frutas e vegetais (X9).

Quadro 1: Consumo diário de proteína, por pessoa (em gramas), de origem animal e vegetal, em 25 países da Europa, segundo a fonte de consumo (FONTE: Adaptado de MANLY, F. J. Bryan, 1986).

País	X1	X2	X3	X4	X5	X6	X7	X8	X9
ALBÂNIA	10,1	1,4	0,5	8	0,2	42,3	0,6	5,5	1,7
ÁUSTRIA	8,9	14	4,3	19,9	2,1	28	3,6	1,3	4,3
BÉLGICA	13,5	9,3	4,1	17,5	4,5	26,6	5,7	2,1	4
BULGÁRIA	7,8	6	1,6	8,3	1,2	56,7	1,1	3,7	4,2
TCHECOSLOVAQUIA	9,7	11,4	2,8	12,5	2	34,3	5	1,1	4
DINAMARCA	10,6	10,8	3,7	25	9,9	21,9	4,8	0,7	2,4
EX-A, ORIENTAL	8,4	11,6	3,7	11,1	5,4	24,6	6,5	0,8	3,6
FINLÂNDIA	9,5	4,9	2,7	33,7	5,8	26,3	5	1	1,4
FRANÇA	18	9,9	3,3	19,5	5,7	28,1	4,8	2,4	6,5
GRÉCIA	10,2	3,8	2,8	17,6	5,9	41,7	2,2	7,8	6,5
HUNGRIA	5,3	12,4	2,9	9,7	0,3	40,1	4	5,4	4,2
IRLANDA	13,9	10	4,7	25,8	2,2	24,8	6,2	1,6	2,9
ITÁLIA	9	5,1	2,9	13,7	3,4	36,8	2,1	4,3	6,7
HOLANDA	9,5	13,6	3,6	23,4	2,5	22,4	4,2	1,8	3,7
NORUEGA	9,4	4,7	2,7	23,3	9,7	23	4,6	1,6	2,7
POLÔNIA	6,9	10,2	2,7	19,3	3	36,1	5,9	2	6,6
PORTUGAL	6,2	3,7	1,1	4,9	14,2	27	5,9	4,7	7,9
ROMÊNIA	6,2	6,3	1,5	11,1	1	49,6	3,1	5,3	2,8
ESPANHA	7,1	3,4	3,1	8,6	7,8	29,2	5,7	5,9	7,2
SUÉCIA	9,9	7,8	3,5	24,7	7,5	19,5	3,7	1,4	2
SUIÇA	13	10,1	3,1	23,8	2,3	25,6	2,8	2,4	4,9
INGLATERRA	17,4	5,7	4,7	20,6	4,3	24,3	4,7	3,4	3,3
EX-URSS	9,3	4,6	2,1	16,6	3	43,6	6,4	3,4	2,9
EX-A, OCIDENTAL	11,4	12,5	4,1	18,8	3,4	18,6	5,2	1,5	3,8
IUGOSLÁVIA	4,4	5	1,2	9,5	0,6	55,9	3	5,7	3,2

<u>RESPOSTA</u>:

Realizou-se primeiro uma análise descritiva dos dados, para que se tenha um prévio conhecimento sobre o comportamento dos valores de consumo de proteínas por pessoa em 25 países da Europa.

Tabela 1: Estatística descritiva dos valores de consumo diário (em grama) de proteína por pessoa

	X1	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**
Min.	4.40	1.40	0.50	4.90	0.20	18.60	0.60	0.70	1.40
1o Qu.	7.80	4.90	2.70	11.10	2.10	24.60	3.10	1.50	2.90
Mediana	9.50	7.80	2.90	17.60	3.40	28.00	4.70	2.40	3.80
Média	9.82	7.93	2.94	17.08	4.32	32.28	4.27	3.07	4.14
3o Qu.	10.60	10.80	3.70	23.30	5.80	40.10	5.70	4.70	4.90
Max.	18.00	14.00	4.70	33.70	14.20	56.70	6.50	7.80	7.90

A analise descritiva que se encontra na Tabela 1, mostra os valores referente as médias de consumo diário em gramas de proteína por pessoa. Observou-se que o maior valor se deu para a variável X6 (cereais), com 32.28 g/pessoa por dia e o menor se deu para a variável X3 (ovos), com 2.93 g/pessoa por dia.

A variável com o menor valor de Mínimo foi X5 (peixe) com 0.20 g/pessoa e o maior valor de mínimo foi X6 (cereais) com 18,60 g/pessoa. Já para os valores de Máximo, a variável com maior valor foi novamente X6 (56,70 g/pessoa) e com o menor valor X3 (ovos, 4,70 g/pessoa).

A distribuição dos dados em relação à média e mediana, ou seja, os quartis, ficam mais evidentes numa análise gráfica. Para isso, a Figura 1 mostra os gráficos *boxplot* das 9 variáveis.

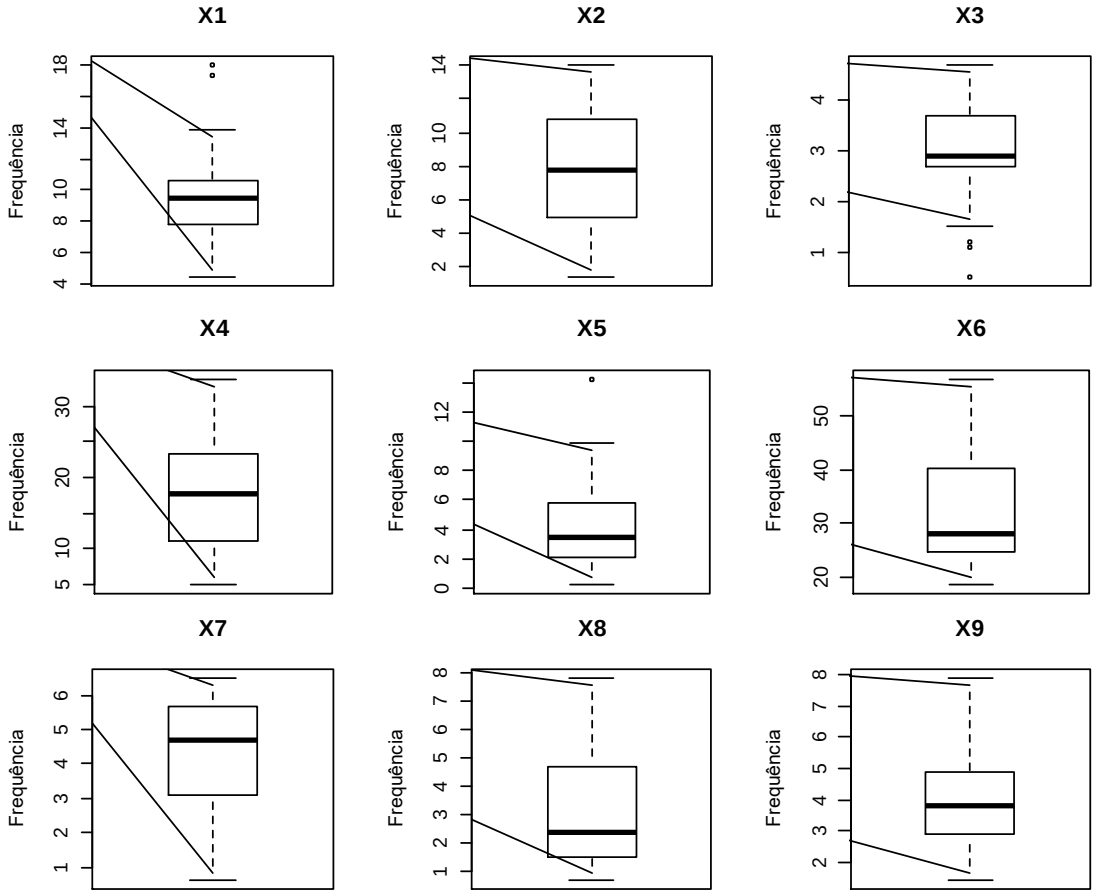

Figura 1: Gráficos boxplot das 9 variáveis

Realizando a análise dos gráficos da Figura 1, nota-se inicialmente a presença de possíveis pontos discrepantes nas variáveis X1, X3 e X5. Nota-se também distribuições mais assimétricas dos quartis nas variáveis X3, X6 e X8, enquanto que nas variáveis X2 e X9 os quartis estão bem distribuídos.

Alternativa A) Faça a análise desses dados pela técnica de componentes principais. Interprete os dados obtidos.

Estabelece-se inicialmente a matriz de Variância e Covariância entre as variáveis a serem analisadas no *software* R, conforme Tabela 2 a seguir.

Tabela 2: Matriz de Variâncias e Covariâncias.

	X1	X2	X3	X4	X5	X6	X7	X8	X9
X1	11.176	1.895	2.190	11.923	0.612	-18.199	0.748	-2.320	-0.451
X2	1.895	13.345	2.557	7.649	-3.038	-16.392	1.837	-4.500	-0.330
X3	2.190	2.557	1.249	4.662	0.255	-8.680	0.827	-1.242	-0.092
X4	11.923	7.649	4.662	51.135	3.204	-46.308	2.651	-8.854	-5.143
X5	0.612	-3.038	0.255	3.204	11.784	-19.748	2.287	-0.900	1.736
X6	-18.199	-16.392	-8.680	-46.308	-19.748	119.922	-9.474	14.138	0.880
X7	0.748	1.837	0.827	2.651	2.287	-9.474	2.664	-1.530	0.260
X8	-2.320	-4.500	-1.242	-8.854	-0.900	14.138	-1.530	3.943	1.343
X9	-0.451	-0.330	-0.092	-5.143	1.736	0.880	0.260	1.343	3.254

A obtenção dos componentes principais ocorre por meio da matriz de variância e covariância (S) ou da matriz de correlação linear (R), cuja equivale à matriz de covariância padronizada. Esses componentes principais são influenciados pelas variáveis de maior variância, sendo de pouca utilidade nos casos em que existe uma discrepância muito grande entre essas variáveis.

Apesar das variáveis originais estarem todas na mesma unidade (g/pessoa), nota-se que entre elas existe uma grande variação entre os maiores valores e menores valores, isto é, amplitude dos dados elevada (menor valor geral 0,2 e maior 56,7). Sendo assim optou-se por usar a matriz R (Tabela 3), que se faz mais confiável para a respectiva análise:

Tabela 3: Matriz de Correlação de Pearson

	X1	X2	X3	X4	X5	X6	X7	X8	X9
X1	**1**	Fraca	Moderada	Moderada	Fraca	Moderada	Fraca	Moderada	Fraca
X2	0.155	**1**	Forte	Moderada	Fraca	Moderada	Moderada	Forte	Moderada
X3	0.586	0.626	**1**	Moderada	Fraca	Forte	Moderada	Moderada	Fraca
X4	0.499	0.293	0.583	**1**	Fraca	Moderada	Fraca	Forte	Moderada
X5	0.053	-0.242	0.066	0.131	**1**	Moderada	Moderada	Fraca	Fraca
X6	-0.497	-0.410	-0.709	-0.591	-0.525	**1**	Moderada	Forte	Fraca
X7	0.137	0.308	0.453	0.227	0.408	-0.530	**1**	Moderada	Fraca
X8	-0.349	-0.620	-0.560	-0.624	-0.132	0.650	-0.472	**1**	Moderada
X9	-0.075	-0.050	-0.046	-0.399	0.280	0.045	0.088	0.375	**1**

Classificações da intensidade da correlação segundo Callegari-Jacques (2003).
Em vermelho = correlações negativas

Segundo Callegari-Jacques (2003), para avaliar de maneira qualitativa o coeficiente de correlação de Pearson, podemos adotar o seguinte critério:

• se 0,00 <| r |< 0,30, existe fraca correlação linear;

• se 0,30 ≤| r |< 0,60, existe moderada correlação linear;
• se 0,60 ≤| r |< 0,90, existe forte correlação linear;
• se 0,90 ≤| r |< 1,00, existe correlação linear muito forte.

Na matriz de correlação, o maior valor obtido foi entre o consumo diário de proteína derivado de cereais com sementes oleaginosas com valor de (0.650) e a menor correlação positiva foi entre cereais e frutas e vegetais (0.044). Os valores positivos sugerem a existência de um relacionamento linear positivo (crescimento) entre as variáveis, ou seja, o consumo entre essas proteínas é diretamente proporcional.

Ainda na matriz de correlação, a maior correlação negativa foi entre cereais e ovos (-0.709) e a menor correlação negativa foi obtida entre ovos e frutas e vegetais (-0.045). Os valores negativos podem indicar a existência de uma relação linear negativa (decrescimento), ou seja, o consumo entre essas proteínas é inversamente proporcional.

Com o auxilio do R, aplicou-se o Teste de Esfericidade de Bartlett, verificando que o mesmo foi não foi significativo a 5% de significância, com p-valor calculado 3.628e-10<0,05, demostrando assim, que a análise dos componentes principais e a análise fatorial são adequados.

Para o teste KMO (Coeficiente Kaiser-Mayer-Olkin), o valor obtido foi de 0.64, que é considerável aceitável e indicando que a análise é apropriada. Para este teste, quando KMO for mais próximo de 1, se assegura a confiabilidade dos dados que serão apresentados pelos métodos de componentes principais e fatorial.

Os Autovalores (**λ**) e Autovetores (**e**) da matriz R se encontram a seguir:

$$\boldsymbol{\lambda} = \begin{vmatrix} 4.000 & 1.653 & 1.118 & 0.948 & 0.467 & 0.326 & 0.274 & 0.114 & 0.101 \end{vmatrix}$$

$$\mathbf{e} = \begin{array}{|ccccccccc|} \multicolumn{9}{c}{\textbf{1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9}} \\ -0.303 & 0.059 & 0.279 & 0.660 & -0.347 & -0.416 & 0.172 & -0.015 & -0.262 \\ -0.310 & 0.236 & -0.628 & -0.051 & 0.290 & -0.120 & -0.022 & -0.004 & -0.594 \\ -0.429 & 0.036 & -0.191 & 0.299 & -0.068 & 0.335 & -0.465 & -0.494 & 0.334 \\ -0.380 & 0.181 & 0.373 & 0.009 & 0.233 & 0.644 & 0.423 & 0.104 & -0.158 \\ -0.132 & -0.651 & 0.322 & -0.192 & 0.282 & -0.130 & -0.098 & -0.441 & -0.343 \\ 0.437 & 0.234 & -0.101 & -0.006 & -0.245 & 0.147 & 0.392 & -0.688 & -0.188 \\ -0.297 & -0.353 & -0.227 & -0.350 & -0.728 & 0.187 & 0.133 & 0.134 & -0.113 \\ 0.419 & -0.145 & 0.035 & 0.336 & -0.125 & 0.445 & -0.443 & 0.227 & -0.478 \\ 0.108 & -0.532 & -0.431 & 0.448 & 0.228 & 0.121 & 0.445 & 0.085 & 0.219 \end{array}$$

Os componentes principais (CPs) são gerados pela combinação linear de cada autovetor com os valores observados de cada variável, o que resulta nos respectivos coeficientes de cada CP. Porém, nesse trabalho, os mesmos foram gerados através do *software* R, e o resultado mostrado na Tabela 4 a seguir (usando a matriz R como explicitado anteriormente):

Tabela 4: Coeficientes dos CPs

	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8	CP9
X1	-0.303	-0.059	-0.279	-0.660	0.347	-0.416	0.172	-0.015	0.262
X2	-0.310	-0.236	0.628	0.051	-0.290	-0.120	-0.022	-0.004	0.594
X3	-0.429	-0.036	0.191	-0.299	0.068	0.335	-0.465	-0.494	-0.334
X4	-0.380	-0.181	-0.373	-0.009	-0.233	0.644	0.423	0.104	0.158
X5	-0.132	0.651	-0.322	0.192	-0.282	-0.130	-0.098	-0.441	0.343
X6	0.437	-0.234	0.101	0.006	0.245	0.147	0.392	-0.688	0.188
X7	-0.297	0.353	0.227	0.350	0.728	0.187	0.133	0.134	0.113
X8	0.419	0.144	-0.035	-0.336	0.125	0.445	-0.443	0.227	0.478
X9	0.108	0.532	0.431	-0.448	-0.228	0.121	0.445	0.085	-0.219

Tendo-se definidos os CPs, procede-se então a tomada de decisão de quantos CPs serão usados na análise. Para isso, a quantidade é determinada em função da representatividade das variâncias originais.

Uma análise possível é o gráfico de Scree Plot (Gráfico de cotovelo, Figura 2), que é uma forma de determinação do número mínimos de CPs consideradas na análise. No entanto, essa verificação é muito subjetiva, uma vez que sua avaliação é visual, onde procura-se no comportamento da curva decrescente sua mudança a intensidade, ou seja, a partir de qual CP o decaimento é visualmente irrelevante.

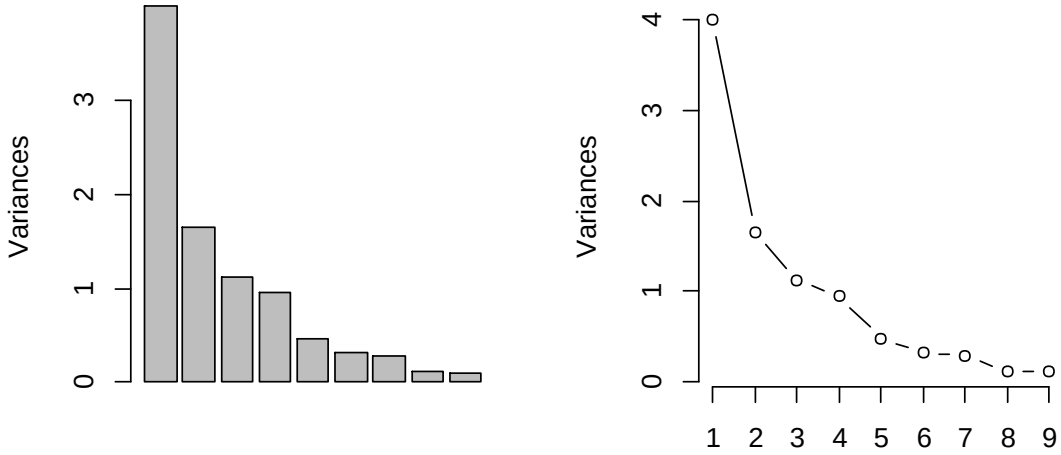

Figura 2: Gráfico Scree Plot

Analisando a Figura 1, o gráfico apresenta uma forma de cotovelo, aparentemente, mais nítida no CP3. Mas para se ter uma definição mais confiável procede-se uma verificação numérica das variâncias acumuladas representadas em cada CP (Tabela 5):

Tabela 5: Estatística de variância dos componentes principais.

	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8	CP9
Variância	4.000	1.653	1.118	0.948	0.467	0.326	0.274	0.114	0.101
Desvio Padrão	2.000	1.286	1.057	0.974	0.683	0.571	0.523	0.337	0.318
Proporção de Variância	0.444	0.184	0.124	0.105	0.052	0.036	0.030	0.013	0.011
Proporção Acumulada	0.444	0.628	0.752	**0.858**	0.909	0.946	0.976	0.989	1.000

Segundo a Regra de Kaiser, os componentes principais retidos devem ter autovalores (variâncias) maiores que 1 (Kaiser, 1960; Silva et al., 2015). Assim, seriam escolhidos o CP1, CP2 e CP3. Porém, a proporção acumulada desses 3 componentes (porcentagem de explicação da variância das variáveis iniciais) seria de 75,22%.

Já para Johnson e Wichern (1999), deve-se considerar os primeiros CP's que expliquem de 80 a 90% da variabilidade total das variáveis. Para atender essa recomendação, então, adotaram-se os 4 primeiros CP's, que juntos representam 85,76% do total.

Na Tabela 6, apresentam-se os coeficientes dos CPs adotados para análise:

Tabela 6: Coeficientes dos CPs adotados:

	CP1	CP2	CP3	CP4
X1	-0.3028	-0.0592	-0.2792	**-0.6597**
X2	-0.3103	-0.2357	**0.6281**	0.0509
X3	**-0.4288**	-0.0360	0.1906	-0.2990
X4	-0.3798	-0.1808	**-0.3730**	-0.0092
X5	-0.1317	**0.6509**	-0.3221	0.1921
X6	**0.4369**	-0.2342	0.1007	0.0060
X7	-0.2972	**0.3529**	0.2274	**0.3501**
X8	**0.4192**	0.1445	-0.0351	-0.3359
X9	0.1075	**0.5319**	**0.4305**	**-0.4477**

Verifica-se assim, com base na Tabela 6, que as variáveis que mais influenciaram o CP1 foram ovos (X3), cereais (X6) e amêndoas e sementes oleaginosas (X8), sendo diretamente proporcional para X6 e X8 e inversamente proporcional para X3.

Já o CP2 foi mais influenciado pelas variáveis peixe (X5), alimentos enriquecidos e frutas e vegetais (X9), sendo todos eles diretamente proporcional.

As variáveis que tiveram mais peso no CP3 foram carne de frango (X2), leite (X4), e frutas e vegetais (X9). Sendo que X2 e X9 foram diretamente proporcionais, enquanto que X4 foi inversamente proporcional.

Por fim, para CP4, as variáveis que mais influenciaram foram carne de gado (X1), alimentos enriquecidos (X7) e frutos e vegetais (X9), sendo que X1 e X7 foram diretamente proporcionais, enquanto X9 foi inversamente proporcional.

Os componentes principais estão representados nas quatro equações apresentadas na sequência. Os valores das equações são referentes aos dados padronizados, uma vez que assim todas as variáveis apresentam o mesmo intervalo de amplitude dos dados.

Equações dos CPs:

$$Y_1 = -0.3028X_1 - 0.3103X_2 - 0.4288X_3 - 0.3798X_4 - 0.1317X_5 + 0.4369X_6 - 0.2972X_7 + 0.4192X_8 + 0.1075X_9$$

$$Y_2 = -0.0592X_1 - 0.2357X_2 - 0.0360X_3 - 0.1808X_4 + 0.6509X_5 - 0.2342X_6 + 0.3529X_7 + 0.1445X_8 + 0.5319X_9$$

$$Y_3 = -0.2792X_1 + 0.6281X_2 + 0.1906X_3 - 0.3730X_4 - 0.3221X_5 + 0.1007X_6 + 0.2274X_7 - 0.0351X_8 + 0.4305X_9$$

$$Y_4 = -0.6597X_1 + 0.0509X_2 - 0.2990X_3 - 0.0092X_4 + 0.1921X_5 + 0.0060X_6 + 0.3501X_7 - 0.3359X_8 - 0.4477X_9$$

Valores positivos para Y1, indicam que o país em questão terá maior destaque para X6, X8. Já os valores negativos indicam um maior consumo para as variáveis X1, X2, X3 e X7. Já para Y2, destaque para as variáveis de maior destaque sendo elas X5, X7 e X9, e valores negativos nas variáveis X2 e X6. Enquanto para Y3, recebe maior destaque as variáveis X2 e X9. Já os valores negativos que terão maior consumo serão X1 e X4. Finalmente para Y4, os valores positivos que terão maior destaque serão X7, enquanto os menores valores apresentam nas variáveis X1 e X9.

O gráfico Biplot, Figuras 3 e 4, pode ser utilizado para visualizar a associação das variáveis com os componentes. Nele, podemos observar que as variáveis X3 e X6 estão mais associadas ao componente principal 1. Já as variáveis X5 e X7 estão mais associadas ao 7 componente principal 2. As variáveis X2, X4 e X8 estão mais associadas ao componente principal 3 e as variáveis X1 e X9 estão mais associadas ao componente 4.

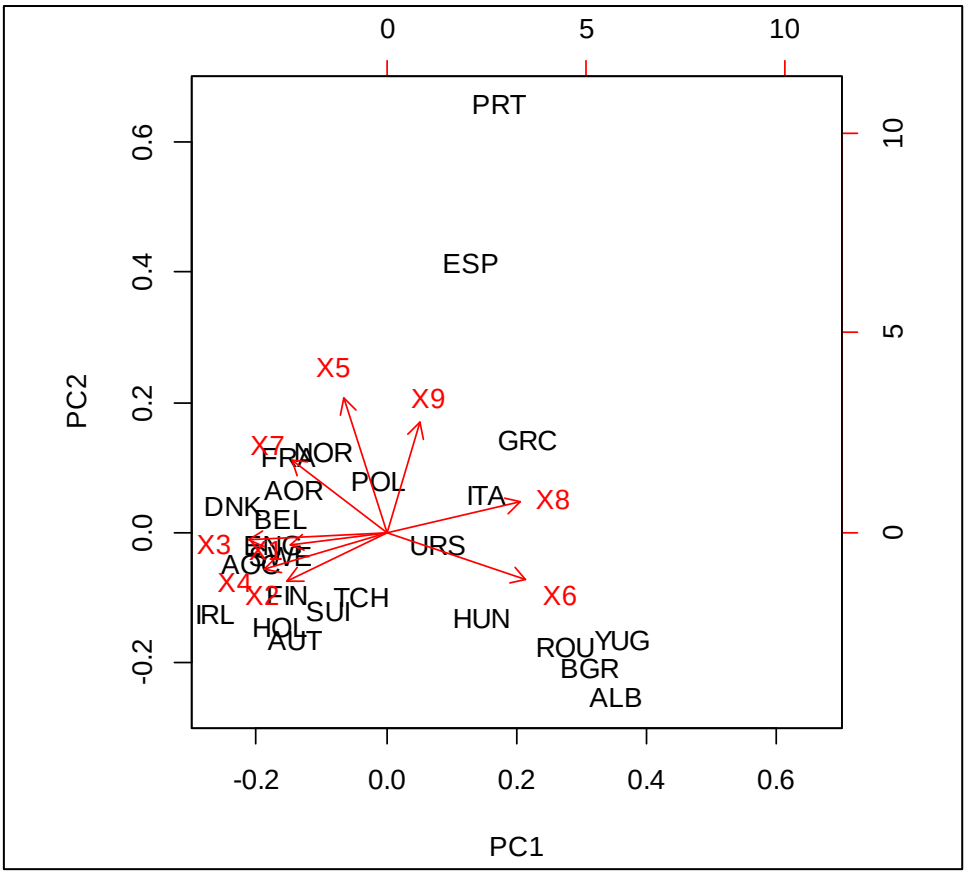

Figura 3: Biplot CP1 e CP2

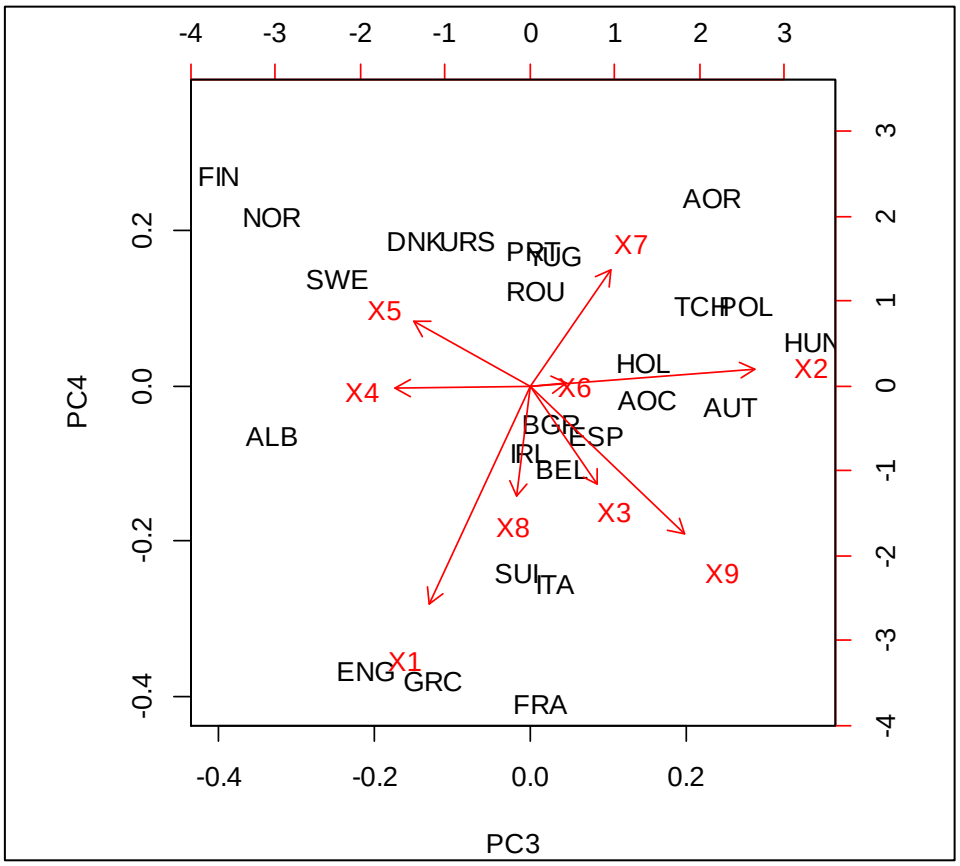

Figura 4: Biplot CP3 e CP4

Com base a combinação linear dos coeficientes de cada CP com os dados originais, calcularam-se os valores dos componentes principais para cada país, conforme mostrado na Tabela 7, sendo esses valores chamados de Scores. Dentro de cada semelhança no valor do Score dentro de cada componente principal, podemos agrupar os países que possuem similaridade no consumo de proteína.

Tabela 7: Scores dos 4 componentes principais.

	Scores			
	CP1	**CP2**	**CP3**	**CP4**
ALB	14.018	-10.259	0.043	-8.919
AUT	-4.549	-9.025	4.474	-7.156
BEL	-4.735	-5.161	0.501	-9.294
BGR	18.226	-12.776	6.053	-7.559
TCH	1.703	-8.302	5.955	-6.589
DNK	-10.255	-3.439	-3.678	-5.374
AOR	-3.120	-3.294	5.244	-4.577
FIN	-8.546	-7.638	-9.144	-5.074
FRA	-5.541	-4.416	-0.651	-13.307
GRC	8.602	-5.348	-1.167	-10.909
HUN	8.625	-9.865	9.714	-5.815
IRL	-9.443	-8.428	-1.935	-9.395
ITA	6.773	-5.796	1.956	-9.504
HOL	-8.168	-8.037	1.784	-7.041
NOR	-5.948	-1.756	-6.503	-5.117
POL	1.427	-7.020	4.579	-5.784
PRT	5.633	7.719	1.699	-4.434
ROU	14.447	-11.530	4.770	-5.776
ESP	5.482	2.044	2.155	-7.051
SWE	-9.068	-4.129	-5.208	-5.919
SUI	-5.859	-8.122	-1.072	-10.630
ENG	-6.445	-5.254	-4.632	-12.784
URS	7.038	-8.678	0.511	-6.041
AOC	-9.060	-5.238	1.982	-7.901
YUG	19.177	-12.317	5.894	-4.938

Na análise dos scores do componente 1 na Tabela 7 pode-se verificar que os maiores consumidores de proteína diária foram a Iugoslávia (19,17) e a Bulgária (18,22). Estes valores positivos indicam que nestes países se destacam o consumo das proteínas referentes as variáveis X6 e X8. A Dinamarca possui o menor valor negativo (-10,25), o que indica um dos maiores consumos para as variáveis X1, X2, X3 e X7.

No componente 2, o maior consumidor foi Potugual (7,71), ou seja, este pais se destaca no consumo de proteína das variáveis X5, X7 e X9. O país que Bulgaria teve o menor valor negativo (-12,77), se destacando com pouco consumos de X5, X7 e X9 e grande consumo de X6.

Para o componente 3, o maior consumidor de proteína diária foi a Hungria (9,71), ou seja, este pais se destaca em relação as variáveis X4 e X8. Os menores valores negativos foram a Finlandia (-9,14), logo estes país se destacam com valores maiores de consumo de proteína referente as variáveis X2 e X9.

Já para o componente 4, o maior valor positivo foi a Inglaterra (12,78), ou seja, este pais teve um bom consumo de proteína para a variável X7 e baixos consumos para as variáveis X1 e X9. O menor valor negativo foi a Grécia (-10,90), logo este pais se destaca com valor maior de consumo para as variáveis X1 e X9.

Alternativa B) Faça a análise dos dados pela técnica de análise fatorial. Interprete os resultados.

A análise fatorial é uma técnica que auxilia na identificação de novas variáveis hipotéticas, com menor número que as presentes no conjunto inicial, acarretando numa redução dos dados originais, mas que não traz perda significativa nas informações existentes no conjunto.

Para isso, é aplicado um método de estimação dessas variáveis. O aplicado aqui é o de Componentes Principais, cuja análise já foi previamente realizada na **Alternativa A**. Sendo assim, a Matriz de correlação (Tabela 3), seus autovalores e autovetores, e as proporções acumuladas de variância dos CPs criados (Tabela 5), necessárias para essa análise fatorial foram reaproveitadas da Alternativa anterior, evitando dados repetitivos.

Para o cálculo da proporção de variância acumulada, considera-se que os autovalores representam à variabilidade de cada componente e os autovetores compõem a base para se obter as cargas fatoriais. Em outras palavras, as p-variáveis originais geram através de suas combinações lineares, p-cargas fatoriais, que tem como principal característica, além da ortogonalidade, obter as cargas fatoriais em ordem decrescente de máxima variância, ou seja, a primeira componente principal detém mais informação estatística que a segunda componente principal, que por sua vez tem mais informação estatística que a terceira componente principal e assim sucessivamente, fazendo a redução da dimensão original das variáveis facilitando a interpretação das análises para o conjunto de dados.

Os valores de cargas fatoriais estão expressos na Tabela 8. As cargas fatoriais são obtidas pela multiplicação da raiz quadrada de cada autovalor pelo autovetor correspondente. A Tabela 9 representa uma matriz de p variáveis por m cargas fatoriais. Os valores presentes nesta tabela indicam o valor da correlação linear entre as variáveis (p) com as cargas fatoriais (m).

Tabela 8: Cargas fatoriais.

	carga1	carga2	carga3	carga4
X1	-0.6057	0.0761	0.2952	0.6423
X2	-0.6206	0.3030	-0.6640	-0.0496
X3	-0.8576	0.0463	-0.2015	0.2912
X4	-0.7596	0.2325	0.3943	0.0090
X5	-0.2634	-0.8369	0.3405	-0.1871
X6	0.8737	0.3011	-0.1065	-0.0059
X7	-0.5945	-0.4538	-0.2404	-0.3409
X8	0.8383	-0.1858	0.0371	0.3271
X9	0.2151	-0.6839	-0.4552	0.4359

Na carga fatorial 1, as variáveis que mais apresentam correlação são X3, X4, X6 e X8. As variáveis X5 e X9 apresentam os menores valores de correlação

para essa carga. Para a carga 2, de forma inversa, X5 e X9 apresentam os maiores valores de correlação, e as variáveis X1, X3 e X8 os menores valores.

Na carga fatorial 3, X2 tem o maior peso de correlação, seguida por X9 com um valor intermediário. X6 e X8 possuem os menores valores nessa carga. Já na ultima carga fatorial, 4, a variável X1 apresenta o maior valor de correlação, enquanto X2, X4 e X6 apresentam valores bem baixos.

Para uma análise visual da distribuição das variáveis em relação às cargas fatoriais, complementando a análise numérica acima, apresentam-se na Figura 5 a seguir os gráficos de dispersão das mesmas.

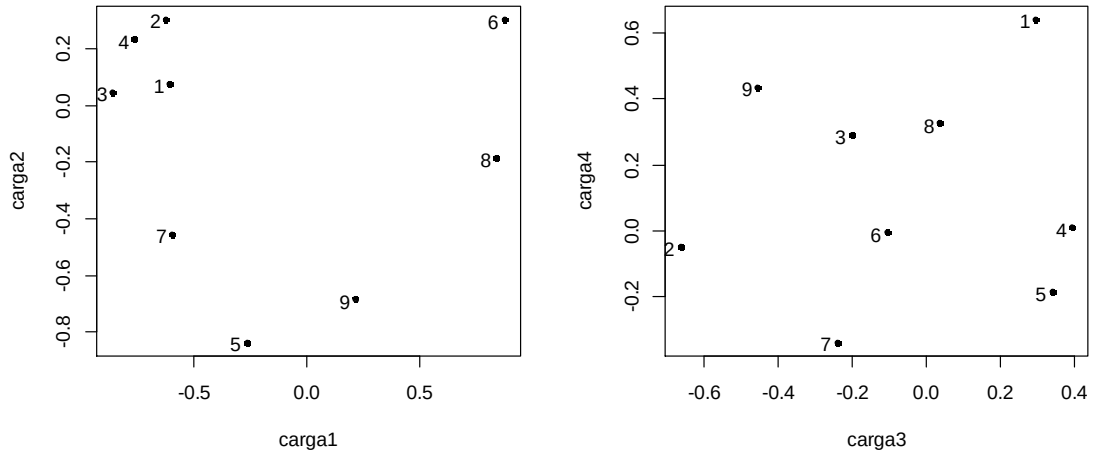

Figura 5: Gráfico de dispersão das cargas fatoriais

De acordo com a Figura 5, pode-se afirmar que as variáveis posicionadas mais distantes do centro influenciam em maior peso cada uma das cargas, sendo no eixo horizontal relativo às cargas 1 e 3, e no eixo vertical às cargas 2 e 4, respectivamente.

Em seguida, na Tabela 9, são mostrados os valores de comunalidade para os quatro fatores. Esses valores representam o quanto estes 4 fatores conseguem explicar cada uma das 9 variáveis. No caso, observou-se que apenas os fatores X4 e X7 apresentaram valores menores que 0,8, ou seja, apenas estas duas variáveis são explicadas em menos de 80 % pelos 4 fatores.

Tabela 9: Valores de Comunalidade para cada uma das variáveis

X1	X2	X3	X4	X5	X6	X7	X8	X9
0.872	0.920	0.863	0.787	0.921	0.865	0.733	0.846	0.911

O erro aleatório de cada variável pode ser observado na diagonal principal da Matriz Psi (gerada pelo software R), na Tabela 10:

Tabela 10: Diagonal da Matriz Psi

X1	X2	X3	X4	X5	X6	X7	X8	X9
0.128	0.080	0.137	0.213	0.079	0.135	0.267	0.154	0.089

Nessa tabela nota-se que os menores erros foram das variáveis X2, X5 e X9, enquanto os maiores foram de X4 e X7, corroborando com os valores de comunalidade, onde foram as variáveis com menor índice de explicação nas 4 cargas fatoriais.

Para facilitar a interpretação dos dados e distribuir de forma mais eficiente as cargas fatoriais, empregam-se rotações dos fatores. O método de rotação aplicado aqui é o Varimax (proposto por Kaiser, 1953), cujo busca em sua rotação maximizar a soma das variâncias das cargas fatoriais e anular o maior número possível de coeficientes (grupos com variáveis similares).

A Tabela 11 apresenta, então, os valores das cargas fatoriais matriz após a transformação Varimax, realizada no software R. Destacam-se nela os maiores valores das variáveis em cada carga. Observa-se que a variável X1 apresenta valor significativo para a carga 1 (0,931), destacando uma alta correlação para este componente. A variável, X2 apresenta valor significativo também para a carga 1 (-0,941), o que caracteriza uma alta correlação negativa. A Variável X5 apresenta destaque na carga 2 (-0,929), o que caracteriza uma alta correlação negativa. A variável X9, possui destaque quanto a carga 2 (-0,938). As variáveis X3, X4, X6 e X7 apresentam valores de destaque para mais de uma carga, indicando pouca eficiência da rotação para as mesmas.

Tabela 11: Cargas fatoriais após Transformação Varimax:

	carga1	carga2	carga3	carga4
X1	0.931	-	-	-
X2	-0.941	0.143	0.119	-
X3	-0.648	-0.151	0.648	-
X4	-0.236	-0.223	0.566	0.601
X5	0.191	-0.929	-0.124	-
X6	0.426	0.619	-0.138	-0.530
X7	-0.538	-0.666	-	-
X8	0.654	0.305	-0.513	-0.250
X9	-0.175	-0.938	-	-

Porém, de modo geral pode-se dizer que a rotação Varimax se apresentou razoavelmente eficiente, uma vez que concentrou determinadas variáveis em cada carga, ou seja, escolhendo apenas o maior valor de cada variável, elas se encontram alocadas nas cargas da seguinte maneira, facilitando a compreensão:

- Carga 1: X1, X2, X3*, X8;
- Carga 2: X5, X6, X7, X9;
- Carga 3: X3*;
- Carga 4: X4.

*valores iguais em módulo

De acordo com a Tabela 12, verifica-se que a proporção da variância para a carga 1 é de 25,5%, para a carga 2 é de 21,2%, para a carga 3 é de 16,7% e para a carga 4 é de 22,4%. Portanto acumulando as quatro cargas, temos um acumulado de 85,8%.

Tabela 12: Proporção de variância e acumulada para cada carga fatorial (CF) após a rotação Varimax.

	carga1	**carga2**	**carga3**	**carga4**
Desv. Pad.	2.299	1.907	1.501	2.012
Proporção Var.	0.255	0.212	0.167	0.224
Var. Acumulada	0.255	0.467	0.634	**0.858**

Por fim, o gráfico das dispersões das variáveis em relação às cargas, agora com a aplicação da rotação Varimax, se encontram na Figura 6. Nota-se uma maior clareza de tendência da distribuição das variáveis em relação à cada carga, em especial na segunda figura:

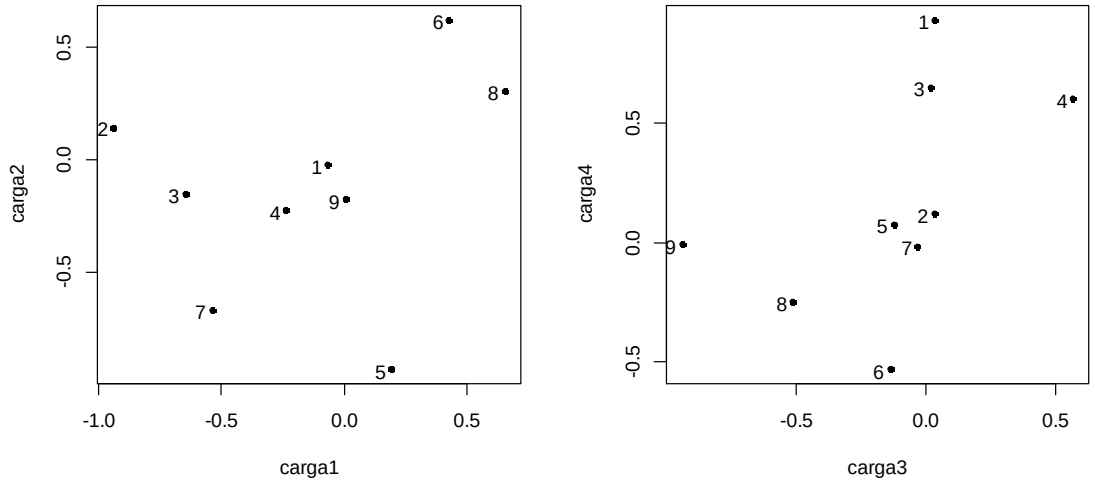

Figura 6: Gráfico de dispersão das cargas fatoriais para os dois fatores aplicando a rotação Varimax.

Analisando o gráfico tem-se que: As variáveis X6 (cereais) e X8 (amêndoas e sementes oleaginosas) são explicadas concomitantemente na carga 1 e 2, com intensidade semelhante e de forma direta, enquanto que na carga 4 ambas influenciam de forma inversa (negativo); X2 (carne de frango) apresenta o maior valor negativo para carga 1, ou seja, influenciando a mesma de forma inversa, enquanto que possui valores próximos à nulidade nas outras três cargas fatoriais; De forma análoga, a variável X1 (carne de gado), apresenta valor relevante e direto apenas para a carga 4; X4 (leite) possui valores relativamente altos e positivos (diretos) para a carga 3 e 4 simultaneamente.

Alternativa C) Faça a análise de agrupamento dos indivíduos, considerando como medida de dissimilaridade a distância euclidiana. Interprete os resultados obtidos.

Análise Agrupamento – Indivíduos

A análise de agrupamento, também chamada de clustering, é o nome dado para o grupo de técnicas computacionais onde o propósito é separar objetos e grupos se baseando nas características que estes objetos possuem.

Na Tabela 13 está distância euclidiana para as 24 parcelas, sendo esta usada como medida de dissimilaridade para a análise de agrupamento.

Tabela 13: Matriz das distâncias euclidianas para 25 países e 9 tipos de proteínas.

	ALB	AUT	BEL	BGR	TCH	DNK	AOR	FIN	FRA	GRC	HUN	IRL	ITA	HOL	NOR	POL	PRT	ROU	ESP	SWE	SUI	ENG	URS	AOC
AUT	23.62																							
BEL	22.02	7.87																						
BGR	15.68	32.30	32.79																					
TCH	15.39	10.31	10.61	24.01																				
DNK	30.65	11.96	11.12	40.33	19.42																			
AOR	22.97	10.74	8.93	33.62	10.61	15.19																		
FIN	31.70	17.41	17.61	40.33	24.02	12.25	23.83																	
FRA	23.60	11.01	6.01	33.26	13.44	12.72	13.86	18.18																
GRC	12.92	19.09	17.99	19.21	14.59	24.22	21.81	24.05	17.97															
HUN	13.24	16.98	18.78	18.40	9.18	26.74	17.52	29.98	21.26	14.44														
IRL	27.95	9.75	8.95	37.68	17.13	9.16	16.17	11.45	9.86	22.13	24.51													
ITA	11.06	14.69	13.57	21.01	8.71	21.60	15.62	23.74	15.16	7.80	10.70	19.54												
HOL	28.77	6.76	9.68	38.53	16.35	8.36	13.27	14.67	12.35	23.72	23.21	7.06	19.87											
NOR	27.30	13.69	10.80	38.18	18.73	6.69	14.99	11.68	13.15	21.37	25.67	10.99	18.77	11.55										
POL	18.19	9.94	12.20	24.49	8.26	17.81	14.88	19.36	14.01	11.91	11.99	15.41	9.07	15.34	16.91									
PRT	22.97	22.93	19.20	33.29	19.06	23.93	15.16	31.18	21.85	22.15	22.03	27.06	17.87	25.39	20.66	21.68								
ROU	10.55	25.26	25.88	8.34	17.31	33.29	26.73	33.39	27.12	13.17	11.64	30.65	14.31	31.25	31.02	17.43	27.65							
ESP	17.49	17.68	14.08	29.06	13.40	21.09	11.92	26.62	17.29	16.34	16.51	21.82	11.20	20.85	17.47	15.72	8.15	22.52						
SWE	30.47	13.03	11.63	41.48	20.43	4.80	15.58	11.90	14.03	25.02	27.66	9.32	21.88	8.50	5.56	19.15	24.08	34.20	20.79					
SUI	25.47	7.53	7.53	35.49	14.95	9.63	14.62	13.01	8.24	19.76	22.11	4.92	16.65	6.24	10.62	13.46	24.82	28.55	19.49	9.51				
ENG	24.75	12.92	6.83	36.42	16.50	11.74	14.78	15.95	6.99	20.23	24.17	8.25	17.26	11.99	10.52	17.07	22.77	29.73	17.80	10.90	8.02			
URS	11.68	19.04	18.42	16.68	12.64	25.34	21.37	24.70	19.43	8.06	12.49	22.34	9.46	24.23	22.83	10.81	24.01	9.88	18.08	26.29	20.96	21.61		
AOC	29.46	10.13	9.07	40.62	17.15	9.90	10.56	18.88	12.46	26.00	24.70	10.18	21.10	6.53	12.15	18.51	22.84	33.44	19.22	9.09	9.61	11.10	26.58	
YUG	15.54	31.95	32.69	4.88	23.98	39.87	33.28	39.33	33.72	18.68	17.62	37.28	20.72	37.97	37.45	23.73	32.83	6.91	28.33	40.83	35.25	36.41	15.79	40.28

Para definir os métodos hierárquicos de agrupamento de indivíduos a serem aplicados nesse trabalho, foram primeiramente calculados seus respectivos coeficientes de correlação cofenética (ccc) (Tabela 14). Esse coeficiente aponta para o grau de ajuste de cada agrupamento.

Tabela 14: Coeficiente de correlação cofenética (ccc).

Método	ccc
LIGAÇÃO SIMPLES	0.557
LIGAÇÃO COMPLETA	0.764
LIGAÇÃO MÉDIA	0.730
CENTRÓIDE	0.664
WARD	0.636

Os métodos com maiores valores foram Ligação Completa (0,76) e Ligação Média (0,73), ambos considerados adequados segundo Rohlf (1970). Porém, o método de ligação completa pode sofrer influências na presença de pontos discrepantes (como visto na Figura 1), então optou-se por usar a Ligação Média. Realizando uma segunda análise, para fruto de comparação de agrupamento, escolheu-se o método seguinte na ordem dos coeficientes, Método do Centróide (0,66). Procedeu-se então a análise dos dendrogramas desses dois métodos.

- **LIGAÇÃO MÉDIA (Average Linkage):**

Na Figura 7, pode ser visto o dendrograma construído através do método de ligação média no software R.

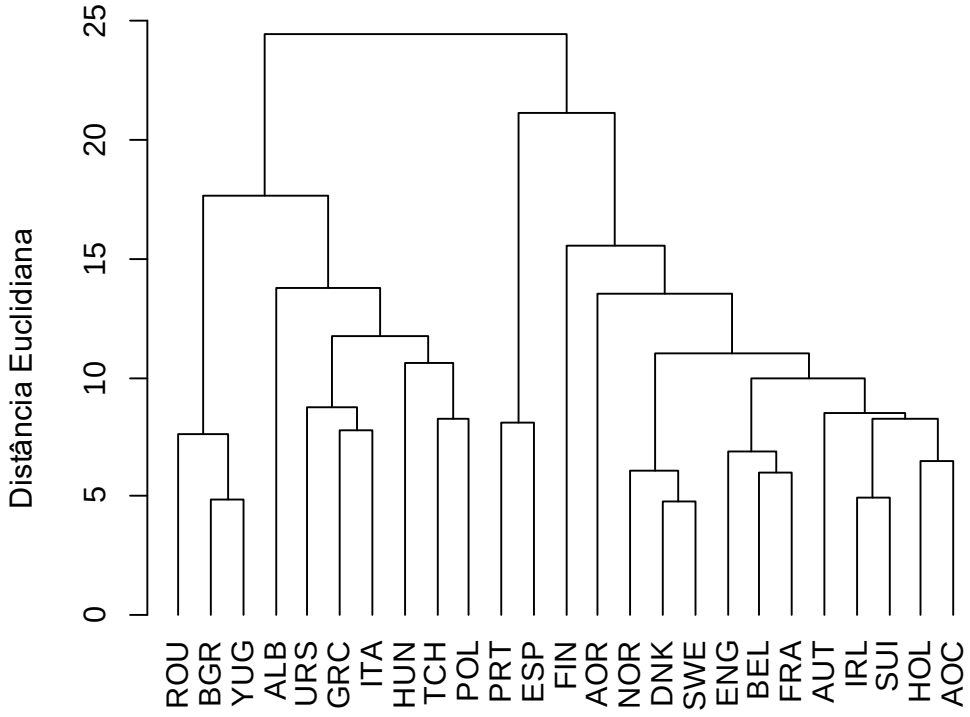

Figura 7: Dendrograma método da Ligação Média

Para definir o número de grupos, isto é, o ponto de corte, verificam-se os valores de junção do dendrograma na Tabela 15, buscando-se os maiores saltos:

Tabela 15: Valores de junção no dendrograma (Ligação Média)

Posição		Diferença	Posição		Diferença
1	4.796		...		
		0.080	13	8.543	
2	4.875				0.217
		0.047	14	8.760	
3	4.922				1.194
		1.085	15	9.953	
4	6.007				0.633
		0.115	16	10.586	
5	6.123				0.416
		0.402	17	11.002	
6	6.525				0.705
		0.383	18	11.707	
7	6.909				1.809
		0.714	19	13.516	
8	7.623				0.232
		0.182	20	13.748	
9	7.804				1.821
		0.342	21	15.568	
10	8.146				**2.042**
		0.109	22	17.610	
11	8.255				**3.466**
		0.016	23	21.076	
12	8.271				**3.335**
		0.272	24	24.411	
13	8.543				

Os maiores valores apontam para o agrupamento entre 2, 3 e 4 clusters, sendo k=3 o valor numericamente mais adequado. Porém, de modo a escolher uma divisão em que os grupos tenham dimensões mais homogêneas, optou-se por usar k=4. Logo, o dendrograma com a divisão entre os 4 clusters se apresenta a seguir (Figura 8).

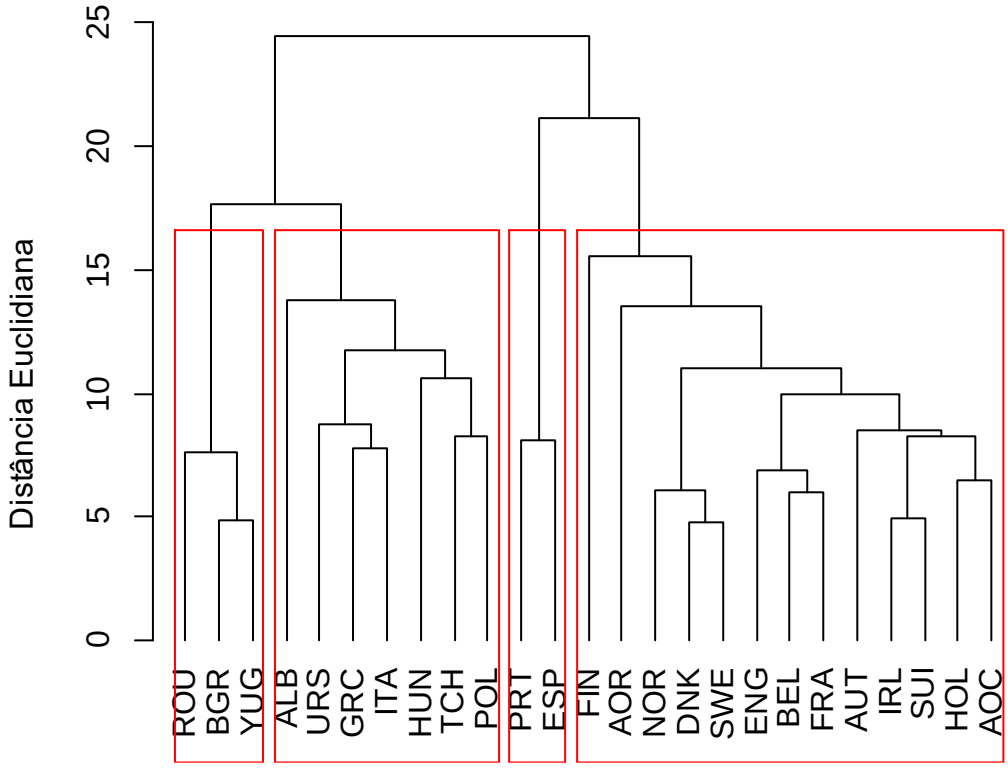

Figura 8: Dendrograma (Ligação Média) com a divisão entre os Clusters

O grupo 1 é formado pelos países Romênia, Bulgária e Iugoslávia, cujos países são os que apresentam os maiores consumos de cereais (X6); o grupo 2 é formado pela Albânia, Ex-URSS, Grécia, Itália, Hungria, Tchecoslováquia, Polônia, e são os países que possuem em comum os maiores valores de consumo de cereais (X6) após os citados no grupo 1; o grupo 3 é formado pelos países Portugal e Espanha, que apresentam os maiores valores de consumo de proteínas de frutas e vegetais (X9); e finalmente o grupo 4 que é formado por: Finlândia, Ex-Alemanha Oriental, Noruega, Dinamarca, Suécia, Inglaterra, Bélgica, França, Áustria, Irlanda, Suíça, Holanda e Ex- Alemanha Ocidental, que apresentam como características comuns os maiores consumos de ovos (X3) e leite (X4), ao mesmo passo que os menores valores de cereais (X6) e amêndoas e sementes oleaginosas (X8).

Nota-se nesse agrupamento que a variável X6 foi responsável por influenciar de forma direta na classificação dos clusters, sendo justamente a que apresentou a maior média entre as variáveis.

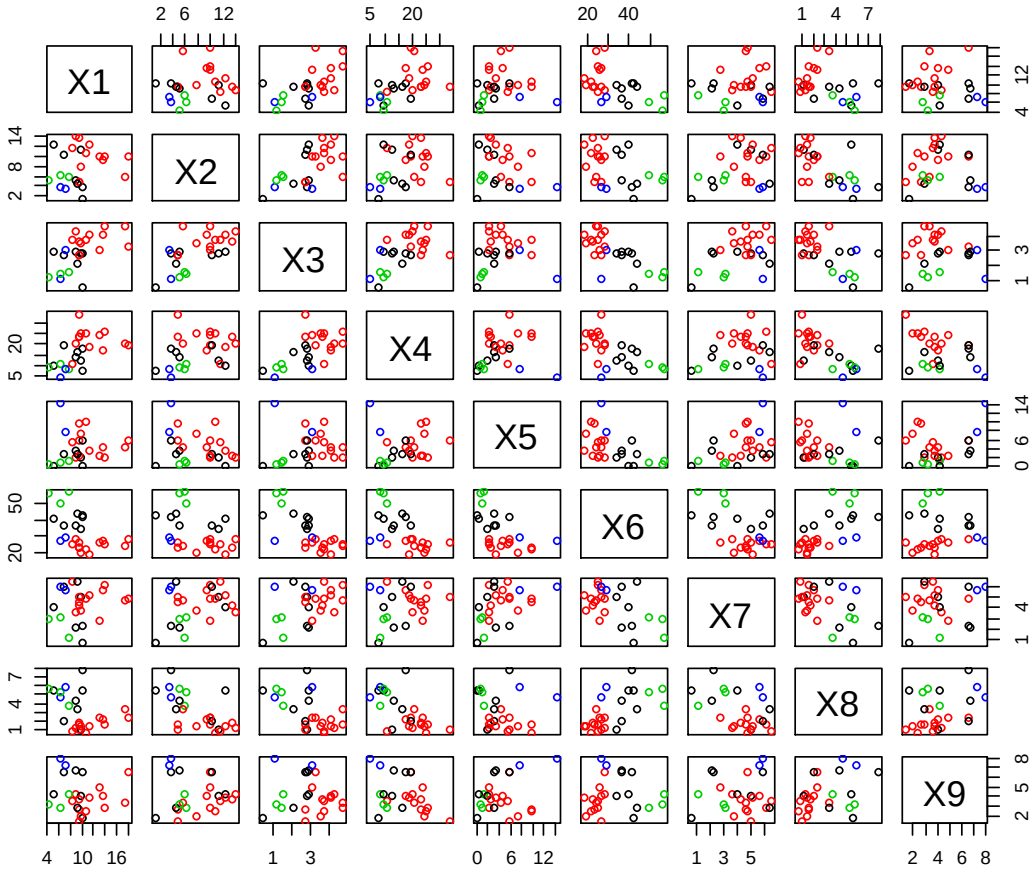

Figura 9: Diagrama de dispersão a cada duas variáveis.

O diagrama de dispersão está apresentado na Figura 9 e nele, pode ser observado a correlação entre as variáveis combinadas duas a duas. As variáveis que representam as fontes de proteína (animal ou vegetal) consumidas diariamente pelas pessoas nos 25 países da Europa são: carne de gado (X1), carne de frango (X2), ovos (X3), leite (X4), peixe (X5), cereais (X6), alimentos enriquecidos (X7), amêndoas e sementes oleaginosas (X8) e frutas/vegetais (X9). Nota-se uma dispersão dos dados, não apresentando tendência linear aparente para as combinações de proteínas X1-X2, X1-X5, X1-X7, X1-X8, X1-X9, X2-X4, X2-X5, X2-X7, X2-X9, X3-X5, X3-X9, X4-X5, X4-X7, X4-X8, X5-X8, X5-X9, X6-X9, X7-X8, X7-X9 e X8-X9.

Assim, a não significância do coeficiente de correção linear de Pearson, indica que não existe dependência estatística entre as variáveis. No diagrama, também pode ser observada uma tendência linear positiva para as correlações entre X1-X3, X1-X4, X2-X3, X3-X4, X3-X7, X6-X8 e X5-X7. E tendência linear negativa para as combinações entre X1-X6, X2-X6, X2-X8, X3-X6, X3-X8, X4-X6, X4-X9, X5-X6 e X6- X7.

- **MÉTODO DO CENTRÓIDE:**

De forma análoga ao método anterior, gerou-se o dendrograma no software R (Figura 10).

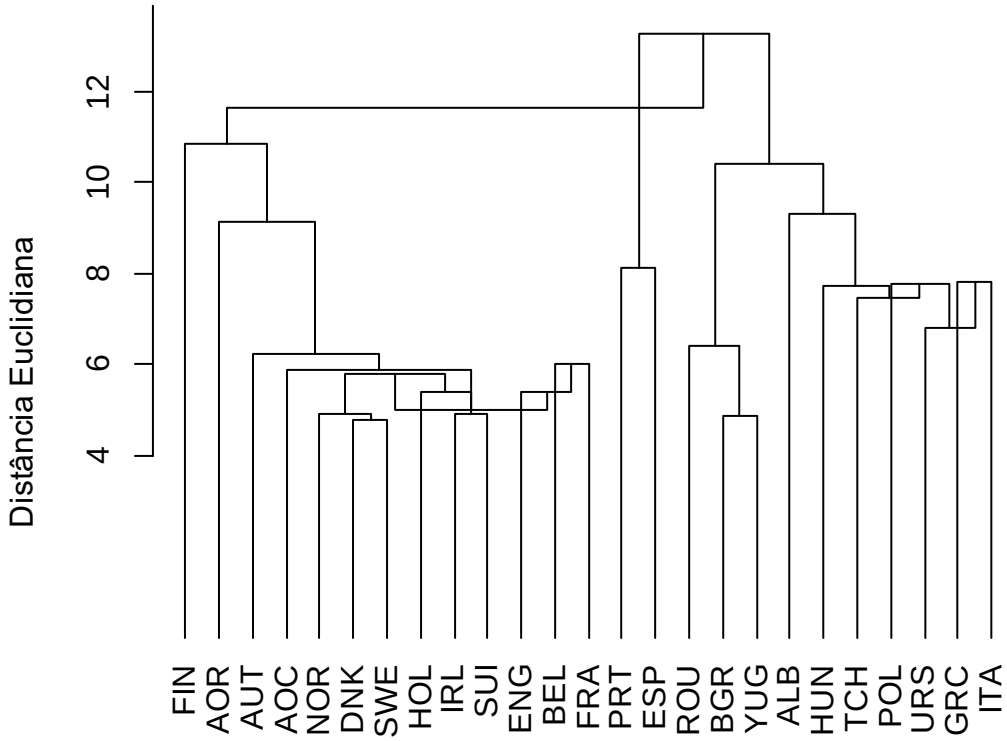

Figura 8: Dendrograma pelo método centroide.

Para definir o número de grupos, isto é, o ponto de corte, verificam-se os valores de junção do dendrograma na Tabela 16, buscando os maiores saltos:

Tabela 16: Valores de junção no dendrograma (Centróide).

Posição		Diferença		Posição		Diferença
1	4.796			...		
		0.080		13	7.804	
2	4.875					-0.996
		0.047		14	6.809	
3	4.922					0.975
		0.002		15	7.783	
4	4.924					-0.305
		0.495		16	7.478	
5	5.419					0.232
		0.398		17	7.710	
6	5.817					0.436
		0.190		18	8.146	
7	6.007					0.998
		-0.601		19	9.144	
8	5.407					0.156
		-0.384		20	9.300	
9	5.022					1.123
		0.861		21	10.423	
10	5.883					0.439
		0.339		22	10.863	
11	6.222					**2.384**
		0.182		23	13.247	
12	6.404					-1.598
		1.400		24	11.649	
13	7.804					

Pela orientação do maior salto do dendrograma e pela maior diferença entre os agrupamentos, foram separados em grupos as parcelas no dendrogama construído através do método da Centróide. Maior valor em k=3 clusters.

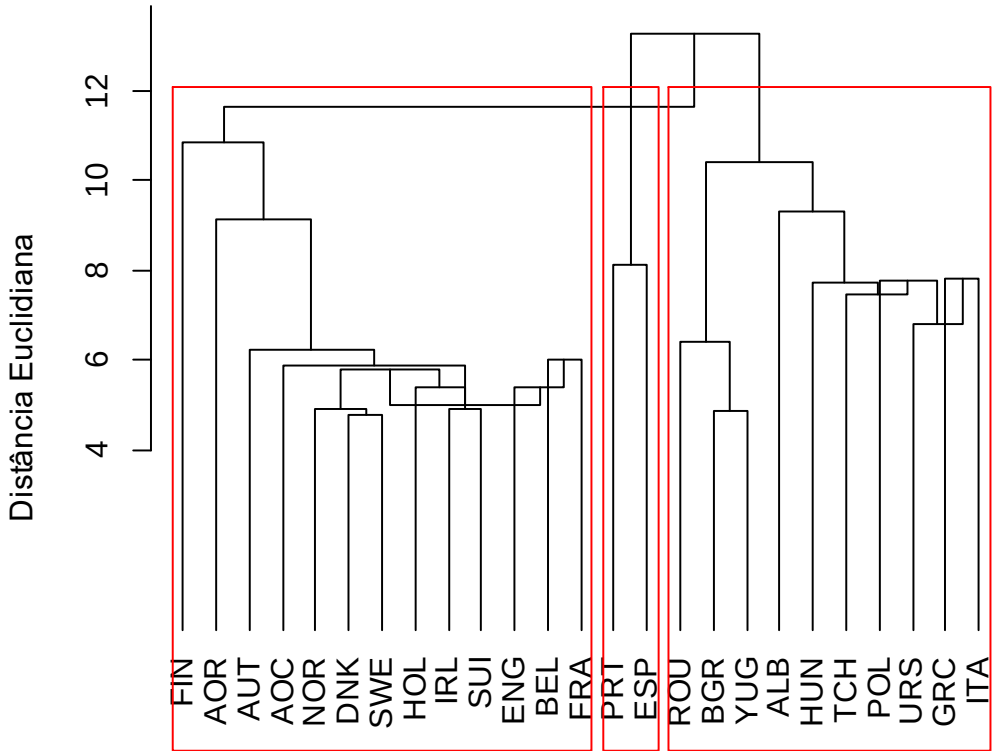

Figura 9: Dendrograma (Método do Centróide) com os grupos separados.

Nesse clustering os grupos de países formados foram:

Grupo 1: Finlândia, Ex-Alemanha Oriental, Áustria, Ex- Alemanha Ocidental, Noruega, Dinamarca, Suécia, Holanda, Irlanda, Suíça, Inglaterra, Bélgica e França (idêntico ao grupo 4 da análise anteior).

Grupo 2: Portugal e Espanha (idêntico ao grupo 3 da análise anterior)

Grupo 3: Romênia, Bulgária e Iugoslávia, Albânia, Hungria, Tchecoslováquia, Polônia, Ex-URSS, Grécia e Itália (equivalente ao grupo 1 + grupo 2 da análise anterior, ambos com os maiores valores de X6 (cereais));

- **MÉTODO K-MEANS (K-Médias):**

Uma terceira análise foi empregada, nessa aplicando-se um método de agrupamento não-hierárquico (K-means). Nesse método, escolhe-se inicialmente o número de grupos em que serão alocados os indivíduos, e para definir o número ideal foram empregados dois métodos no software R. O primeiro método utilizado foi o de Elbow, que é basicamente a analise do gráfico do "cotovelo". O segundo método utilizado foi o do gráfico de índice de Hubert e D. Ambos aplicados no software R e com os resultados na Figura 10 e 11, respectivamente.

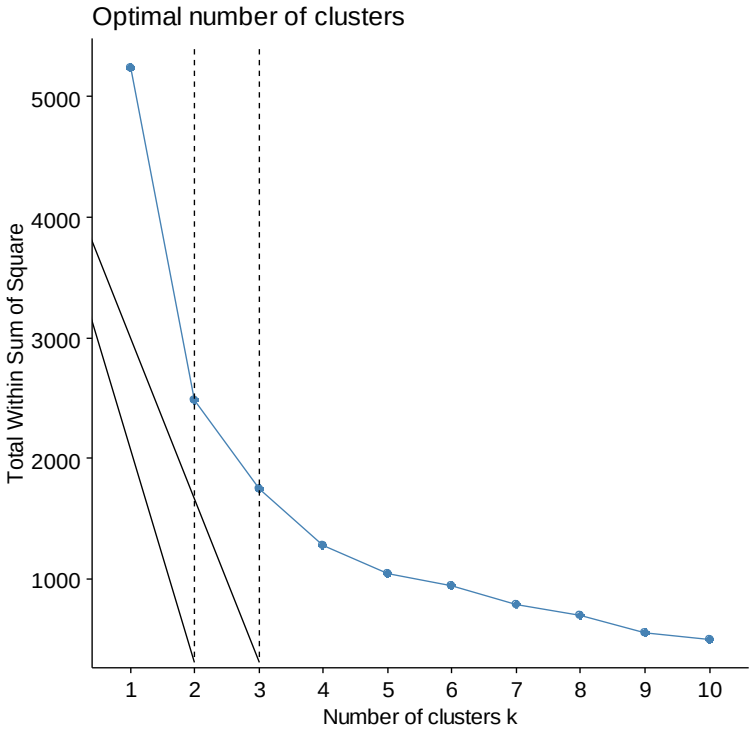

Figura 10: Gráfico do número ótimo de agrupamento de Elbow.

Busca-se a forma de cotovelo nessa curva. Porém, como a análise é subjetiva, o número de clusters poderia ser entre 2 e 3. Uma análise numérica pode se mostrar mais apropriada para a definição.

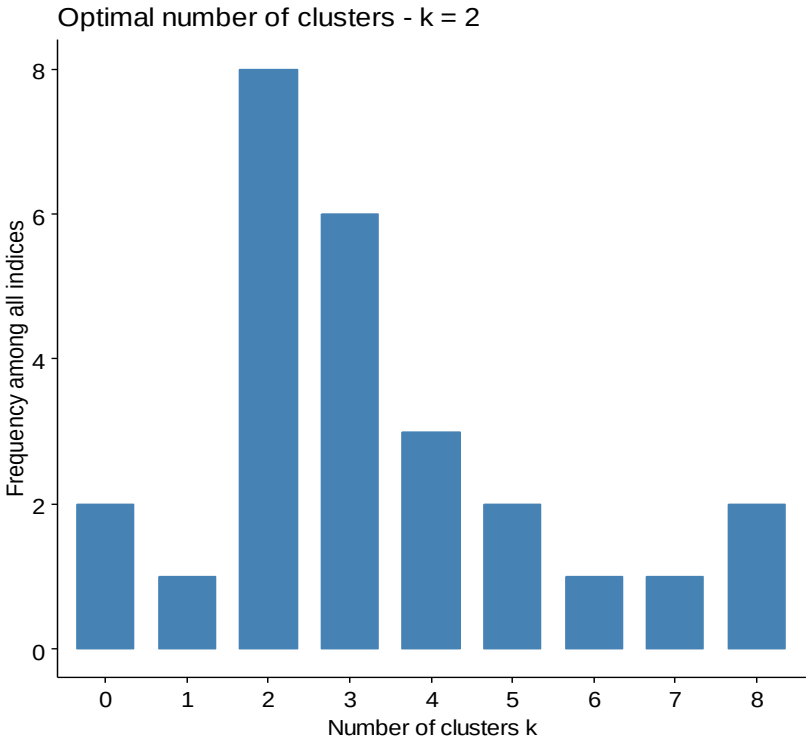

Figura 11: Gráfico do número ótimo de agrupamento de Índice Hubert e D,

Nesse método, a conclusão da quantidade de grupos mais apropriada para a análise desses dados é k=2. Segue-se então a aplicação do método K-means, empregando-se a divisão dos indivíduos em 2 clusters.

Após aplicar o método no software R, a divisão dos indivíduos (países) é apresentada na tabela 17:

Tabela 17: Divisão dos indivíduos (Método K-means).

País	Cluster	País	Cluster
AUT	1	ALB	2
BEL	1	BGR	2
DNK	1	TCH	2
AOR	1	GRC	2
FIN	1	HUN	2
FRA	1	ITA	2
IRL	1	POL	2
HOL	1	PRT	2
NOR	1	ROU	2
SWE	1	ESP	2
SUI	1	URS	2
ENG	1	YUG	2
AOC	1		

	Cluster 1	Cluster 2
TOTAL	**13**	**12**

A análise visual da divisão dos países nos clusters, de acordo com a Tabela 19, pode ser feita na Figura 12 a seguir:

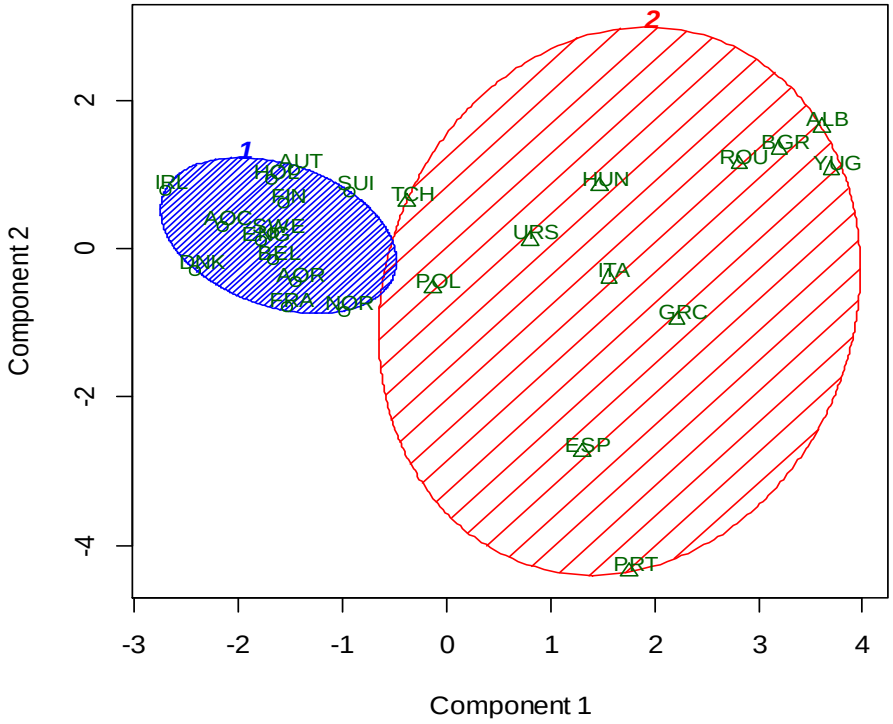

Figura 12: Gráfico de agrupamento de clusters pelo K-Means

Foram formados 2 grupos então com os seguintes países:

- Grupo 1: Áustria, Bélgica, Dinamarca, Ex. Alemanha Oriental, Finlândia, França, Holanda, Irlanda, Noruega, Suécia, Suíça, Inglaterra e Ex. Alemanha Ocidental. Grupo esse caracterizado majoritariamente por altos consumos de X3 (ovos) e X4 (leite) e baixos valores de consumo de X6 (cereais) e X8 (amêndoas e sementes oleaginosas).

- Grupo 2: Albânia, Bulgária, Tchecoslováquia, Grécia, Hungria, Itália, Polônia, Portugal, Romênia, Espanha, Ex. URSS e Iugoslávia. Grupo esse caracterizado majoritariamente por altos valores de consumo de X6 (cereais) e X8 (amêndoas e sementes oleaginosas) e baixos consumos de X3 (ovos) e X4 (leite).

Os dois grupos se mostraram inversamente proporcionais nas variáveis de influência, em termos das fontes de ingestão de proteína em g/pessoa. As variáveis que mais influenciaram nesse agrupamento (X3, X4, X6 e X7) foram as mesmas de destaque nos métodos anteriores.

Alternativa D) Faça a análise de agrupamento das variáveis, considerando como medida de similaridade baseada na correlação linear de Pearson. Interprete os resultados obtidos.

Inicialmente, calcula-se como medida de similaridade a matriz de correlação de Pearson (R), que já foi apresentada na Tabela 3 da Alternativa A.

A matriz de dissimilaridade é obtida por meio da seguinte equação, sugerida por Rencher (2002):

$$\boldsymbol{r} = 1 - R^2$$

A matriz **r** resultante está relacionada com as distâncias das variáveis da matriz de correlação de Pearson. Tal matriz está representada na Tabela 18.

Tabela 18: Matriz de distância das variáveis de Rencher (r):

	X1	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**
X2	0.976							
X3	0.656	0.608						
X4	0.751	0.914	0.660					
X5	0.997	0.941	0.996	0.983				
X6	0.753	0.832	0.497	0.650	0.724			
X7	0.981	0.905	0.795	0.948	0.833	0.719		
X8	0.878	0.615	0.687	0.611	0.983	0.577	0.777	
X9	0.994	0.997	0.998	0.841	0.921	0.998	0.992	0.859

Sabendo que quanto maior a distância, menor a similaridade entre os valores, pode-se afirmar, de acordo com a matriz r supracitada, que as maiores dissimilaridades (em vermelho) são:

1. Carne de gado e peixe (0,997)
2. Carne de gado e frutas e vegetais (0,994)
3. Carne de frango e frutos e vegetais (0,995
4. Ovos e peixe (0,995) e ovos com frutas e vegetais (0,997)
5. Cereais e frutas e vegetais (0,998)
6. Alimentos enriquecidos e frutas e vegetais (0,992)

Em contrapartida, as maiores medidas de similaridade (em verde) ficam a cargo das seguintes variáveis:

1. Ovos e cereais (0,497)
2. Cereais e amêndoas e sementes oleaginosas (0,5777)

De forma análoga à alternativa anterior (Alternativa C), para a definição dos métodos hierárquicos de agrupamento de variáveis a serem empregados, calcularam-se os respectivos coeficientes de correlação cofenética (ccc) (Tabela 19).

Tabela 19: Coeficiente de correlação cofenética (ccc).

Método	ccc
LIGAÇÃO SIMPLES	0.728
LIGAÇÃO COMPLETA	0.699
LIGAÇÃO MÉDIA	0.788
CENTRÓIDE	0.737
WARD	0.714

Aqui, pode-se afirmar que praticamente todos os outros métodos são adequados para serem empregados no agrupamento, segundo Rencher (2002), sendo acima do valor 0,70 (com exceção de Ligação Completa, porém por valor de diferença ínfimo). O método com maior Coeficiente de correlação cofenética e escolhido para análise foi Ligação Média (0,78). Optou-se aqui também por empregar outros métodos, à fins de comparação, escolhendo-se o Método de Ligação Simples, pois foi o segundo maior ccc, e o Método de Ward que, segundo estudos, junto com o primeiro método escolhido, tendem a apresentarem melhores desempenhos.

- **LIGAÇÃO MÉDIA (Average Linkage):**

Na Figura 13, pode ser visto o dendrograma construído através do método de ligação média no software R.

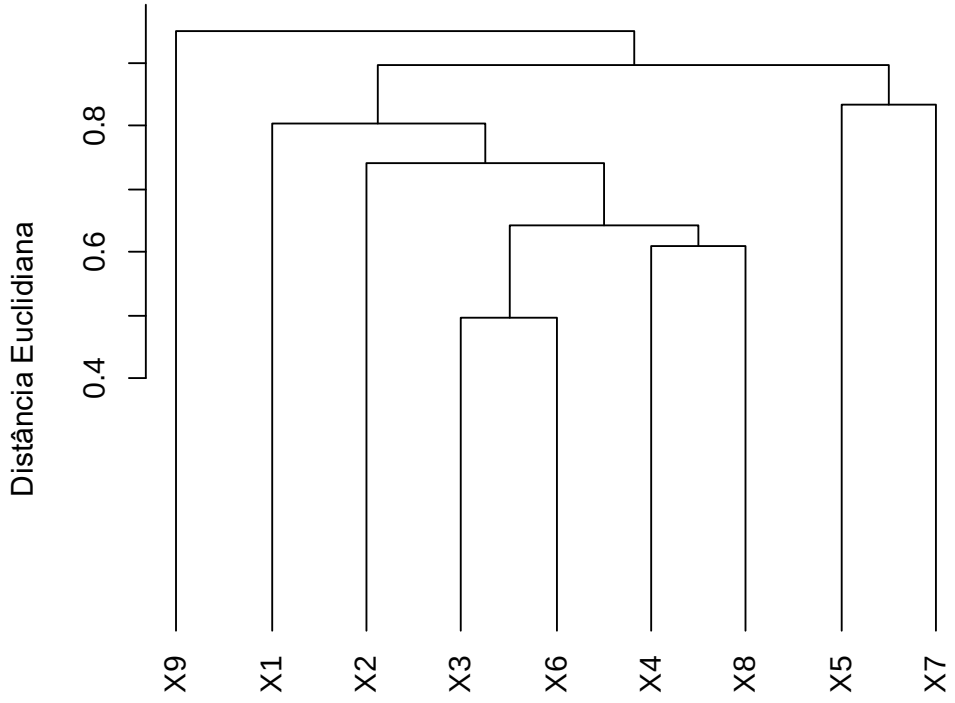

Figura 13: Dendrograma método da Ligação Média

Para definir o número de grupos, verificam-se os valores de junção do dendrograma na Tabela 20, buscando-se novamente os maiores saltos:

Tabela 20: Valores de junção no dendrograma (Ligação Média)

Posição		Diferença
1	0.497	
		0.114
2	0.611	
		0.032
3	0.643	
		0.099
4	0.742	
		0.061
5	0.803	
		0.031
6	0.833	
		0.062
7	0.896	
		0.055
8	0.950	

Os maiores saltos foram para os respectivos k=8 e, por segundo, k=6. Como o objetivo do método é agrupar variáveis em grupos semelhantes, optou-se pelo valor menor de grupos entre as duas opções e, consequentemente, mais variáveis agrupadas. Assim, procedeu-se o agrupamento no dendograma com 6 clusters (Figura 14).

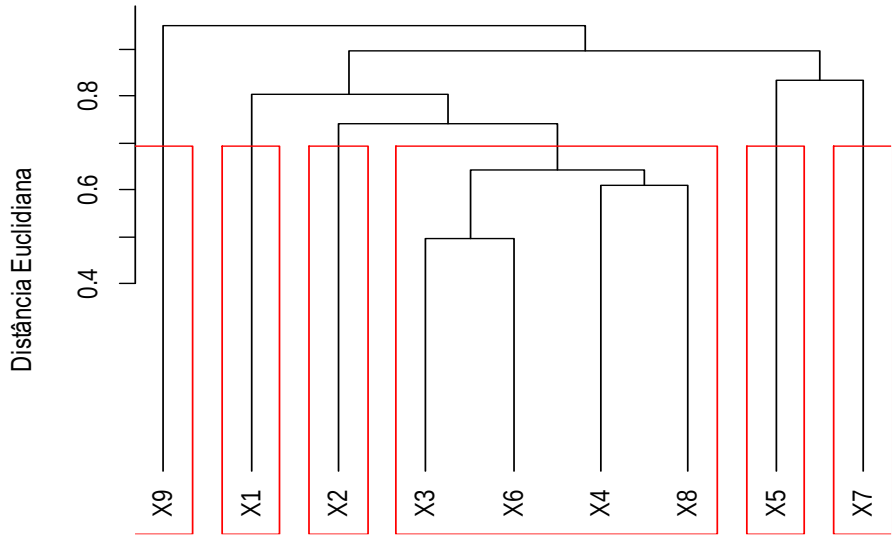

Figura 14: Dendrograma (Ligação Média) com a divisão entre os Clusters

- Grupo 1 formado por frutas e vegetais;
- Grupo 2 formado por carne de gado;
- Grupo 3 formado por carne de frango;

- Grupo 4 formado por: ovos, cereais, leite e amêndoas e sementes oleaginosas;
- Grupo 5 formado por peixe;
- Grupo 7 formado por alimentos enriquecidos;

Nota-se que nesta análise apenas ocorreu o agrupamento das variáveis X3, X4, X6 e X8, mantendo todas as outras em grupos separados e unitários.

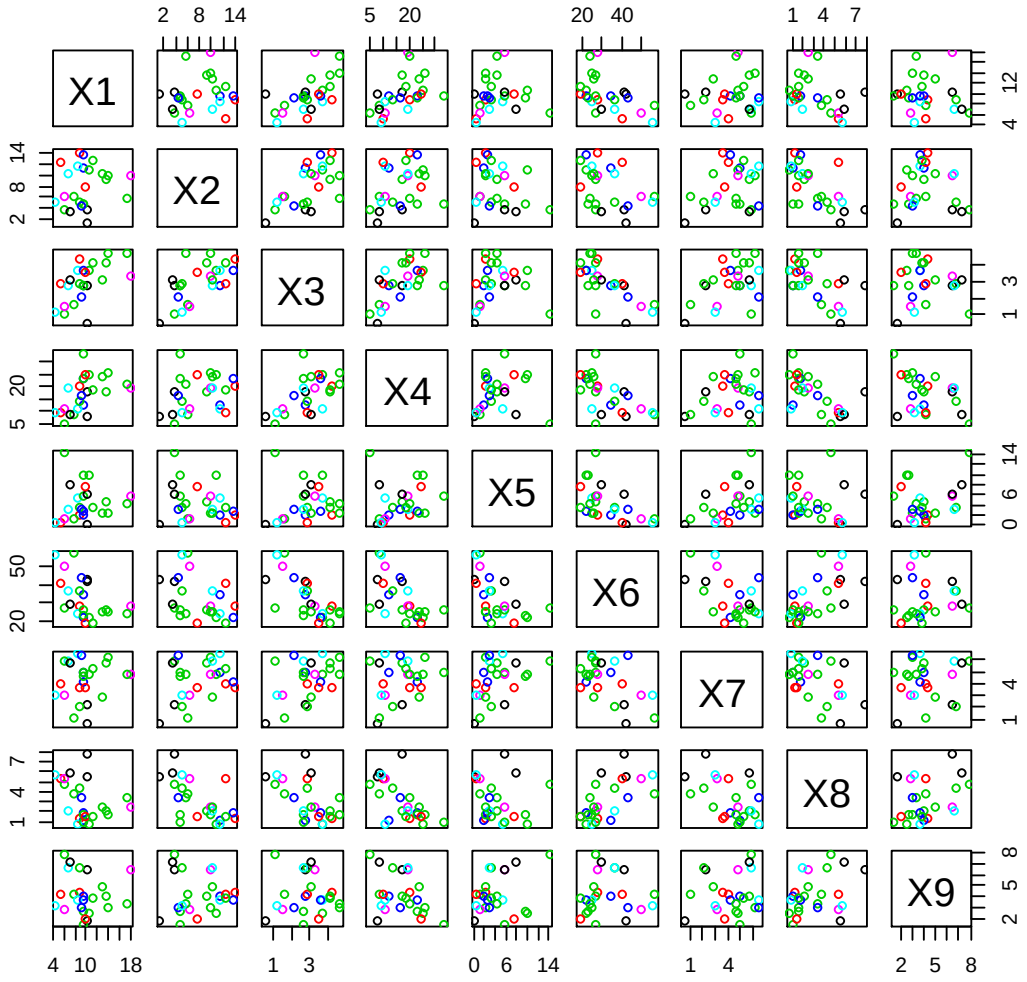

Figura 15: Diagrama de dispersão a cada duas variáveis.

O diagrama de dispersão está apresentado na Figura 15 e nele, pode ser observado a correlação entre as variáveis combinadas duas a duas. As variáveis que representam as fontes de proteína (animal ou vegetal) consumidas diariamente pelas pessoas nos 25 países da Europa são: carne de gado (X1), carne de frango (X2), ovos (X3), leite (X4), peixe (X5), cereais (X6), alimentos enriquecidos (X7), amêndoas e sementes oleaginosas (X8) e frutas/vegetais (X9). Nota-se uma dispersão dos dados, não apresentando tendência linear aparente para as combinações de proteínas X1-X2, X1-X5, X1-X7, X1-X8, X1-

X9, X2-X4, X2-X5, X2-X7, X2-X9, X3-X5, X3-X9, X4-X5, X4-X7, X4-X8, X5-X8, X5-X9, X6-X9, X7-X8, X7-X9 E X8-X9.

Assim, a não significância do coeficiente de correção linear de Pearson, indica que não existe dependência estatística entre as variáveis. No diagrama, também pode ser observada uma tendência linear positiva para as correlações entre X1-X3, X1-X4, X2-X3, X3-X4, X3-X7, X6-X8 e X5-X7. E tendência linear negativa para as combinações entre X1-X6, X2-X6, X2-X8, X3-X6, X3-X8, X4-X6, X4-X9, X5-X6 e X6- X7.

- **LIGAÇÃO SIMPLES (Single Linkage):**

Na Figura 16, pode ser visto o dendrograma construído através do método de ligação simples no software R.

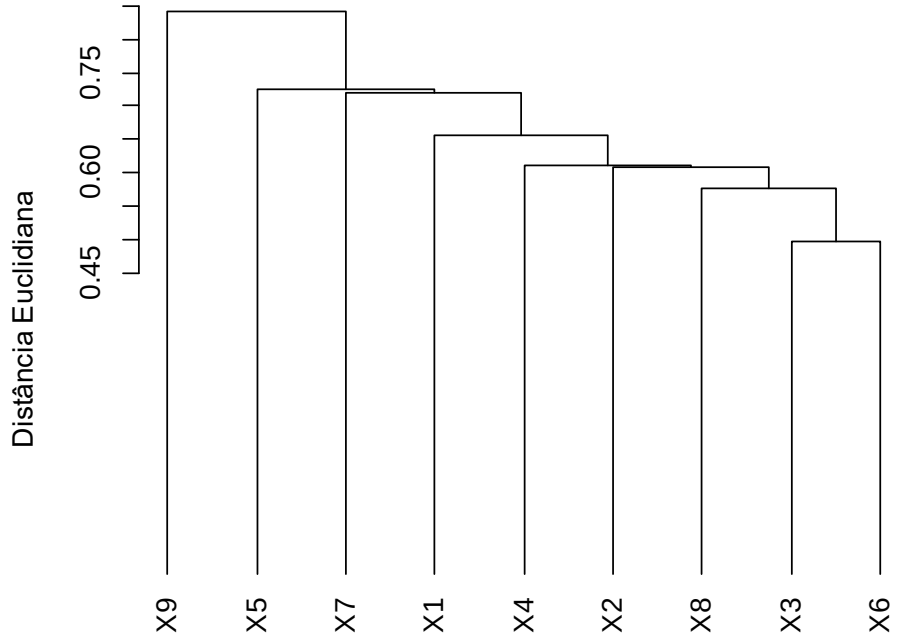

Figura 16: Gráfico dendograma considerando o método de Ligação Simples

Da mesma forma ao método anterior, a Tabela 21 apresenta os valores dos saltos no dendograma para definição do número ideal de grupos.

Tabela 21: Valores de junção no dendrograma (Ligação Simples)

Posição		Diferença
1	0.497	
		0.080
2	0.577	
		0.031
3	0.608	
		0.003
4	0.611	
		0.045
5	0.656	
		0.063
6	0.719	
		0.005
7	0.724	
		0.117
8	0.841	

Nesse caso o maior salto ocorre em k=2. Logo, o agrupamento do dendograma com 2 clusters ocorre na Figura 17 a seguir:

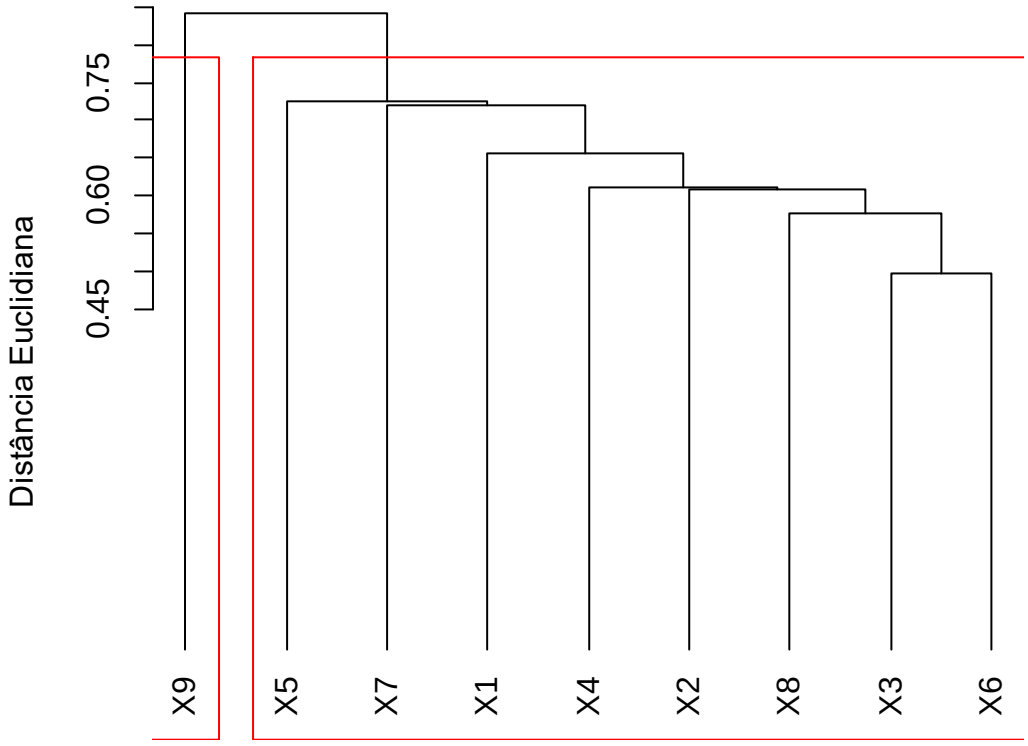

Figura 17: Dendograma (Ligação Simples) com 2 clusters.

Segundo a análise feita do agrupamento desse método, a única variável que ficou separa foi X9 (frutas e cereais), mostrando-se como uma variável com influências diferentes das demais.

- **MÉTODO DE WARD:**

Último método de análise, o dendrograma construído através do método de Ward no software R está exposto na Figura 18.

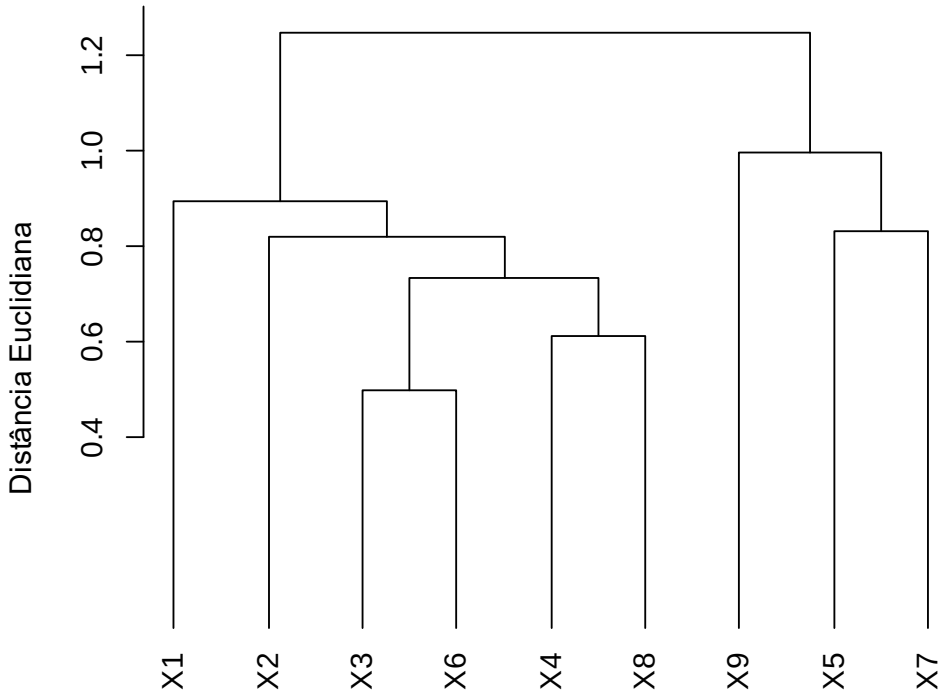

Figura 18: Gráfico dendograma considerando o método de Ward.

A Tabela 22 trás os valores das alturas das junções do dendograma para definir-se a quantidade de grupos.

Tabela 22: Valores de junção no dendrograma (Método de Ward)

Posição		Diferença
1	0.497	
		0.114
2	0.611	
		0.122
3	0.733	
		0.087
4	0.820	
		0.014
5	0.833	
		0.061
6	0.895	
		0.103
7	0.998	
		0.249
8	1.247	

Para esse método, o maior salto ocorreu em k=2, e o segundo maior salto em k=7. Como os valores de k foram nos dois extremos, optou-se por nesse método realizar os dois pontos de corte no dendograma para avaliar as duas diferentes separações de grupos. O agrupamento do dendograma com 2 clusters está na Figura 19 e o agrupamento com 7 clusters está na Figura 20:

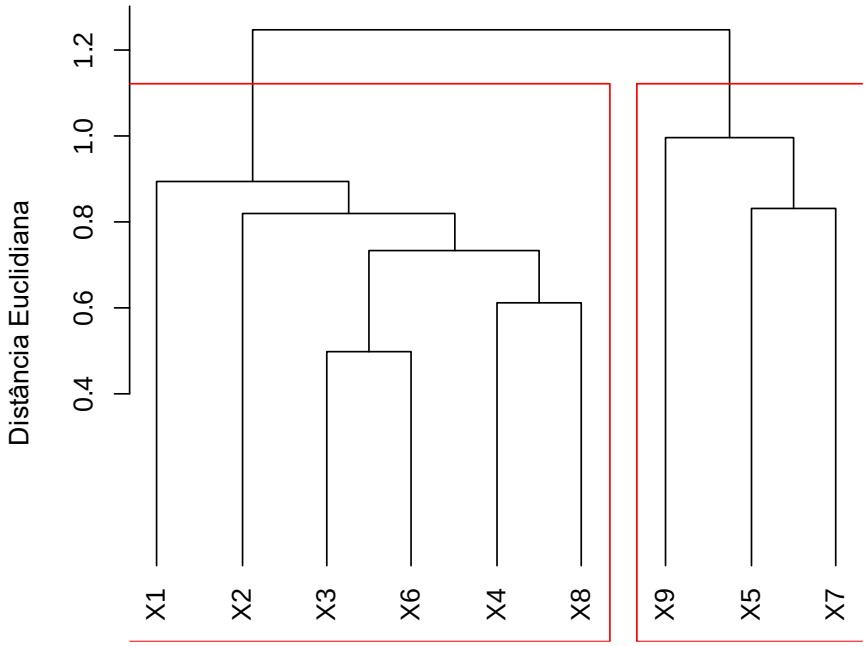

Figura 19: Agrupamento do dendograma (Método Ward) com 2 clusters.

Nesse agrupamento, as variáveis X1, X2, X3, X6, X4 e X8 fazem parte do grupo 1, enquanto as variáveis X9, X5 e X7, fazem parte do grupo 2.

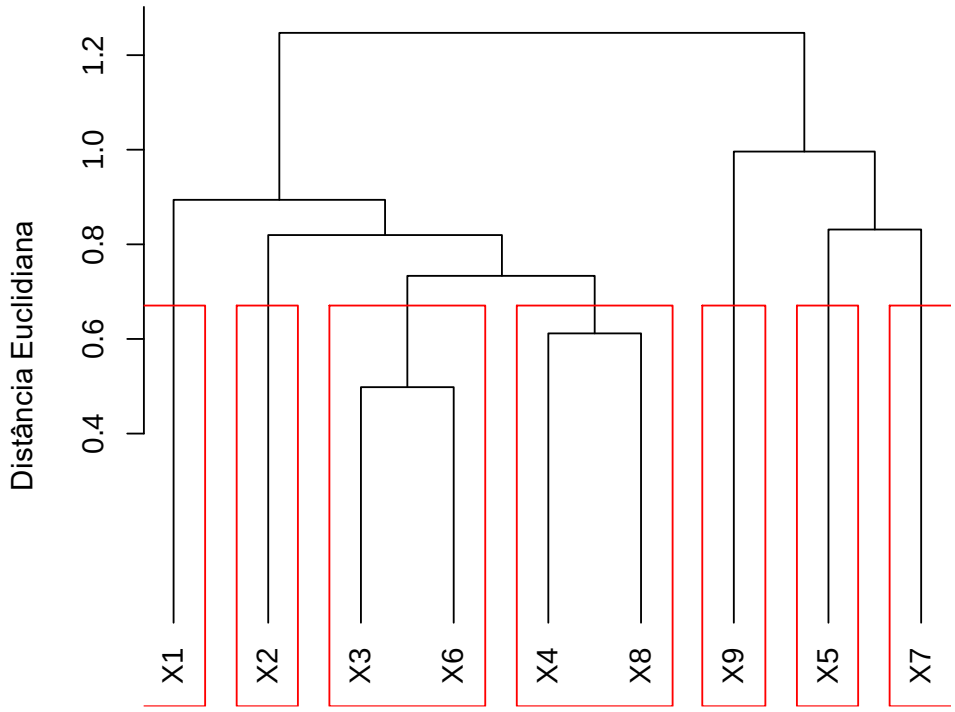

Figura 20: Agrupamento do dendograma (Método Ward) com 7 clusters.

Já nessa divisão com mais grupos, X3 e X6 ficaram agrupadas juntas e X4 e X8 também. As outras variáveis ficaram separadas em grupos unitários. Essa divisão dos grupos se assemelha muito com os clusters gerados no Método de Ligação Média gerados anteriormente (Figura 14).

Alternativa E) Compare os resultados obtidos.

Realizando algumas comparações entre as análises multivariadas de Componente Principal, Fatorial e Agrupamento empregadas nesse trabalho, notam-se algumas semelhanças e diferenças:

i) Entre a análise de Componentes Principais e Análise Fatorial, as variáveis de maior influência em cada respectivo Cp (componente principal) e Cf (carga fatorial) são as mesmas:

CP1 (Cf1): X3, X6 e X8
CP2 (Cf2): X5, X7 e X9
CP3 (Cf3): X2, X4 e X9
CP4 (Cf4): X1, X7 e X9

O que já era esperado, uma vez que como método de estimação das cargas fatoriais foi empregado o próprio método de Componentes Principais.

ii) Já aplicando a rotação Varimax, dentro da própria análise fatorial, ocorrem algumas mudanças (diferenças) nas variáveis de maior influência nas cargas fatoriais.

Cf1: **X3**, X6 e **X8** –> X1, X2, **X3** e **X8**
Cf2: **X5**, **X7** e X8 –> **X5**, **X7** e X9
Cf3: X2, **X4** e X9 –> X3 e **X4**
Cf4: X1, X7 e X9 –> X4 e X6

iii) Na comparação entre os métodos de agrupamento aplicado para os indivíduos, houve grande semelhança nos clusters formados pelos métodos de Ligação Média e Centróide, sendo que o coeficiente de correlação cofenético do primeiro foi mais alto e considerado adequado à análise. Em comparação desses métodos hierárquicos com o método não hierárquico do K-médias, também é possível visualizar grande coincidência na formação dos clusters. Apesar de no K-médias os indivíduos terem sido separados apenas em 2 grupos, eles equivalem à junção dos grupos dos métodos anteriores. Isto é, com 2 grupos nos métodos hierárquicos ocorreria a mesma separação de países ocorrida no K-means. Com três métodos apontando para um mesmo caminho de agrupamento, pode-se dizer que essas foram divisões adequadas dos países nos grupos.

iv) As variáveis que mais influenciaram tanto na formação dos clusters do agrupamento de variáveis quanto com o maior valor de correlação na formação do primeiro componente principal (este com maior porcentagem de representação da variância total) foram X3 (ovos), X6 (cereais), e X8 (amêndoas

e sementes oleaginosas), podendo-se afirmar, de forma geral, que essas foram as variáveis de maior relevância para essa análise multivariada desse banco de dados.

v) O agrupamento de variáveis não gerou, aparentemente, uma divisão de grupos que estabeleça alguma relação com um agrupamento natural das variáveis de forma exploratória, não levando a nenhuma separação das mesmas por classe (por exemplo: origem vegetal / origem animal). Isso pode ser explicado, possivelmente, pelo fato de o banco de dados ser relativo à ingestão de gramas de proteína sem levar em consideração as proporções de proteína em cada fonte, mostrando então que não há uma correlação entre a origem (característica) da proteína e o consumo em gramas nos países europeus estudados.

REFERÊNCIAS BIBLIOGRÁFICAS

CALLEGARI-JACQUES, S.M. **Bioestatística – Princípios e Aplicações**. Porto Alegre: Artmed, 2003.

COHEN, J. (1988). **Statistical power analysis for the behavioral sciences** (2nd ed.). New Jersey: Lawrence Erlbaum.

KAISER, H. F. The application of electronic computers to factor analysis. Educational and Psychological Measurement, v. 20, p. 141 – 151, 1960. *apud.* SILVA, M.C.; SILVA J.D.G.; BORGES, E.F. ANÁLISES DE COMPONENTES PRINCIPAIS PARA ELABORAR ÍNDICES DE DESEMPENHO NO SETOR PÚBLICO. **Rev. Bras. Biom**., São Paulo, v.33, n.3, p.291-309, 2015.

LINDEN, R. Técnicas de agrupamento. **Revista de Sistemas de Informação da FSMA**, n.4, p.18-36, 2009.

ANEXO A – Script software R

```
############################################################
##### QUESTÃO 2 #####

tabela <- read.table("dados_prova1.txt", header = T)
tabela

dados <- tabela[,2:10]
dados

row.names(dados) = tabela[,1]
dados

attach(dados)
names(dados)

is.data.frame(dados)

dim(dados)

colMeans(dados)

summary(dados)

## BOX PLOTs ##
par(mfrow = c(1,3))
boxplot(X1, main = "X1", ylab = "Frequência")
boxplot(X2, main = "X2", ylab = "Frequência")
boxplot(X3, main = "X3", ylab = "Frequência")
par(mfrow = c(1,3))
boxplot(X4, main = "X4", ylab = "Frequência")
boxplot(X5, main = "X5", ylab = "Frequência")
boxplot(X6, main = "X6", ylab = "Frequência")
par(mfrow = c(1,3))
boxplot(X7, main = "X7", ylab = "Frequência")
boxplot(X8, main = "X8", ylab = "Frequência")
boxplot(X9, main = "X9", ylab = "Frequência")


############################################################
### Alternativa A ###

## COMPONENTES PRINCIAPAIS ##

S <- cov(dados,dados)
S

R <- cor(dados,dados)
R

## teste de esfericidade de Bartlett ##
n <- nrow(dados)
require(psych)
cortest.bartlett(R,n)  ## teste de esfericidade de bartlett

KMO(R)  ## índice KMO pelo pacote psych ##
```

```
## COMANDO ESPECÍFICO DE COMPONENTES PRINCIPAIS ##


## USANDO A MATRIZ S padronizada = MATRIZ R ##

eigen(R)

cpR <- prcomp(dados, scale = T)   ## cria os componentes principais usando R ##
cpR

par(mfrow = c(1,2))
screeplot(cpR)
screeplot(cpR, type = "lines")     ## gráfico de cotovelo ##

require(factoextra)
fviz_eig(cpR)

names(cpR)
summary(cpR)          ## desvio padrão, proporção e proporção acumulada ##

cpR$sdev              ## desvio padrão dos CP's: raíz quadrada autovalores ##
cpR$rotation          ## coeficientes cada componente principal: autovetores ##
cpR$center            ## coordenada central: média amostral ##

cpR$rotation[,1]      ## coeficientes do 1º CP ##
score1 <- t(cpR$rotation[,1]) %*% t(dados)
t(score1)             ## score para cada indivíduo no CP1 ##

cpR$rotation[,2]      ## coeficientes do 2º CP ##
score2 <- t(cpR$rotation[,2]) %*% t(dados)
t(score2)             ## score para cada indivíduo no CP2 ##

cpR$rotation[,3]      ## coeficientes do 3º CP ##
score3 <- t(cpR$rotation[,3]) %*% t(dados)
t(score3)             ## score para cada indivíduo no CP3 ##

cpR$rotation[,4]      ## coeficientes do 4º CP ##
score4 <- t(cpR$rotation[,4]) %*% t(dados)
t(score4)             ## score para cada indivíduo no CP4 ##


cpR$x                 ## scores com variáveis centradas ##


biplot(cpR, choices=c(1,2))          ## gráfico biplot CP1,CP2##

biplot(cpR, choices=c(3,4))          ## gráfico biplot CP3,CP4##



#########################################################
### Alternativa B ###

## FATORIAL ##

## Matriz de Correlação ##
R <- cor(dados)
R
```

```
## Autovalores e Autovetores ##
autov <- eigen(R)
autov

## Da análise anterior (ACP), adotam-se os 4 primeiros CPs ##
## Proporção acumulada ##
prop_acu <- sum(autov$values[1:4]) / sum(autov$values)
prop_acu*100 #(%)


## Cargas fatoriais pelo método de CP ##

carga1 <- sqrt(autov$values[1]) * autov$vectors[,1]
carga2 <- sqrt(autov$values[2]) * autov$vectors[,2]
carga3 <- sqrt(autov$values[3]) * autov$vectors[,3]
carga4 <- sqrt(autov$values[4]) * autov$vectors[,4]

carga1
carga2
carga3
carga4

L <- cbind(carga1,carga2,carga3,carga4)
L

##Comunalidade: qualidade da análise##
com <- carga1^2 + carga2^2 + carga3^2 + carga4^2
com

## Matriz psi ##
psi <- R - L %*% t(L)
diag(psi)

# grafico de dispersao entre as cargas
par(mfrow = c(1,2))
plot(carga1,carga2, pch = 20)
text(carga1,carga2, adj=1.5)

plot(carga3,carga4, pch = 20)
text(carga3,carga4, adj=1.5)


## Transformação Varimax ##
V <- varimax(L, normalize = F)
V

par(mfrow = c(1,2))
plot(V$loading[,c(1,2)], pch = 20)
text(V$loadings[,c(1,2)], adj=1.5)

plot(V$loading[,c(3,4)], pch = 20)
text(V$loadings[,c(3,4)], adj=1.5)


#######################################################
### Alternativa C ###

## AGRUPAMENTO - INDIVIDUOS ##
```

```
disteuclid <- dist(dados)
disteuclid
print(disteuclid, digits = 3)

## Cálculo dos Coeficientes Cor. Cofenéticos para escolha dos métodos ##

## LIGAÇÃO SIMPLES (Single Linkage) - VIZINHO MAIS PRÓXIMO ##
cluster_single <- hclust(disteuclid, "single")
cluster_single
help(hclust)
d1 <- cophenetic(cluster_single)
cor1 <- cor(disteuclid,d1)            ## coeficiente de correlação cofenética ##

## LIGAÇÃO COMPLETA (Complete Linkage) - VIZINHO MAIS DISTANTE ##
cluster_complete <- hclust(disteuclid, "complete")
cluster_complete
d2 <- cophenetic(cluster_complete)
cor2 <- cor(disteuclid,d2)            ## coeficiente de correlação cofenética ##

## LIGAÇÃO MÉDIA (Average Linkage) ##
cluster_average <- hclust(disteuclid, "average")
cluster_average
d3 <- cophenetic(cluster_average)
cor3 <- cor(disteuclid,d3)            ## coeficiente de correlação cofenética ##

## CENTRÓIDE ##
cluster_centroid <- hclust(disteuclid, "centroid")
cluster_centroid
d4 <- cophenetic(cluster_centroid)
cor4 <- cor(disteuclid,d4)            ## coeficiente de correlação cofenética ##

## WARD ##
cluster_ward <- hclust(disteuclid, "ward")
cluster_ward
d5 <- cophenetic(cluster_ward)
cor5 <- cor(disteuclid,d5)            ## coeficiente de correlação cofenética ##

ccc <- c(cor1, cor2, cor3, cor4, cor5)
cbind(ccc)                  ## Métodos escolhidos: Ligação Média e Centróide ##


        ## DENDOGRAMAS ##

## LIGAÇÃO MÉDIA ##

cluster_average$height      ## valores de junção no dendograma (maior salto)##
## definir numero de clusters ##
## h>17,61
## k=4
plot(cluster_average, xlab = "Parcelas", ylab = "Distância Euclidiana", main = "",
hang = -1)
rect.hclust(cluster_average, k=4)

## classifica os elementos em cada grupo ##
c <- cutree(cluster_average, k = 4)
## diagrama de dispersão a cada 2 variáveis, identificando os grupos ##
plot(dados, col = c)
```

```
## CENTRÓIDE ##

cluster_centroid$height       ## valores de junção no dendograma (maior salto)##
## definir numero de clusters ##
## h>10,86
## k= 3
plot(cluster_centroid, xlab = "Parcelas", ylab = "Distância Euclidiana", main = "",
hang = -1)
rect.hclust(cluster_centroid, k=3)


## MÉTODO K-MÉDIAS(K-Means) - Não Hierárquico ##

## Número de grupos ##
help(fviz_nbclust)
library("factoextra")
library("NbClust")
## Hubert and D index graphical method ##
nb <- NbClust(dados, distance = "euclidean", min.nc = 2,
max.nc = 8, method = "kmeans")
fviz_nbclust(nb)

# Elbow method for kmeans
fviz_nbclust(dados, kmeans, method = "wss") +
geom_vline(xintercept = 2, linetype = 2) +
geom_vline(xintercept = 3, linetype = 2)

## k=2 ##

kmedias <- kmeans(dados,2)        ## método kmeans ##
kmedias
kmedias$size
kmedias$cluster
cbind(kmedias$cluster)

require(cluster)
clusplot(dados, kmedias$cluster, color=T, shade=T, labels=2,
lines=0,cex.txt=0.8, main="")
help(clusplot)


#########################################################
### Alternativa D ###

## AGRUPAMENTO - VARIÁVEIS ##

## matriz de correlação Pearson ##
R <- cor(dados)
R

## matriz de distancia Rencher ##
r <- as.dist((1-(R^2)))  ## SUGERIDA POR RENCHER (2002) ##
r

## Cálculo dos Coeficientes Cor. Cofenéticos para escolha dos métodos ##

## LIGAÇÃO SIMPLES (Single Linkage) - VIZINHO MAIS PRÓXIMO ##
cluster_single <- hclust(r, "single")
cluster_single
help(hclust)
```

```
d1 <- cophenetic(cluster_single)
cor1 <- cor(r,d1)            ## coeficiente de correlação cofenética ##

## LIGAÇÃO COMPLETA (Complete Linkage) - VIZINHO MAIS DISTANTE ##
cluster_complete <- hclust(r, "complete")
cluster_complete
d2 <- cophenetic(cluster_complete)
cor2 <- cor(r,d2)            ## coeficiente de correlação cofenética ##

## LIGAÇÃO MÉDIA (Average Linkage) ##
cluster_average <- hclust(r, "average")
cluster_average
d3 <- cophenetic(cluster_average)
cor3 <- cor(r,d3)            ## coeficiente de correlação cofenética ##

## CENTRÓIDE ##
cluster_centroid <- hclust(r, "centroid")
cluster_centroid
d4 <- cophenetic(cluster_centroid)
cor4 <- cor(r,d4)            ## coeficiente de correlação cofenética ##

## WARD ##
cluster_ward <- hclust(r, "ward")
cluster_ward
d5 <- cophenetic(cluster_ward)
cor5 <- cor(r,d5)            ## coeficiente de correlação cofenética ##

ccc <- c(cor1, cor2, cor3, cor4, cor5)
cbind(ccc)                   ## Métodos escolhidos: Ligação Média e Ward ##


          ## DENDOGRAMAS ##

## LIGAÇÃO MÉDIA ##

cluster_average$height       ## valores de junção no dendograma (maior salto)##
## definir numero de clusters ##
## h>0,64
## k=6
plot(cluster_average, xlab = "Parcelas", ylab = "Distância Euclidiana", main = "",
hang = -1)
rect.hclust(cluster_average, k=6)

## classifica os elementos em cada grupo ##
c <- cutree(cluster_average, k = 6)
## diagrama de dispersão a cada 2 variáveis, identificando os grupos ##
plot(dados, col = c)

## LIGAÇÃO SIMPLES ##

cluster_single$height       ## valores de junção no dendograma (maior salto)##
## definir numero de clusters ##
## h>0,72
## k= 2
plot(cluster_single, xlab = "Parcelas", ylab = "Distância Euclidiana", main = "",
hang = -1)
rect.hclust(cluster_single, k=2)


## WARD ##
```

```
cluster_ward$height       ## valores de junção no dendograma (maior salto)##
## definir numero de clusters ##
## h>0,99
## k= 2
plot(cluster_ward, xlab = "Parcelas", ylab = "Distância Euclidiana", main = "",
hang = -1)
rect.hclust(cluster_ward, k=2)
plot(cluster_ward, xlab = "Parcelas", ylab = "Distância Euclidiana", main = "",
hang = -1)
rect.hclust(cluster_ward, k=7)


################## FIM #################
```