Introdução a Métodos Estatísticos para a Bioinformática

Profa. Júlia Maria Pavan Soler

pavan@ime.usp.br

IBI 5086 – Bioinformática - IME/USP

2º Sem/2020

Programa

- ➢ Álgebra linear básica: cálculo matricial, determinantes, sistemas lineares, produto interno, norma, ortogonalidade, autovalores e autovetores
- ✓ Estrutura de Dados: variáveis (resposta, explicativa), unidades amostrais e experimentais

- ✓ 1.1. Comparação de 2 ou mais grupos: Testes Clássicos (teste t, Wilcoxon, ANOVA), Testes de Aleatorização, Comparações Múltiplas, Efeitos Genéticos
- ✓ 1.2. Análise de Tabelas de Contingência: Testes Qui-Quadrado, Regressão Logística.

2. **Análise Multivariada de Dados**: Componentes Principais, Análise Discriminante e Classificação, Correlação Canônica, modelos MANOVA

3. Simulação de Monte Carlo, Intervalos de Confiança Bootstrap

Análise Multivariada

$$Y_{n\times p} = \left(Y_{ij}\right) \in \Re^{n\times p}$$

matriz retangular

Já vimos ☺

Matriz de Dados: **Estatísticas descritivas multivariadas**

Centróide: $\bar{Y}_{p\times 1} = \left(\bar{Y}_1, \bar{Y}_1, ..., \bar{Y}_p\right)' \in \Re^p$

Matrix de Covariância: $S_{p\times p} \in \Re^{pxp}$ *matriz quadrada*

Matriz de Correlação: $R_{p\times p} \in \Re^{pxp}$ *matriz quadrada*

Matriz de Distâncias: $D_{n\times n} \in \Re^{nxn}$ *matriz quadrada*

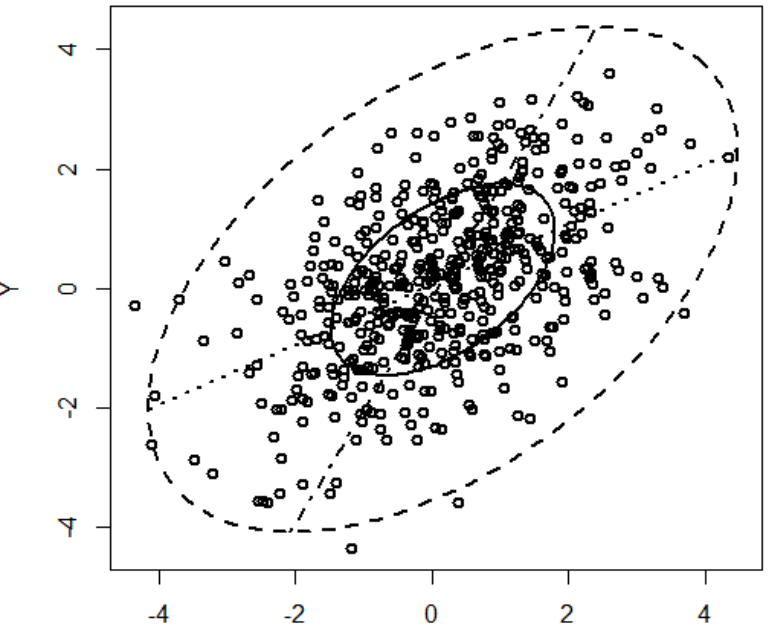
Boxplot Bivariado $\Rightarrow$ Elipses de Concentração $\Rightarrow$ Diagnóstico de Outliers em $\Re^p$

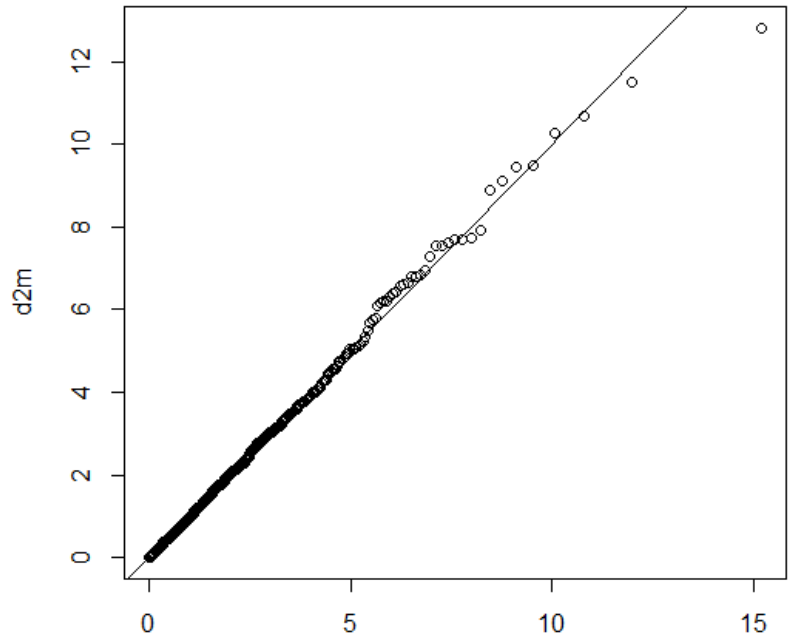
Técnicas de Redução de Dimensionalidade: $\Re^p \rightarrow \Re^m; \quad m \leq p$

Análises Clássicas: n > p, observações *iid* , variáveis quantitativas

Caso n << p (Big-p): soluções esparsas (regularizadas, penalizadas)

Caso n>>p (Big-n): soluções por reamostragem





Dados Bivariados (p=2)
Elipse de Concentração dos Dados
(Distância de Mahalanobis)

```
library(MASS)
mu<-c(0,0)
sigma<-matrix(c(2,1,1,2),ncol=2)
n<-500
y<-mvrnorm(n,mu,sigma)
mi<-colMeans(y)
s<-cov(y)
par(mfrow=c(1,2))
bivbox(y, method ="O")
# Copy Everitt's bivbox function
d2m<-mahalanobis(y,mi,s)
quantis <- qchisq(ppoints(length(y)),df=2)
qqplot(quantis, d2m)
abline(0,1)
```

Qual dos dois eixos da elipse de concentração dos dados descreve a direção de maior variabilidade dos dados?

Redução de Dimensionalidade em $\Re^p$

	Variáveis					
Unidades Amostrais	1	2	…	j	…	p
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
…	…	…	…	…	…	…
n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

$$Y_{n\times p};\quad n > p \qquad \Re^p \rightarrow \Re^m, m < p$$

Redução de Dimensionalidade $\Rightarrow$ obter m vetores (m<p) que são combinações lineares das p variáveis originais e atendem a critérios de otimalidade

$Y_{n\times p}$ p vetores das respostas para n indivíduos $\Rightarrow$ Vetores de **Escores** para os n indivíduos (CP)

Vetores de **Cargas** (pesos) às p variáveis

Ténicas Multivariadas de Redução de Dimensionalidade

Como obter vetores reducionistas de dados?

Depende:

- Estrutura dos Dados
- Objetivo da análise

- Análise de Componentes Principais: $Y_{n\times p} \Rightarrow \Re^{p\times p}$
- Escalonamento Multidimensional: $Y_{n\times p} \Rightarrow D^{n\times n}$

Análises (equivalentes) em espaços duais (linhas e colunas)

- Análise de Correspondência: $Y_{n\times p} \Rightarrow [0,1]^{I\times J}$ Em tabelas de contingência (proporções)
- Análise Discriminante $Y_{n\times(p+1)} \Rightarrow \Re^{p\times p}$ $\Rightarrow$ MANOVA

Análise supervisionada

- Análise de Agrupamento: $Y_{n\times p} \Rightarrow D^{n\times n}$
- Análise de Correlação Canônica: $Y_{n\times(p+q)} \Rightarrow \Re^{p\times q}(\Re^{p\times p}, \Re^{q\times q})$

N-Integração de bancos de dados

Análise de Componentes Principais

Análise Clássica

n > p
Observações iid
(respostas quantitativas)

Análise de Componentes Principais

(Pearson, 1901)

	Variáveis					
Unidades Amostrais	1	2	…	j	…	p
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
…	…	…	…	…	…	…
n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

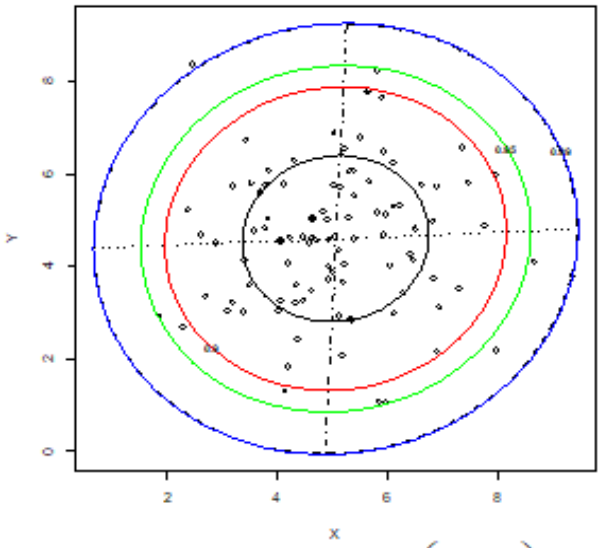
$$Y_{n\times p};\quad n > p \Rightarrow Y_{i_{p\times 1}} \overset{iid}{\sim} (\mu;\Sigma)$$

Premissa: Dados de Uma única População
Observações *iid*
Matriz de covariâncias "válida" ($\Sigma \in \Re^{pxp}$)

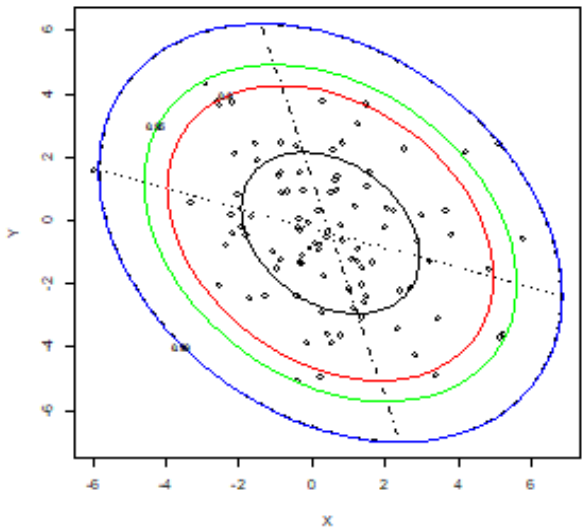
- A variável Y_j pode ser eliminada da análise?
- Como as variáveis podem ser ordenadas segundo sua "importância" na análise?

A análise de CP considera a estrutura de Σ
(análise no espaço das variáveis)

$$\mu' = (5,5) \quad \Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

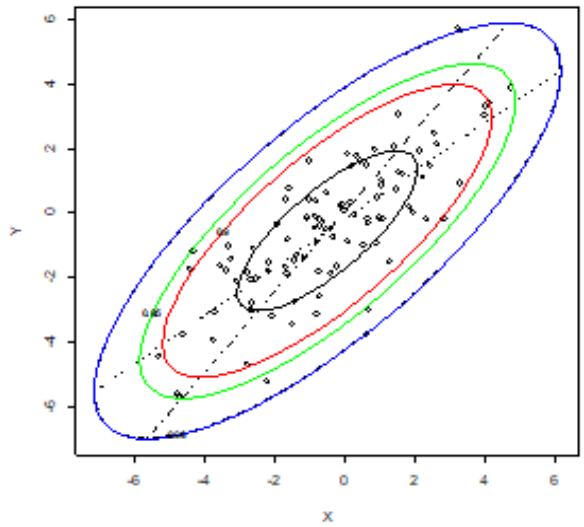


$$\mu' = (0,0) \quad \Sigma = \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix}$$

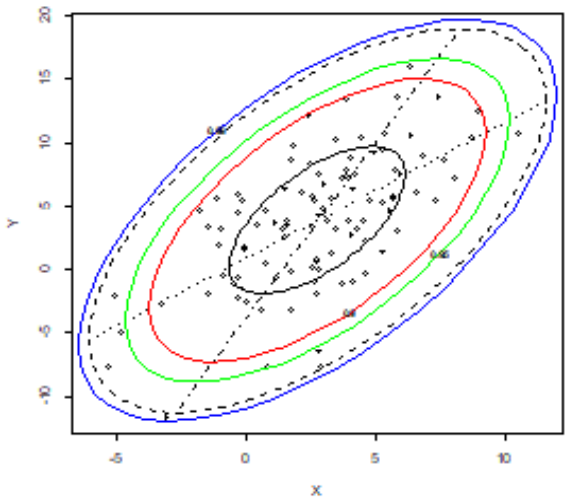


BoxPlot Bivariado
Elipses de concentração: representação da matriz de covariância Σ

$$\mu' = (0,0) \quad \Sigma = \begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix}$$



$$\mu' = (3,4) \quad \Sigma = \begin{pmatrix} 9 & 10 \\ 10 & 25 \end{pmatrix}$$



Identifique as direções de maior variabilidade dos dados?

$\Sigma \Leftrightarrow R$

Análise de Componentes Principais
Estruturas de Σ e R

Como proceder com a redução de dimensionalidade nos seguintes casos?

Σ_1: Estrutura *apropriada* para a redução: ordenar as variáveis de acordo com a variância e calcular a contribuição para a variância total.

$$\Sigma_1 = \begin{pmatrix} \sigma_{11} & 0 & \dots & 0 \\ 0 & \sigma_{22} & \dots & 0 \\ \dots & 0 & \dots & \dots \\ 0 & 0 & 0 & \sigma_{pp} \end{pmatrix}; \qquad R_1 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

R_1: Não há como reduzir a dimensionalidade de espaços formados por variáveis não correlacionadas e homocedásticas

$$\Sigma_2 = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \dots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \dots & \rho\sigma^2 \\ \dots & 0 & \dots & \dots \\ \rho\sigma^2 & \rho\sigma^2 & \dots & \sigma^2 \end{pmatrix}; \qquad R_2 = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ \rho & \rho & \dots & 1 \end{pmatrix} = (1-\rho) I_p + \rho 1_p 1_p'$$

Correlação uniforme.
Se ρ for alto, um único CP deve explicar bem a (co)variância dos dados e ele é uma média ponderada que atribui pesos iguais à todas as variáveis.

$$\Sigma_3 = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ & \sigma_{22} & \dots & \sigma_{2p} \\ \sim & & \dots & \dots \\ & & \dots & \sigma_{pp} \end{pmatrix}; \qquad R_3 = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ & 1 & \dots & \rho_{2p} \\ \sim & & \dots & \dots \\ & & \dots & 1 \end{pmatrix}$$

Análise de Componentes Principais
Dados Nutricionais

	energia	proteina	gordura	calcio	ferro
[1,]	340	20	28	9	2.6
[2,]	245	21	17	9	2.7
[3,]	420	15	39	7	2.0
[4,]	375	19	32	9	2.5
[5,]	180	22	10	17	3.7
[6,]	115	20	3	8	1.4
[7,]	170	25	7	12	1.5
[8,]	160	26	5	14	5.9
[9,]	265	20	20	9	2.6
[10,]	300	18	25	9	2.3
[11,]	340	20	28	9	2.5
[12,]	340	19	29	9	2.5
[13,]	355	19	30	9	2.4
[14,]	205	18	14	7	2.5
[15,]	185	23	9	9	2.7
[16,]	135	22	4	25	0.6
[17,]	70	11	1	82	6.0
[18,]	45	7	1	74	5.4
[19,]	90	14	2	38	0.8
[20,]	135	16	5	15	0.5
[21,]	200	19	13	5	1.0
[22,]	155	16	9	157	1.8
[23,]	195	16	11	14	1.3
[24,]	120	17	5	159	0.7
[25,]	180	22	9	367	2.5
[26,]	170	25	7	7	1.2
[27,]	110	23	1	98	2.6

Como construir um Escore Nutricional que characterize a variabilidade dos produtos?
Como representar graficamente os produtos em $\Re^2$?

Centróide:

energia	proteína	gordura	cálcio	ferro
207.41	19.00	13.48	43.96	2.38

Matriz de covariância (S)

10243.02	74.81	1124.57	-2530.29	-14.75
74.81	18.08	1.19	-28.23	-1.08
1124.57	1.19	126.72	-270.6 7	-1.00
-2530.29	-28.23	-270.67	6089.34	5.05
-14.75	-1.08	-1.00	5.05	2.13

Matriz de correlação (R)

1.00	0.17	0.99	-0.32	-0.10
0.17	1.00	0.02	-0.09	-0.17
0.99	0.02	1.00	-0.31	-0.06
-0.32	-0.09	-0.31	1.00	0.04
-0.10	-0.17	-0.06	0.04	1.00

R sugere um padrão não estruturado de correlação entre as variáveis

Análise de Componentes Principais
Dados dos Cães

$Y_{7\times6}$

Cães pré-históricos da Tailândia (Manly, 2005).

Grupo	X1	X2	X3	X4	X5	X6
G1	9.7	21.0	19.4	7.7	32.0	36.5
G2	8.1	16.7	18.3	7	30.3	32.9
G3	13.5	27.3	26.8	10.6	41.9	48.1
G4	11.5	24.3	24.5	9.3	40.0	44.6
G5	10.7	23.5	21.4	8.5	28.8	37.6
G6	9.6	22.6	21.1	8.3	34.4	43.1
Cão Pré-h	10.3	22.1	19.1	8.1	32.2	35.0

Quais variáveis mais contribuem para a variabilidade entre os indíduos (cães)?
Como representar graficamente os cães em $\Re^2$?

Dados dos Cães Pré-históricos

Centróide

X1	X2	X3	X4	X5	X6
10.48571	22.50000	21.51429	8.50000	34.22857	39.68571

Matriz de Covariância

	X1	X2	X3	X4	X5	X6
X1	2.881429	5.251667	4.846905	1.933333	6.527143	7.739762
X2	5.251667	10.556667	8.895000	3.593333	11.456667	15.583333
X3	4.846905	8.895000	9.611429	3.508333	13.427857	16.305238
X4	1.933333	3.593333	3.508333	1.356667	4.863333	5.920000
X5	6.527143	11.456667	13.427857	4.863333	24.362381	24.680476
X6	7.739762	15.583333	16.305238	5.920000	24.680476	31.518095

Sugere um padrão de correlação uniforme

Matriz de Correlação

	X1	X2	X3	X4	X5	X6
X1	1.0000000	0.9522036	0.9210148	0.9778365	0.7790392	0.8121639
X2	0.9522036	1.0000000	0.8830567	0.9495056	0.7143894	0.8543129
X3	0.9210148	0.8830567	1.0000000	0.9715615	0.8775116	0.9368136
X4	0.9778365	0.9495056	0.9715615	1.0000000	0.8459362	0.9053263
X5	0.7790392	0.7143894	0.8775116	0.8459362	1.0000000	0.8906636
X6	0.8121639	0.8543129	0.9368136	0.9053263	0.8906636	1.0000000

Análise de Componentes Principais

$Y_{i\,p\times 1}$: vetor de respostas do indivíduo i (i=1,...,n); $Cov(Y_i) = \Sigma_{p\times p}$

$$Y_i \in \Re^p \quad \rightarrow \quad Z_i = V'_{m\times p}\, Y_{i_{p\times 1}} \in \Re^m$$

$$Cov(Y_i) = \Sigma_{p\times p} \qquad Cov(Z_i) = \Lambda_{m\times m} = Diag(\lambda_j)$$

$$tr\,\Sigma = \sum_{j=1}^{p} \lambda_j \;\cong\; \sum_{j=1}^{m} \lambda_j = tr\Lambda_{m\times m}$$

Redução de dimensionalidade: transformar Y em Z, reduzindo de p para m variáveis (m<p), mas preservando ao máximo a variância total

$$Y_{i\,p\times 1} \Rightarrow Z_{i\,m\times 1} = V'_{m\times p}\, Y_i$$

$$Z_{i1} = \nu_{11}Y_{i1} + \nu_{21}Y_{i2} + \ldots + \nu_{p1}Y_{ip} = \sum_{i=1}^{p} \nu_{j1}Y_{ij}$$

$$Z_{i2} = \nu_{12}Y_{i1} + \nu_{22}Y_{i2} + \ldots + \nu_{p2}Y_{ip} = \sum_{i=1}^{p} \nu_{j2}Y_{ij}$$

...

$$Z_{im} = \nu_{1m}Y_{i1} + \nu_{2m}Y_{i2} + \ldots + \nu_{pm}Y_{ip} = \sum_{i=1}^{p} \nu_{jm}Y_{ij}$$

$$Y_{n\times p} \Rightarrow Z_{n\times m} = Y_{n\times p}V_{p\times m}$$

$$V_{p\times m} = \left(\nu_{jk}\right)$$

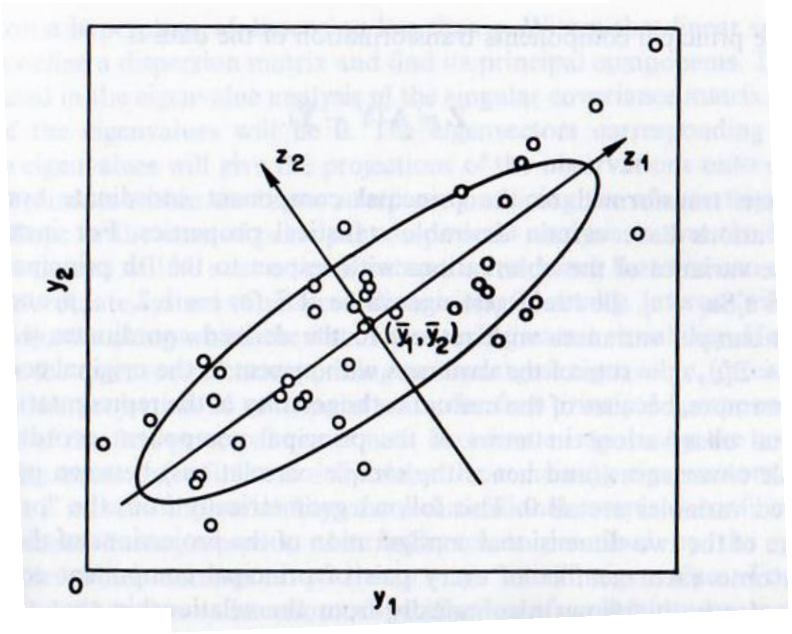
Como obter a matriz ***V*** de tais cargas (pesos) ?

Análise de Componentes Principais

Técnica de Redução Linear de Dimensionalidade de Variáveis

$(y-\bar{y})'\Sigma^{-1}(y-\bar{y})=c^2$ define

uma família de elipses de concentração aos dados



Transformação que preserva a variância total (Rotação ortogonal dos Eixos)

$$Y \in \Re^2 \quad \Rightarrow \quad Z = YV \in \Re^2$$
$$(y_1, y_2) \quad \Rightarrow \quad (z_1, z_2)$$

Z_1 : primeiro componente principal

Z_2 : segundo componente principal

$$Z_1 = V_1'Y\ ; \quad Var(Z_1) = V_1'\Sigma V_1 = \lambda_1$$
$$Z_2 = V_2'Y; \quad Var(Z_2) = V_2'\Sigma V_2 = \lambda_2$$
$$\lambda_1 \geq \lambda_2$$
$$Cov(Z_1, Z_2) = V_1'\Sigma V_2 = 0$$

Como obter V e λ?

Decomposição espectral de Σ: $\Sigma_{p\times p} = V\Lambda V'$

V: matriz de autovetores

$\Lambda=(\lambda_k)$: são os autovalores

Análise de Componentes Principais

Obtenção dos Componentes Principais dos Dados Nutricionais

Decomposição Espectral da Matriz de Covariância S:

Autovalores de S

```
11552.53  4903.92    20.43     2.07     0.35
```

Autovetores de S

	V1	V2	V3	V4	V5
[1,]	0.90	0.42	-0.03	-0.01	0.10
[2,]	0.01	0.00	-0.92	0.10	-0.37
[3,]	0.10	0.05	0.37	0.09	-0.92
[4,]	-0.42	0.91	0.00	0.00	0.00
[5,]	0.00	0.00	0.06	0.99	0.12

$$tr\ S = 16479.3 = \sum_{j=1}^{p} \lambda_j = tr\Lambda$$

Obtenha Z1? $PC1 = Z_1 = YV_1$

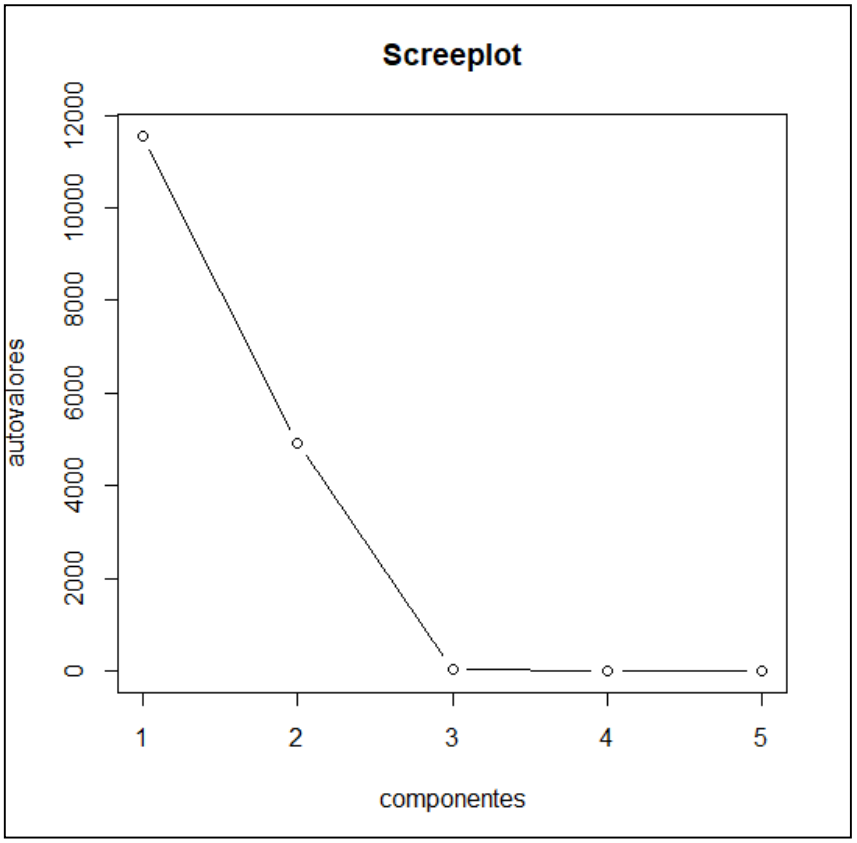
Z2? $PC2 = Z_2 = YV_2$

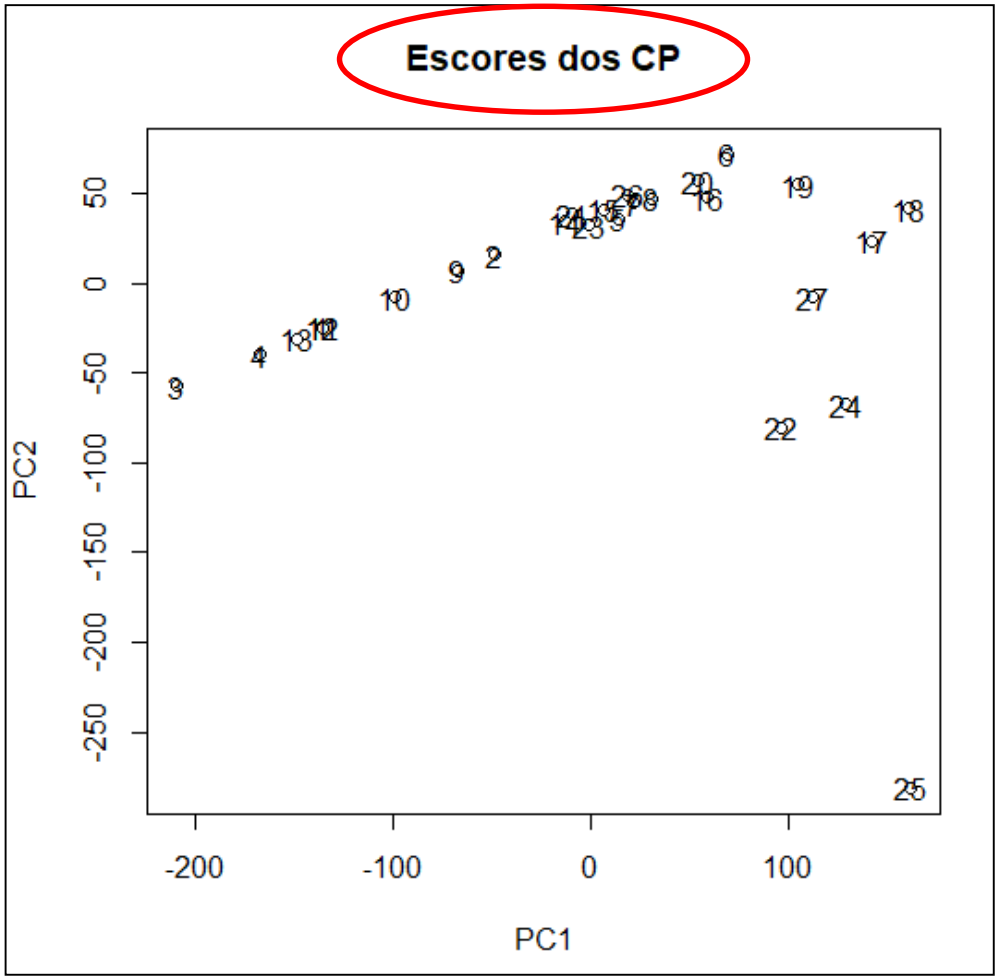
Importância dos Componentes Principais:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	107.483	70.0280	4.51941	1.43767	0.59303
Proportion of Variance	0.701	0.2976	0.00124	0.00013	0.00002
Cumulative Proportion	0.701	0.9986	0.99985	0.99998	1.00000

Análise de Componentes Principais

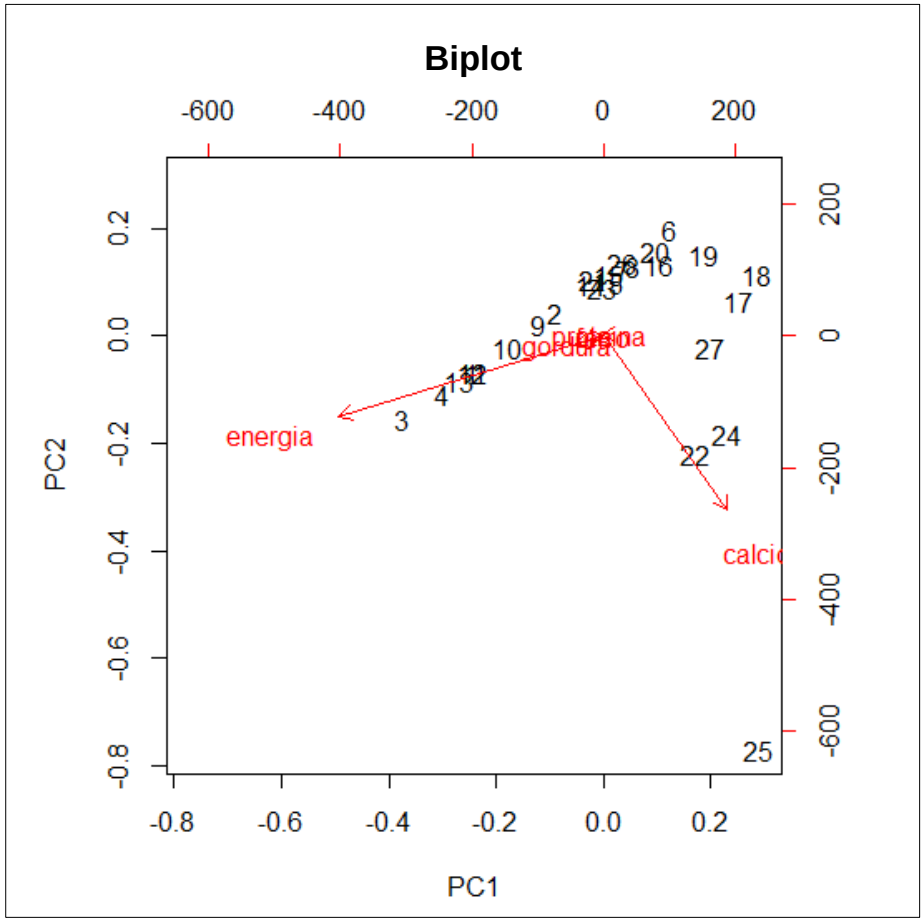
Dados Nutricionais (n=27; p=5)





Análise de Componentes Principais

Dados Nutricionais (n=27; p=5)



Biplot: Representação simultânea dos escores dos CP e das cargas (pesos) das variáveis

As variáveis Energia e Cálcio dominam a análise: atribuem os maiores pesos na combinação linear das variáveis

A observação 25 é atípica em relação às demais.

Dados Nutricionais e os Escores dos Dois Primeiros Componentes Principais

	energia	proteina	gordura	calcio	ferro	PC1	PC2
[1,]	340	20	28	9	2.6	-135.68	-24.63
[2,]	245	21	17	9	2.7	-49.00	15.75
[3,]	420	15	39	7	2.0	-209.66	-56.90
[4,]	375	19	32	9	2.5	-167.60	-39.51
[5,]	180	22	10	17	3.7	13.64	36.10
[6,]	115	20	3	8	1.4	69.11	71.87
[7,]	170	25	7	12	1.5	20.81	44.97
[8,]	160	26	5	14	5.9	30.86	47.45
[9,]	265	20	20	9	2.6	-67.31	7.22
[10,]	300	18	25	9	2.3	-99.32	-7.70
[11,]	340	20	28	9	2.5	-135.68	-24.63
[12,]	340	19	29	9	2.5	-135.77	-24.68
[13,]	355	19	30	9	2.4	-149.38	-31.02
[14,]	205	18	14	7	2.5	-13.48	34.49
[15,]	185	23	9	9	2.7	5.84	41.30
[16,]	135	22	4	25	0.6	58.15	48.02
[17,]	70	11	1	82	6.0	141.17	23.78
[18,]	45	7	1	74	5.4	160.34	41.52
[19,]	90	14	2	38	0.8	104.44	55.22
[20,]	135	16	5	15	0.5	53.87	57.04
[21,]	200	19	13	5	1.0	-9.73	38.45
[22,]	155	16	9	157	1.8	95.41	-80.26
[23,]	195	16	11	14	1.3	-1.21	32.49
[24,]	120	17	5	159	0.7	128.18	-67.20
[25,]	180	22	9	367	2.5	161.52	-281.12
[26,]	170	25	7	7	1.2	18.70	49.50
[27,]	110	23	1	98	2.6	111.79	-7.52

CP obtidos de S

$\Re^5 \Rightarrow \Re^2$

Análise de Componentes Principais

Dados Nutricionais (n=27; p=5) : Redução para os 2 primeiros Componentes Principais (CP1 e CP2)

Matriz de correlação dos CP com as variáveis originais

	PC1	PC2
energia	-0.95693047	-0.29032625
proteina	-0.17411811	-0.01973317
gordura	-0.94229427	-0.29494848
calcio	0.58159624	-0.81346912
ferro	0.09893007	0.01624411

Proporção da variância das variáveis explicada pelos CP

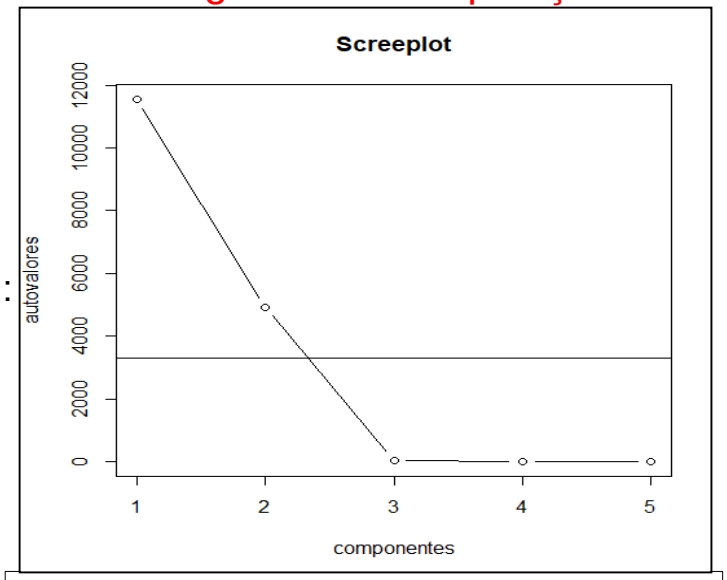
	PC1	PC2	variância
energia	0.915715932	0.0842893318	10243.019943
proteina	0.030317116	0.0003893978	18.076923
gordura	0.887918489	0.0869946033	126.720798
calcio	0.338254192	0.6617320111	6089.344729
ferro	0.009787159	0.0002638710	2.134103

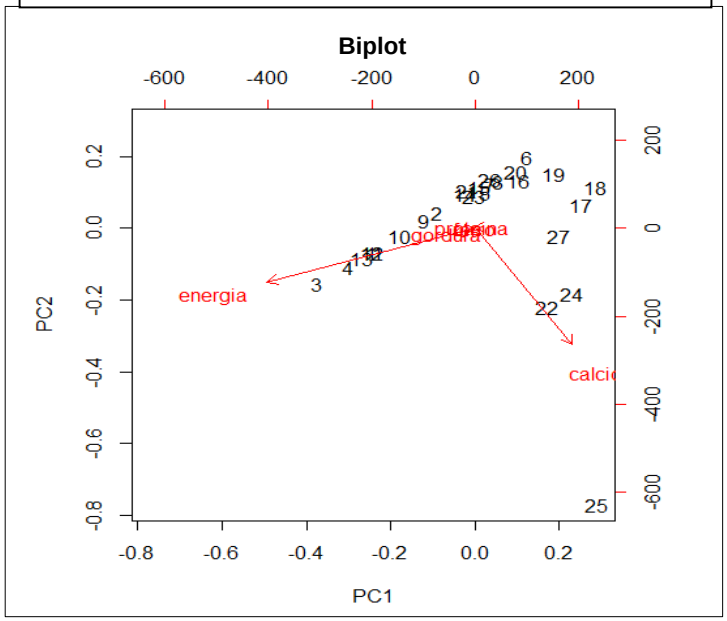
Análise de Componentes Principais

NÃO é invariante por padronização dos dados

Dados Originais: decomposição de S

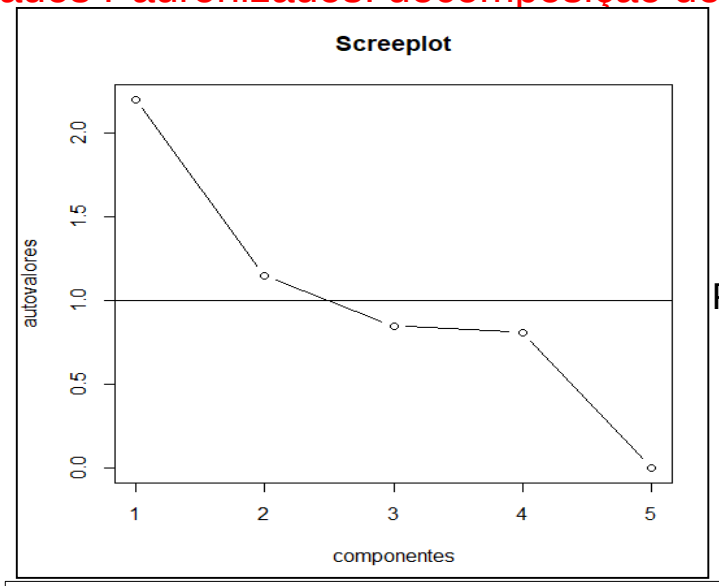
Prop.Ac.:
0.701
0.998

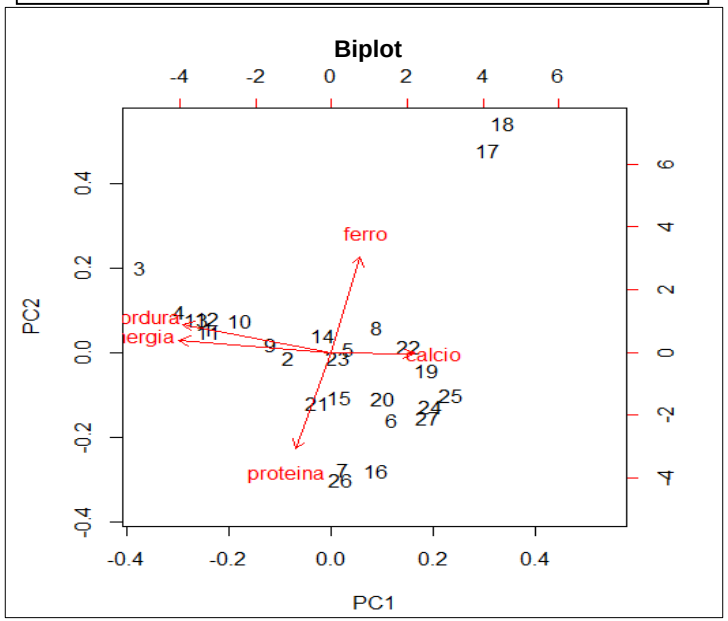




Dados Padronizados: decomposição de R

Prop.Ac.Expl.:
0.44
0.67
0.84





Componentes Principais
Quantos Componentes Reter na Análise?

$$Y_{n\times p};\quad S = V\Lambda V'$$

$$Y_i \in \Re^p \quad \rightarrow \quad Z_i = V'_{m\times p}\ Y_{i_{p\times 1}} \in \Re^m \qquad m?$$

- Preservar “grande” parte da variância total dos dados:

Para variáveis padronizadas: $\lambda_j \geq 1$

$$\frac{\lambda_1 + \lambda_2 \ldots + \lambda_m}{tr\Sigma} \geq ? \quad {}^{0,70}$$

Devem ser retidos todos os CPj, com variância maior que a média:

$$\lambda_j \geq \frac{tr\Sigma}{p}$$

Critério de corte no *ScreePlot*: quando a variação entre os autovalores (λ) passa a ser pequena (*cotovelo do gráfico*)

- Garantir Correlações “Altas” entre as variáveis Originais e as CP:

- Garantir “grande” parte da variabilidade de cada variável original

Análise de Componentes Principais

Na prática, S e R são estimativas (MVS ou estimadores robustos) a serem usadas na decomposição espectral.

- Variáveis originais (Y) em escalas diferentes (com heterocedasticidade) podem ser padronizadas, o que equivale aos CP via R. Os resultados via S ou R NÃO são os mesmos e não há uma função relacionando-os.

- Quando o objetivo é o Agrupamento de observações, em geral, não há necessidade de padronização das variáveis. Contudo, se o objetivo é a construção de índices (Ex., ancestralidade, escore de qualidade de vida, escore de desempenho do atleta, etc.), recomenda-se padronizar as variáveis.

- A interpretação das CP é fundamental (termos como “média ponderada” e “diferença entre médias ponderadas” das variáveis são comumente utilizados). Os coeficientes/cargas/pesos (“ν_{jk}”) e as correlações (r_{YjZk}) das variáveis originais com os CP são úteis na interpretação dos componentes principais.

- A estrutura de S (ou R) é decisiva na análise de CP. Sob a estrutura uniforme, as p variáveis originais ganham pesos “próximos” na construção do CP1.

Escalonamento Multidimensional
ou
Análise de Coordenadas Principais

Dados Multivariados

$$D = \begin{pmatrix} 0 & & & \\ d_{21} & 0 & & \\ \dots & \dots & \dots & \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix}_{n \times n} \qquad C = \begin{pmatrix} 1 & & & \\ r_{21} & 1 & & \\ \dots & \dots & \dots & \\ r_{n1} & r_{n2} & \dots & 1 \end{pmatrix}_{n \times n}$$

Matriz de Distâncias entre indivíduos

Matriz de Similaridades entre indivíduos

D é conhecida, Y_{nxp} não é conhecida

Objetivos:

- A partir de matrizes de distância (D) ou similaridade (C) entre *n* objetos (ou observações), obter uma representação das correspondentes observações Y_{nxp} que geraram D ou C;
- Obter Eixos Principais (Coordenadas Principais) $\Rightarrow$ Identificar dimensões não conhecidas de observações multivariadas

Escalonamento Multidimensional

Análise baseada no espaço das observações (linhas da matriz de dados)

Escalonamento Multidimensional

Matriz de Distância Euclidiana entre os 7 cães
(considerando as p=6 variáveis)

Matriz de Distância Euclidiana

	1	2	3	4	5	6	7
1	0						
2	6.21	0					
3	18.70	24.34	0				
4	13.13	18.55	5.99	0			
5	4.83	9.44	18.38	13.64	0		
6	7.43	12.94	12.50	7.26	7.98	0	
7	2.03	6.62	19.20	13.78	5.09	8.67	0

Com base somente em D, como representar os 7 pontos em um gráfico? Como obter “mensurações” que gerariam D?

$$D_{7\times 7} \xrightarrow{?} Y_{7\times k}; \quad k = 2$$

Escalonamento Multidimensional

Distâncias (em km) entre 12 cidades ⇒ matriz de "distância" empírica

Não é matriz de distância Euclidiana é uma distância empírica!

	1	2	3	4	5	6	7	8	9	10	11	12
1	0											
2	244	0										
3	218	350	0									
4	284	77	369	0								
5	197	167	347	242	0							
6	312	444	94	463	441	0						
7	215	221	150	236	279	245	0					
8	469	583	251	598	598	169	380	0				
9	166	242	116	257	269	210	55	349	0			
10	212	53	298	72	170	392	168	531	190	0		
11	253	325	57	340	359	143	117	264	91	273	0	
12	270	168	284	164	277	378	143	514	174	111	256	0

Como representar os 12 pontos em um gráfico?

$$D_{12\times 12} \xrightarrow{?} Y_{12\times k}; \quad k = 2$$

Escalonamento Multidimensional

Matriz de Distância* (“postos”) entre 6 Pacientes

	A	B	C	D	E	F
A	-					
B	2	-				
C	13	12	-			
D	4	6	9	-		
E	3	5	10	1	-	
F	8	7	11	14	15	-

Não é matriz de distância Euclidiana é uma distância empírica!

*1: é o par mais similar 15: é o par menos similar (6(6-1)/2)

Como representar os 6 pacientes em um gráfico (a partir de D)?

$$D_{6\times 6} \xrightarrow{?} Y_{6\times k}; \quad k = 2$$

Escalonamento Multidimensional

Notação:

Dada uma matriz de distâncias, *D*,

$$D = \left(d_{ij}\right)_{n \times n}$$

o objetivo do Escalonamento Multidimensional é estimar pontos (Y_1, Y_2, …, Y_n), k-dimensionais, tal que, se $\hat{d}_{ij}$ é a distância Euclidiana entre Y_i e Y_j, então $\hat{D} = \left(\hat{d}_{ij}\right)$ é uma "aproximação" para D.

Solução (para obter as coordenadas principais Y_1,...,Y_n):

- Métodos métricos $\Rightarrow$ os pontos P são obtidos tal que $\hat{D} \cong D$
- Métodos não métricos $\Rightarrow$ baseados na ordenação das n(n-1)/2 distâncias e minimização de funções objetivo como o "stress"

Escalonamento Multidimensional

Solução Clássica em k dimensões

(Mardia, 1979)

Dado *D*, matriz de distância Euclidiana $\Leftrightarrow$ Existe $Y_{n\times p}$ matriz de dados tal que:

$$D=\left(d_{ij}\right)_{n\times n};\quad d_{ij}^2=\sum_{k=1}^{p}\left(y_{ik}-y_{jk}\right)^2$$

d_{ij} : conhecido

y_{ik} : desconhecido e queremos "conhecer"

Deve existir: $B_{n\times n}=Y_{n\times p}\,Y'_{p\times n} \Rightarrow b_{ij}=\sum_{k=1}^{p} y_{ik}y_{jk}$

$$\begin{cases} d_{ij}^2=b_{ii}+b_{jj}-2b_{ij} \\ b_{ij}=-\frac{1}{2}\left(d_{ij}^2-d_{i.}^2-d_{.j}^2+d_{..}^2\right) \end{cases}$$

B:calculada de D

Escalonamento Multidimensional

Solução Clássica em k dimensões

$$D = \left(d_{ij}\right)_{n\times n} \quad \Leftrightarrow \quad Y_{n\times p}\ ? \qquad d_{ij}^2 = \sum_{k=1}^{p}\left(y_{ik} - y_{jk}\right)^2$$

$$B = YY' \quad \Rightarrow \quad B = \left(b_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2\right)\right)$$

Dado D, calcular B e então obter Y da decomposição spectral de B:

$B_{n\times n} = Y\,Y'$ Matriz p.s.d. (n>p) e sua Decomposição Espectral é:

$$= U\,\Lambda\,U' = U\Lambda^{1/2}\,\Lambda^{1/2}U' = \left(U\Lambda^{1/2}\right)\left(U\Lambda^{1/2}\right)' \quad \Rightarrow Y = \left(U\Lambda^{1/2}\right)$$

- Quando n>p, o posto de D é p. Logo, há (n-p) autovalores nulos.
- Podemos escolher uma representação para Y em uma dimensão k (k<p) .

$$Y = U\Lambda^{1/2} = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ u_{21} & u_{22} & \dots & u_{2n} \\ & & \dots & \\ u_{n1} & u_{n1} & \dots & u_{nn} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & \sqrt{\lambda_n} \end{pmatrix}$$

$$n > p \Rightarrow \lambda_{(p+1)} = \dots = \lambda_n = 0$$

Escalonamento Multidimensional

Solução Clássica em k dimensões

$$D = \left(d_{ij}\right)_{n\times n} \quad \Leftrightarrow \quad Y_{n\times p} \; ?$$

$$B_{n\times n} = \left(b_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2\right)\right)$$

Matriz de distâncias

Obter os “k” primeiros componentes da decomposição espectral de B:

Autovalores: $\lambda_1 > \lambda_2 > \ldots > \lambda_k > \lambda_{k+1} > \ldots > \lambda_n > 0$

Autovetores normalizados: $U = \left(U_1, U_2, \ldots, U_k, U_{k+1}, \ldots, U_p, U_{p+1}, \ldots, U_n\right)$

$\Rightarrow$ As coordenadas do vetor de resposta Y_i são obtidas a partir da *i*-ésima linha da matriz $U=(u_{ij})$

$$Y_i = \left(U_1 \ldots U_k\right)_i \; \Lambda^{1/2} = \left(u_{i1}\sqrt{\lambda_1}, u_{i2}\sqrt{\lambda_2}, \ldots, u_{ik}\sqrt{\lambda_k}\right)$$

Escalonamento Multidimensional

Matriz de Distância Euclidiana entre os 7 cães (p=6)

Matriz de distância Euclidiana entre as observações

	1	2	3	4	5	6	7
1	0.00	6.21	18.70	13.13	4.83	7.43	2.03
2	6.21	0.00	24.34	18.55	9.44	12.94	6.62
3	18.70	24.34	0.00	5.99	18.38	12.50	19.20
4	13.13	18.55	5.99	0.00	13.64	7.26	13.78
5	4.83	9.44	18.38	13.64	0.00	7.98	5.09
6	7.43	12.94	12.50	7.26	7.98	0.00	8.67
7	2.03	6.62	19.20	13.78	5.09	8.67	0.00

Matriz B

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
[1,]	23.09	48.94	-66.33	-38.97	17.32	-9.68	25.62
[2,]	48.94	113.39	-142.41	-79.68	29.54	-20.68	50.89
[3,]	-66.33	-142.41	194.13	114.78	-54.35	25.24	-71.07
[4,]	-38.97	-79.68	114.78	71.28	-39.90	15.65	-43.17
[5,]	17.32	29.54	-54.35	-39.90	34.88	-8.09	20.62
[6,]	-9.68	-20.68	25.24	15.65	-8.09	12.69	-15.14
[7,]	25.62	50.89	-71.07	-43.17	20.62	-15.14	32.25

Escalonamento Multidimensional

Dados dos Cães (n=7; p=6)

Matriz de distância Euclidiana entre as observações

	1	2	3	4	5	6	7
1	0.00	6.21	18.70	13.13	4.83	7.43	2.03
2	6.21	0.00	24.34	18.55	9.44	12.94	6.62
3	18.70	24.34	0.00	5.99	18.38	12.50	19.20
4	13.13	18.55	5.99	0.00	13.64	7.26	13.78
5	4.83	9.44	18.38	13.64	0.00	7.98	5.09
6	7.43	12.94	12.50	7.26	7.98	0.00	8.67
7	2.03	6.62	19.20	13.78	5.09	8.67	0.00

Decomposição Espectral da Matriz B

Autovalores

435.64	29.17	12.78	3.95	0.15	0.03	0.00

Dimensão: k=2 explica 96,49%

Autovetores

	U1	U2	U3	U4	U5	U6	U7
[1,]	-0.23	-0.05	0.05	0.25	-0.02	0.86	-0.38
[2,]	-0.49	-0.52	-0.06	-0.52	-0.22	-0.18	-0.38
[3,]	0.67	0.03	0.23	-0.19	-0.57	0.04	-0.38
[4,]	0.40	-0.31	0.07	-0.11	0.77	-0.03	-0.38
[5,]	-0.19	0.80	-0.05	-0.39	0.18	-0.05	-0.38
[6,]	0.10	0.00	-0.79	0.40	-0.11	-0.21	-0.38
[7,]	-0.25	0.05	0.55	0.55	-0.03	-0.43	-0.38

Coordenadas Principais:

	$\hat{Y}_1$	$\hat{Y}_2$
1	-4.76	-0.28
2	-10.23	-2.78
3	13.90	0.18
4	8.26	-1.68
5	-3.98	4.29
6	2.01	0.00
7	-5.21	0.26

$\sqrt{435.64}$ (–0.23)

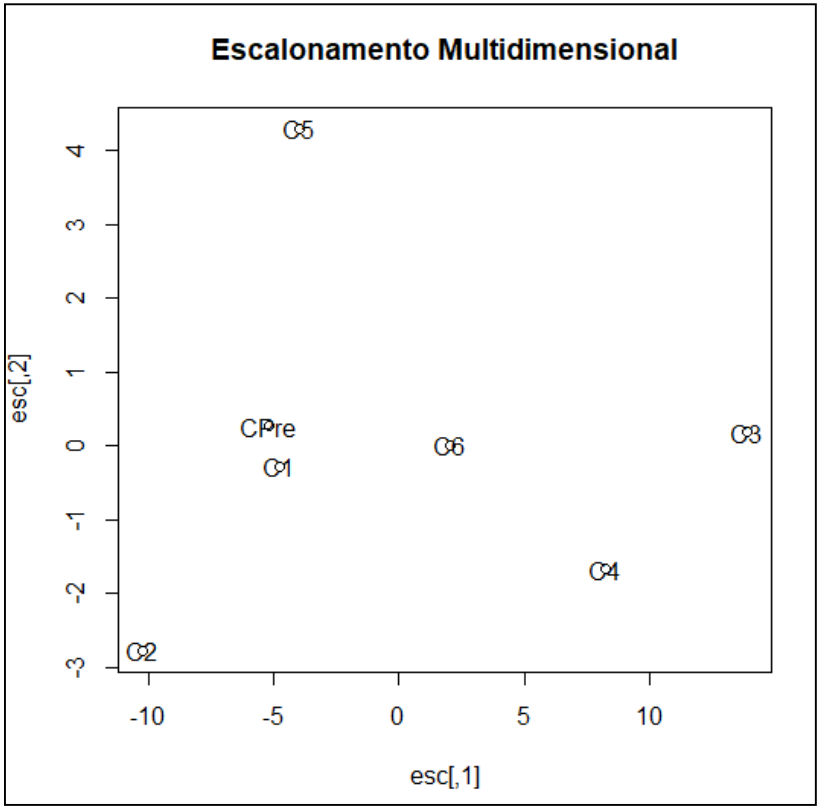
$\sqrt{29.17}$ (–0.05)

Escalonamento Multidimensional

Representação dos 7 cães obtida da matriz de dist6ancia Euclidiana (D)

Coordenadas Principais:

	$\hat{Y}_1$	$\hat{Y}_2$
1	-4.76	-0.28
2	-10.23	-2.78
3	13.90	0.18
4	8.26	-1.68
5	-3.98	4.29
6	2.01	0.00
7	-5.21	0.26



Escalonamento Multidimensional

Matriz de Distância Euclidiana (observada e predita) entre os 7 cães

Triangular superior: Matriz de distância Euclidiana observada (D)
Triangular inferior: Matriz de distância Euclidiana predita das Coordenadas Principais ($\hat{D}$)

	1	2	3	4	5	6	7
1	0	6.21	18.70	13.13	4.83	7.43	2.03
2	6.01	0	24.34	18.55	9.44	12.94	6.62
3	18.67	24.31	0	5.99	18.38	12.50	19.20
4	13.10	18.52	5.94	0	13.64	7.26	13.78
5	4.64	9.44	18.35	13.62	0	7.98	5.09
6	6.78	12.55	11.89	6.48	7.36	0	8.67
7	0.70	5.87	19.11	13.61	4.21	7.22	0

$= D$

$= \hat{D}$

Avaliar a qualidade da representação das observações em $\Re^2$

Análise de Componentes Principais e Coordenadas Principais

$$\hat{\Sigma}_{p\times p} = S$$

Análise de CP
Decomposição espectral de Σ

$$Y_{n\times p};\ \Sigma_{p\times p} = V_{p\times p}\Lambda_p V'_{p\times p} \Rightarrow Z_{n\times p} = YV$$

Decomposição de Y em valores singulares

$$Y = U_{n\times n}\Lambda_n^{1/2}V'_{p\times p};\quad n \geq p \Rightarrow \lambda_{p+1} = \ldots = \lambda_n = 0$$

$$YV = U\,\Lambda^{1/2}V'V = U\Lambda^{1/2}$$

↓ Componentes principais ↓ Coordenadas principais

Análise de EM
Decomposição espectral de B

$$Y_{n\times p}; D_{n\times n} \rightarrow B_{n\times n} = U_{n\times n}\Lambda_n U'_{n\times n} \Rightarrow Y_{n\times p} = U_{n\times p}\Lambda^{1/2}$$

- Os k primeiros Componentes Principais são "ótimos" ⇒ a soma das variâncias é maior do que qualquer outro conjunto de k combinações lineares não correlacionadas
- as k primeiras Coordenadas Principais são "ótimas" ⇒ a projeção de Y no sub-espaço de dimensão k de $\Re^p$ é mais próxima (em distância Euclidiana) da configuração original do que qualquer outra ($\hat{D} \cong D$)

Componentes Principais – Coordenadas Principais
Solução via Espaços Duais

$Y_{n\times p}$: Matriz de dados ("padronizados") multivariados de posto *r=min(n,p)*

Análise no espaço das variáveis: $\Re^{pxp}$

$$Y'Y = \Sigma_{p\times p} = V_{p\times p}\begin{pmatrix}\Lambda_r & 0\\ 0 & 0\end{pmatrix}V'_{p\times p} \Rightarrow \underset{\approx Y}{Y_{n\times p}V_{p\times r}}$$

r Componentes Principais

Análise no espaço dos indivíduos: $\Re^{nxn}$

$$D_{n\times n} \Rightarrow B_{n\times n} = YY' = U_{n\times n}\begin{pmatrix}\Lambda_r & 0\\ 0 & 0\end{pmatrix}U'_{n\times n} \Rightarrow \underset{\approx Y}{U\Lambda_r^{1/2}}$$

Escalonamento Multidimensional:
r Coordenadas Principais
obtidas da Matriz de Distâncias

Análise no espaço $\Re^{nxp}$

$$Y_{n\times p} = U_{n\times n}\begin{pmatrix}\Lambda_r^{1/2} & 0\\ 0 & 0\end{pmatrix}V'_{p\times p} \Rightarrow \boxed{Y_{n\times p}V_{p\times r} = U_{n\times n}\Lambda_r^{1/2}}$$

Equivalência entre os Componentes Principais e as Cordenadas Principais

n<<p : Componentes Principais de Y podem ser obtidos da decomposição espectral da matriz de distâncias D (nxn), de dimensão muito menor que Σ (pxp)

Coordenadas Principais

Dados Cães: n=7; p=6 variáveis

	X1	X2	X3	X4	X5	X6
[1,]	9.7	21.0	19.4	7.7	32.0	36.5
[2,]	8.1	16.7	18.3	7.0	30.3	32.9
[3,]	13.5	27.3	26.8	10.6	41.9	48.1
[4,]	11.5	24.3	24.5	9.3	40.0	44.6
[5,]	10.7	23.5	21.4	8.5	28.8	37.6
[6,]	9.6	22.6	21.1	8.3	34.4	43.1
[7,]	10.3	22.1	19.1	8.1	32.2	35.0

Matriz de Distância Euclidiana (D)

	1	2	3	4	5	6	7
1	0.00	6.21	18.70	13.13	4.83	7.43	2.03
2	6.21	0.00	24.34	18.55	9.44	12.94	6.62
3	18.70	24.34	0.00	5.99	18.38	12.50	19.20
4	13.13	18.55	5.99	0.00	13.64	7.26	13.78
5	4.83	9.44	18.38	13.64	0.00	7.98	5.09
6	7.43	12.94	12.50	7.26	7.98	0.00	8.67
7	2.03	6.62	19.20	13.78	5.09	8.67	0.00

Coordenadas Principais

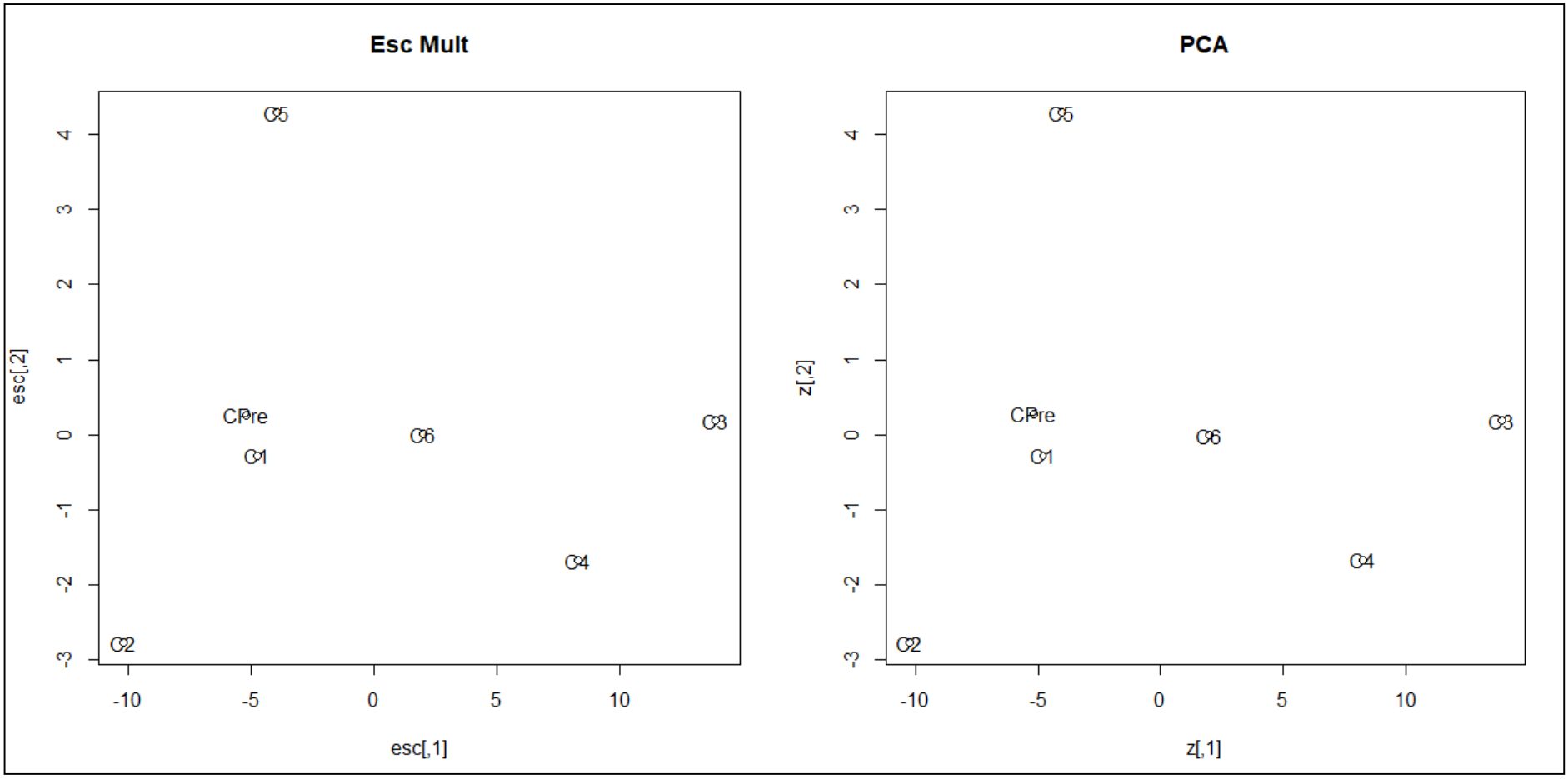
	$\hat{Y}_1$	$\hat{Y}_2$
1	-4.76	-0.28
2	-10.23	-2.78
3	13.90	0.18
4	8.26	-1.68
5	-3.98	4.29
6	2.01	0.00
7	-5.21	0.26

Escalonamento Multidimensional

As Coordenadas Principais representam uma escala (dados) construída a partir da informação das distâncias (D)

Componentes Principais e Coordenadas Principais
Equivalência

Coordenadas Principais obtidas de D $\Rightarrow$ Representação (em $\Re^2$) equivalente aos Componentes Principais obtidos de S



Escalonamento Multidimensional

A análise de Coordenadas Principais (Escalonamento Multidimensional) é baseada em uma matriz de Distâncias (nxn) entre observações enquanto a análise de Componentes Principais é baseada em uma matriz de covariâncias (pxp) entre variáveis.
Equivalências entre essas análises:

1. A análise de Coordenadas Principais da matriz de distâncias Euclidianas é equivalente à análise de Componentes Principais da matriz de covariâncias.
2. A análise de Coordenadas Principais da matriz de distâncias de Penrose é equivalente à análise de Componentes Principais da matriz de correlação.

A análise de Coordenadas Principais pode ser aplicada de maneira mais geral, para diferentes escolhas de matriz de distâncias entre observações (Mahalanobis, Manhattan, entre outras). Neste caso, NÃO está garantida a equivalência entre as duas análises.

Escalonamento Multidimensional

Métodos Não-Métricos

- **D** é considerada uma matriz de "dissimilaridade" geral (não precisa ser de distância Euclidiana)
- Os elementos de **D** podem ser ordenados

$$d_{ij}^{(1)} \le d_{ij}^{(2)} \le \ldots \le d_{ij}^{(m)}; \quad m = n(n-1)/2$$

- Seja $\hat{D}$, tal que os elementos $\hat{d}_{ij}$ estão monotonicamente relacionados aos elementos d_{ij}

$$d_{ij} < d_{rs} \quad \Rightarrow \quad \hat{d}_{ij} \le \hat{d}_{rs} \quad ;\mathrm{i} < \mathrm{j}, \mathrm{r} < \mathrm{s}$$

- Seja Y uma configuração em $\Re^k$ com distâncias $\hat{d}_{ij}$. Y é ótima no sentido de minimizar a seguinte medida:

$$S^2(Y) = \frac{\sum_{i<j} \left(d_{ij} - \hat{d}_{ij} \right)^2}{\sum_{i<j} \left(d_{ij} - \bar{d} \right)^2}$$

Medida de stress de Y: mede quanto da variância de d_{ij} NÃO é explicada pelas k coordenadas principais

Escalonamento Multidimensional

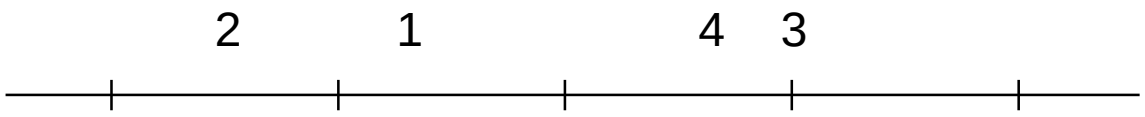
Distância Euclidiana entre os 7 cães (considerando as p=6 variáveis)

Matriz de Distância Euclidiana

	1	2	3	4	5	6	7
1	0						
2	6.21	0					
3	18.70	24.34	0				
4	13.13	18.55	5.99	0			
5	4.83	9.44	18.38	13.64	0		
6	7.43	12.94	12.50	7.26	7.98	0	
7	2.03	6.62	19.20	13.78	5.09	8.67	0

$$d_{34} < d_{12} < d_{14} < d_{24} < d_{13} < d_{23}$$

Tente localizar os cães 1, 2, 3 e 4 em uma única dimensão:

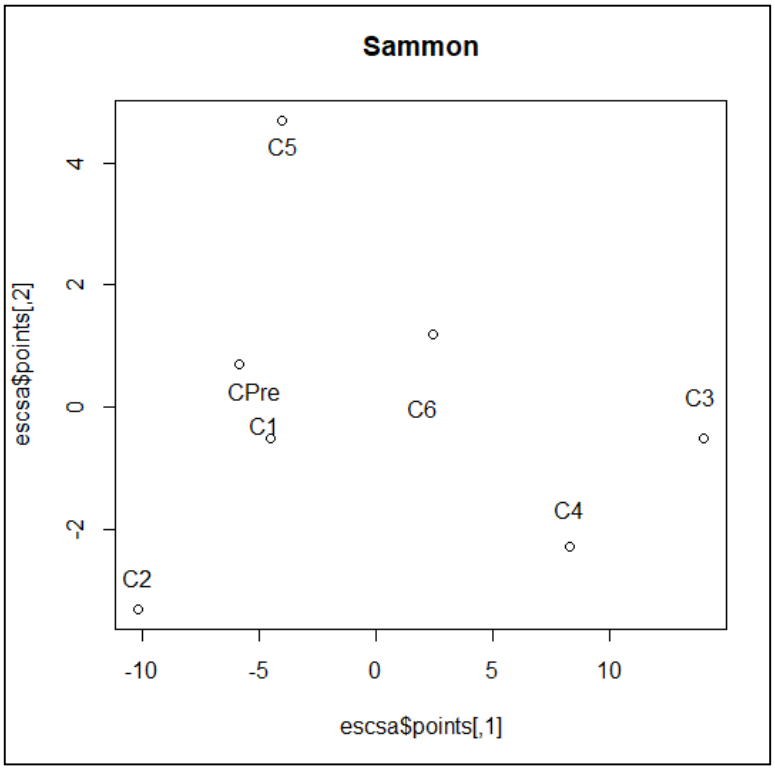


Obter uma escala que minimize o stress ?

Escalonamento Multidimensional

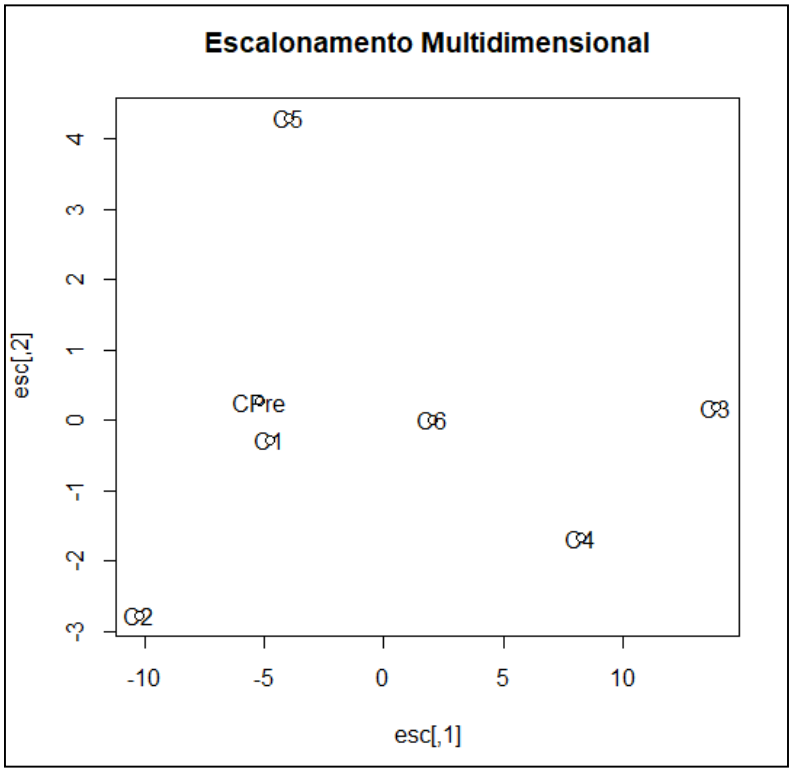
Solução Não Métrica: Sammon

	[,1]	[,2]
1	-4.53	-0.51
2	-10.20	-3.31
3	14.02	-0.51
4	8.27	-2.28
5	-4.07	4.70
6	2.40	1.20
7	-5.90	0.71



Solução Métrica

	[,1]	[,2]
1	-4.76	-0.28
2	-10.23	-2.78
3	13.90	0.18
4	8.26	-1.68
5	-3.98	4.29
6	2.01	0.00
7	-5.21	0.26



Análise de Correspondência

Análise de Correspondência

(Everitt, 2004)

	Variável Coluna					
u.a. / Variável Linha	1	2	…	j	…	J
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1J}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2J}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{iJ}
…	…	…	…	…		…
I	Y_{I1}	Y_{I2}		Y_{Ij}		Y_{IJ}

Identificar a estrutura dos dados multivariados com “Tabelas de Contingência”

Objetivos:

- Representar graficamente os dados dispostos em tabelas de contingência, de tal forma a ter uma visualização do padrão de associação entre variáveis $\Rightarrow$ os vetores linha e os vetores coluna da tabela são visualizados como pontos em um espaço vetorial
- Decompor a estatística χ^2 do teste de independência em tab. de contingência **TÉCNICA GRÁFICA MULTIDIMENSIONAL (similar ao Escalonamento!!)**

Análise de Correspondência Representação Simplex

Pense em como representar graficamente as populações Trinomiais (L1 a L5)!

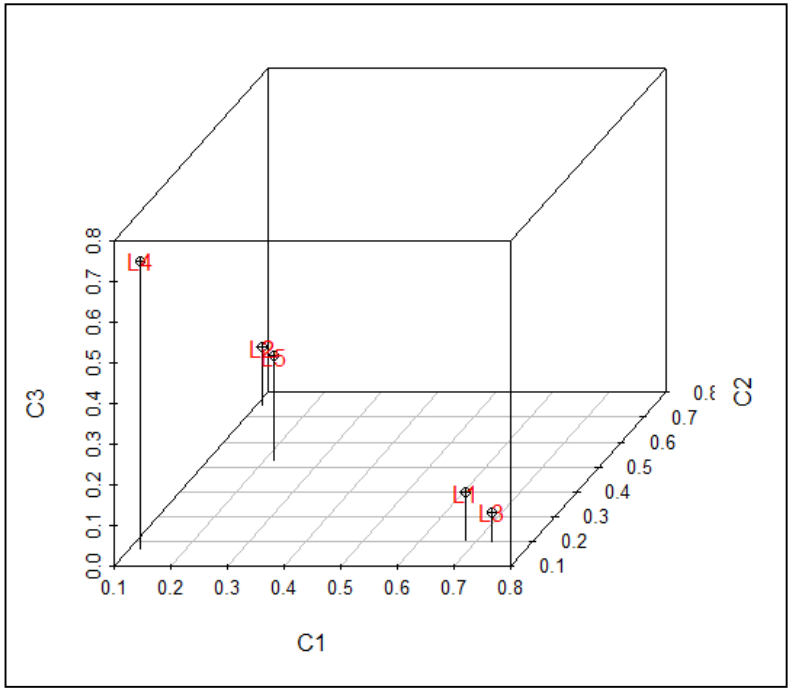
Exemplo 1

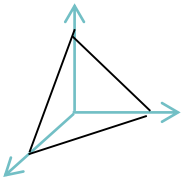
Tabela: Distribuição de 5 populações de acordo com o genótipo (Trinomiais)

	C1	C2	C3
L1	17	5	3
L2	3	20	4
L3	19	5	2
L4	6	8	35
L5	5	12	6

Tabela: Proporções (Linhas)

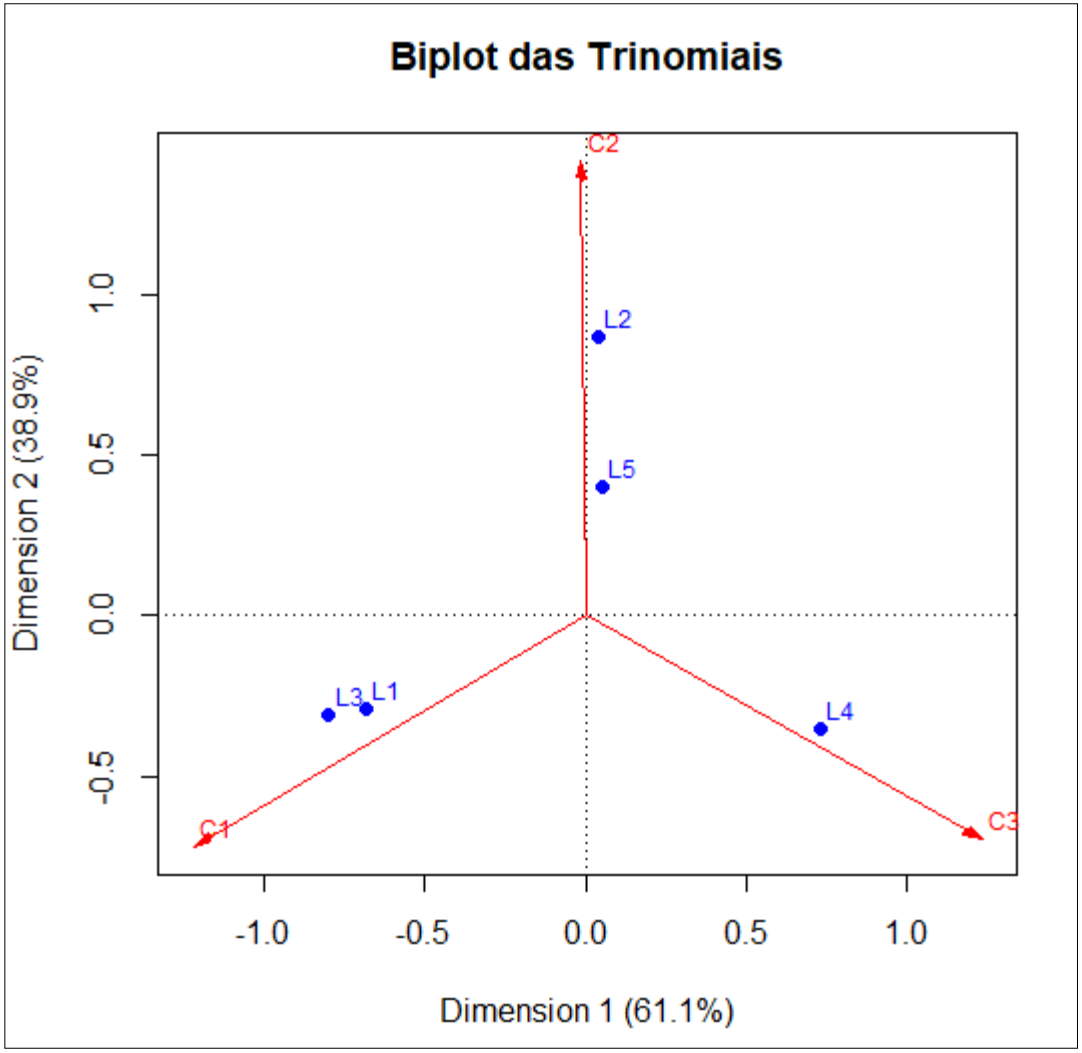
	C1	C2	C3
L1	0.68	0.20	0.12
L2	0.11	0.74	0.15
L3	0.73	0.19	0.08
L4	0.12	0.16	0.71
L5	0.22	0.52	0.26





Sob a restrição $p_1+p_2+p_3=1$, as Trinomiais podem ser representadas no plano $\Re^2$ (simplex) sem qualquer perda de informação. Os genótipos (C1, C2 e C3) definem os eixos do gráfico.

Representação biplot das populações Trinomiais (L1 a L5)!



Biplot: gráfico de Escores e Cargas

Este biplot representa um mapa assimétrico, ideal para tabelas com totais Linha fixos.

Biplot (`map=rowprincipal`)

A distribuição dos pontos (neste caso trinomiais) corresponde à informação da estatística Qui-Quadrado de homegeneidade entre populações (multinomiais)!!

Análise de Correspondência
Representação Simplex

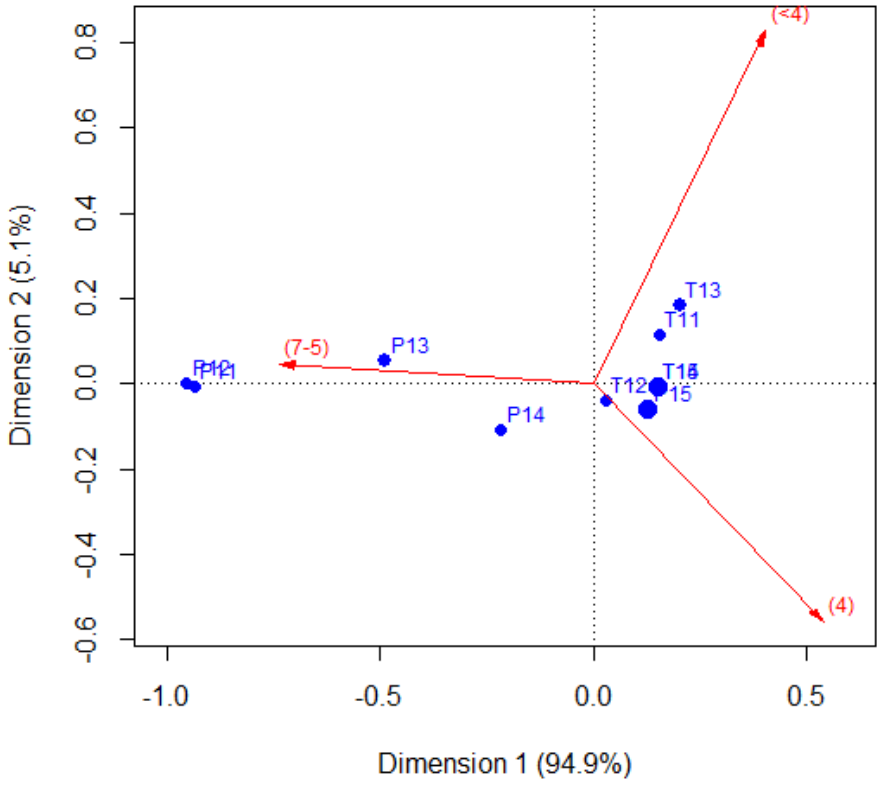
Exemplo 2: Distribuição do número de bulbilhos de alho de acordo com o tamanho (7-5, 4 e <4), tratamento e ano de plantio.

Ano	Tratamento*	Tamanho dos bulbilhos			Total
		7-5	4	<4	
2011	Padrão	417	36	0	453
	Teste	164	176	90	400
2012	Padrão	357	27	0	384
	Teste	169	161	54	384
2013	Padrão	800	240	103	1143
	Teste	412	458	274	1144
2014	Padrão	273	176	39	488
	Teste	185	220	83	488
2015	Padrão	1521	1794	585	3900
	Teste	1420	1681	635	3736

Represente as 10 trinomiais (variáveis nas linhas da tabela de contingência) no simplex. Este gráfico permite visualizar o padrão de heterogeneidade entre as populações trinomiais de acordo com o tamanho dos bulbilhos de alho. Interprete. Quais anos e qual tratamento produz os maiores bulbilhos?

Análise de Correspondência - Representação Simplex

Distribuição do número de bulbilhos de alho de acordo com tamanho, tratamento e ano de plantio.



Biplot: A variável Tamanho do bulbilho de alho está em Coordenadas Padrão (eixos do simplex) e Tratamento_Ano está em Coordenadas Principais.

Neste caso (trinomial), nenhuma informação é perdida nesta representação dos dados.

Biplot (`map=rowprincipal`)

Bulbilhos de tamanho 7-5 estão mais associados ao tratamento Padrão em 2011 (P11) e 2012 (P12), seguidos de 2013 (P13). O tratamento P14 mostra associação (mais fraca) com bulbilhos tanto de tamanho 7-5 e 4. Já os tratamentos Teste de 2011 (T11) a 2015 (T15), bem como o tratamento P15, estão mais associados com bulbilhos de tamanho menor (4 e <4).

Análise de Correspondência e Escalonamento Multidimensional

$Y_{I\times J}$

Variável Linha	Variável Coluna 1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

Perfis Linha de Proporções (Y^L)

Variável Linha	Variável Coluna 1	…	J	Total
1	$p_{11}=n_{11}/n_{1.}$		$p_{1J}=n_{1J}/n_{1.}$	1
…	…	…	…	…
I	$p_{I1}=n_{I1}/n_{I.}$		$p_{IJ}=n_{IJ}/n_{I.}$	1

$$\Rightarrow p_{ij}^{L} = \frac{n_{ij}}{n_{i.}}; \qquad \bar{p}_{.j} = \frac{n_{.j}}{n}$$

Perfis Coluna de Proporções (Y^C)

Variável Linha	Variável Coluna 1	…	J
1	$p_{11}=n_{11}/n_{.1}$		$p_{1J}=n_{1J}/n_{.J}$
…	…	…	…
I	$p_{I1}=n_{I1}/n_{.1}$		$p_{IJ}=n_{IJ}/n_{.J}$
Total	1	…	1

$$\Rightarrow p_{ij}^{C} = \frac{n_{ij}}{n_{.j}}; \qquad \bar{p}_{i.} = \frac{n_{i.}}{n}$$

Análise de Correspondência e Escalonamento Multidimensional

Análise das Matrizes Quadradas D^L e D^C

$Y_{I\times J}$

Variável Linha	Variável Coluna 1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

$$Y_{I\times J} \rightarrow Y^L_{I\times J} \rightarrow D^L_{I\times I}$$

Distância Qui-Quadrado dos Perfis Linha

$$D^L_{I\times I};\quad d_{ij}^{2\,Linhas} = \sum_{k=1}^{J} \frac{\left(p^L_{ik} - p^L_{jk}\right)^2}{\bar{p}_{.k}}$$

Distância Euclidiana ponderada

$$p^L_{ij} = \frac{n_{ij}}{n_{i.}} \qquad i = 1,2,\ldots,I$$

$$Y_{I\times J} \rightarrow Y^C_{I\times J} \rightarrow D^C_{J\times J}$$

Distância Qui-Quadrado dos Perfis Coluna

$$D^C_{J\times J};\quad d_{ij}^{2\,Colunas} = \sum_{k=1}^{I} \frac{\left(p^C_{ki} - p^C_{kj}\right)^2}{\bar{p}_{k.}}$$

Distância Euclidiana ponderada

$$p^c_{ij} = \frac{n_{ij}}{n_{.j}} \qquad j = 1,2,\ldots,J$$

Obter as **Coordenadas Principais** das Matrizes de distância Qui-Quadrado

$D^L_{I\times I}$ e $D^C_{J\times J}$

Os resultados são equivalentes à solução via *dvs (decomposição em valores singulares)* de Y^L e Y^C .

Análise de Correspondência - Escalonamento Multidimensional

Distribuição de funcionários de acordo com o tabagismo e nível funcional.

	F0	F1	F2	F3	Total
N1	4	2	3	2	11
N2	4	3	7	4	18
N3	25	10	12	4	51
N4	18	24	33	13	88
N5	10	6	7	2	25
Total	61	45	62	25	193

Proporção Linha: Y^L

	F0	F1	F2	F3	Total
N1	0.36	0.18	0.27	0.18	1
N2	0.22	0.17	0.39	0.22	1
N3	0.49	0.20	0.24	0.08	1
N4	0.20	0.27	0.38	0.15	1
N5	0.40	0.24	0.28	0.08	1
Total	0.32	0.23	0.32	0.13	1

Proporção Coluna: Y^C

	F0	F1	F2	F3	Total
N1	0.07	0.04	0.05	0.08	0.06
N2	0.07	0.07	0.11	0.16	0.09
N3	0.41	0.22	0.19	0.16	0.26
N4	0.30	0.53	0.53	0.52	0.46
N5	0.16	0.13	0.11	0.08	0.13
Total	1	1	1	1	1

Coord. Padrão: Autovetores de $D^L(Y^L)$

	Dim1	Dim2	Dim3
N1	-0.24	-1.94	3.49
N2	0.95	-2.43	-1.66
N3	-1.39	-0.11	-0.25
N4	0.85	0.58	0.16
N5	-0.74	0.79	-0.40

Coord. Padrão:Autovetores de $D^C(Y^C)$

	Dim1	Dim2	Dim3
F0	-1.44	-0.30	-0.04
F1	0.36	1.41	1.08
F2	0.72	0.07	-1.26
F3	1.07	-1.98	1.29

Inércias(autovalores de $D^L(D^C)$):

Dim 1:	0.07	87.80%
Dim 2:	0.01	11.76%
Dim3:	0.00	0.49%

dimensão máxima: min(I-1,J-1)

Análise de Correspondência - Escalonamento Multidimensional

Distribuição de funcionários de acordo com o tabagismo e nível funcional.

	F0	F1	F2	F3	Total
N1	4	2	3	2	11
N2	4	3	7	4	18
N3	25	10	12	4	51
N4	18	24	33	13	88
N5	10	6	7	2	25
Total	61	45	62	25	193

Outras representações:
Teste Qui-Quadrado de Homogeneidade (das Linhas)

χ^2=16.442 p=0.1718

Coord. Principais de D^L:

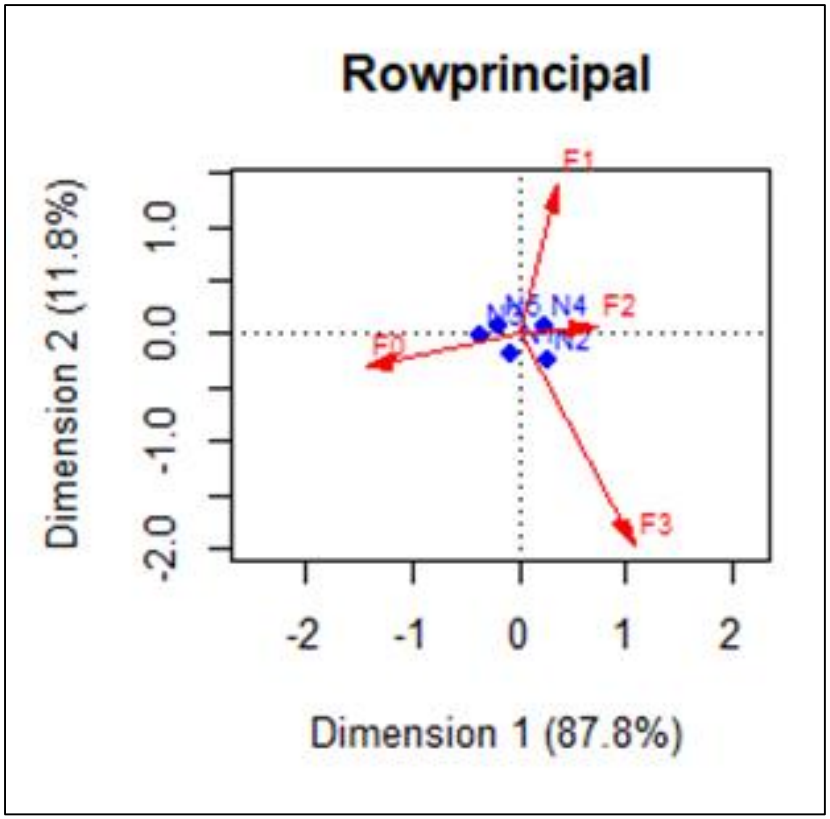
autovetor*sqrt(autovalor)

	Dim1	Dim2
N1	-0.07	-0.19
N2	0.26	-0.24
N3	-0.38	-0.01
N4	0.23	0.06
N5	-0.20	0.08

autovetor

Coordenadas Padrão de D^C:

	Dim1	Dim2
F0	-1.44	-0.30
F1	0.36	1.41
F2	0.72	0.07
F3	1.07	-1.98



Análise de Correspondência - Escalonamento Multidimensional

Distribuição de funcionários de acordo com o tabagismo e nível funcional.

	F0	F1	F2	F3	Total
N1	4	2	3	2	11
N2	4	3	7	4	18
N3	25	10	12	4	51
N4	18	24	33	13	88
N5	10	6	7	2	25
Total	61	45	62	25	193

Outras representações:

Teste Qui-Quadrado de Homogeneidade (das colunas)

χ^2=16.442 p=0.1718

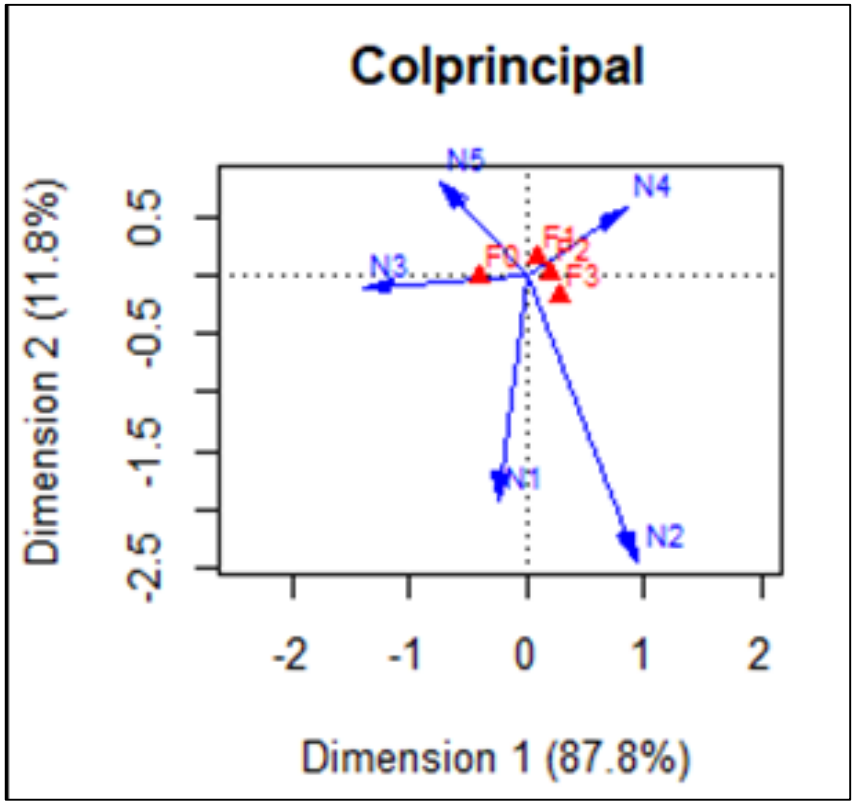
Coord. Padrão: Autovetores de $D^L(Y^L)$

	Dim1	Dim2	Dim3
N1	-0.24	-1.94	3.49
N2	0.95	-2.43	-1.66
N3	-1.39	-0.11	-0.25
N4	0.85	0.58	0.16
N5	-0.74	0.79	-0.40

Coord. Principais de D^C:

autovetor*sqrt(autovalor)

	Dim1	Dim2
F0	-0.39	-0.03
F1	0.10	0.14
F2	0.20	0.01
F3	0.29	-0.20



Análise de Correspondência - Escalonamento Multidimensional

Distribuição de funcionários de acordo com o tabagismo e nível funcional.

	F0	F1	F2	F3	Total
N1	4	2	3	2	11
N2	4	3	7	4	18
N3	25	10	12	4	51
N4	18	24	33	13	88
N5	10	6	7	2	25
Total	61	45	62	25	193

Teste Qui-Quadrado de Independência:

χ^2=16.442 p=0.1718

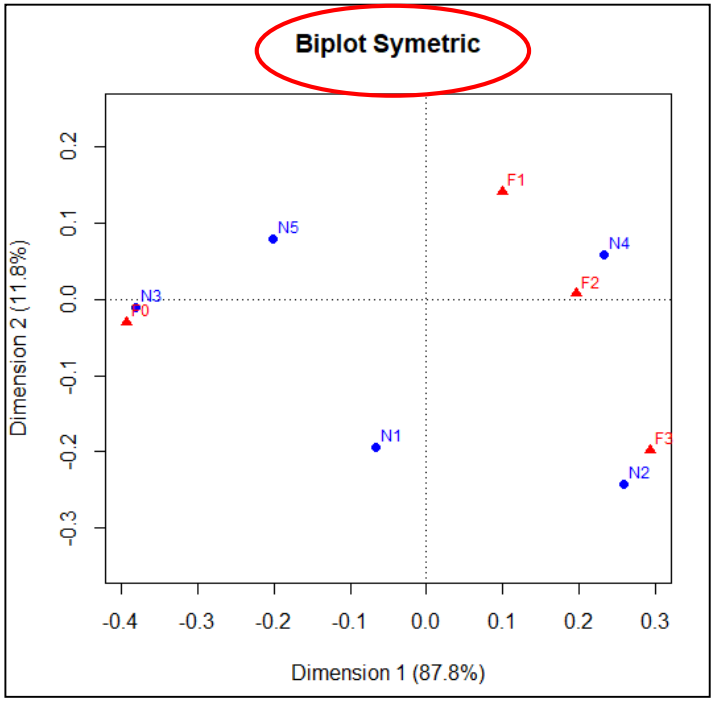
$in(I) = \chi^2 / n$ Inércia total: soma dos autovalores da decomposição de D^L (D^C)

Coord. Principais de D^L:

	Dim1	Dim2
N1	-0.07	-0.19
N2	0.26	-0.24
N3	-0.38	-0.01
N4	0.23	0.06
N5	-0.20	0.08

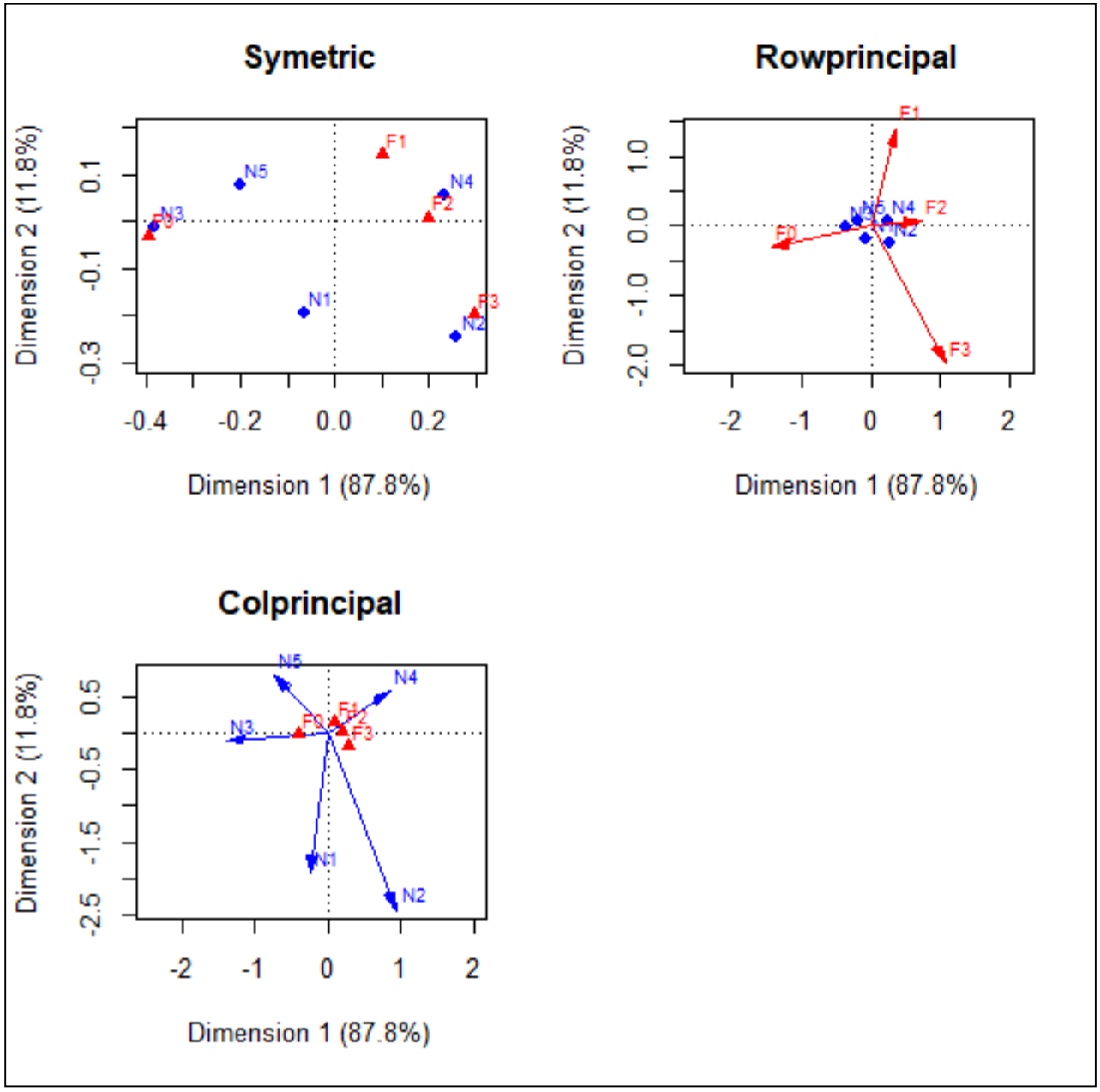
Coord. Principais de D^C:

	Dim1	Dim2
F0	-0.39	-0.03
F1	0.10	0.14
F2	0.20	0.01
F3	0.29	-0.20



Biplot simétrico: ambas as variáveis (Linha e Coluna) em coordenadas principais

Padrão de associação: Nível funcional N3 não fuma e N2 é o que fuma mais (F3)



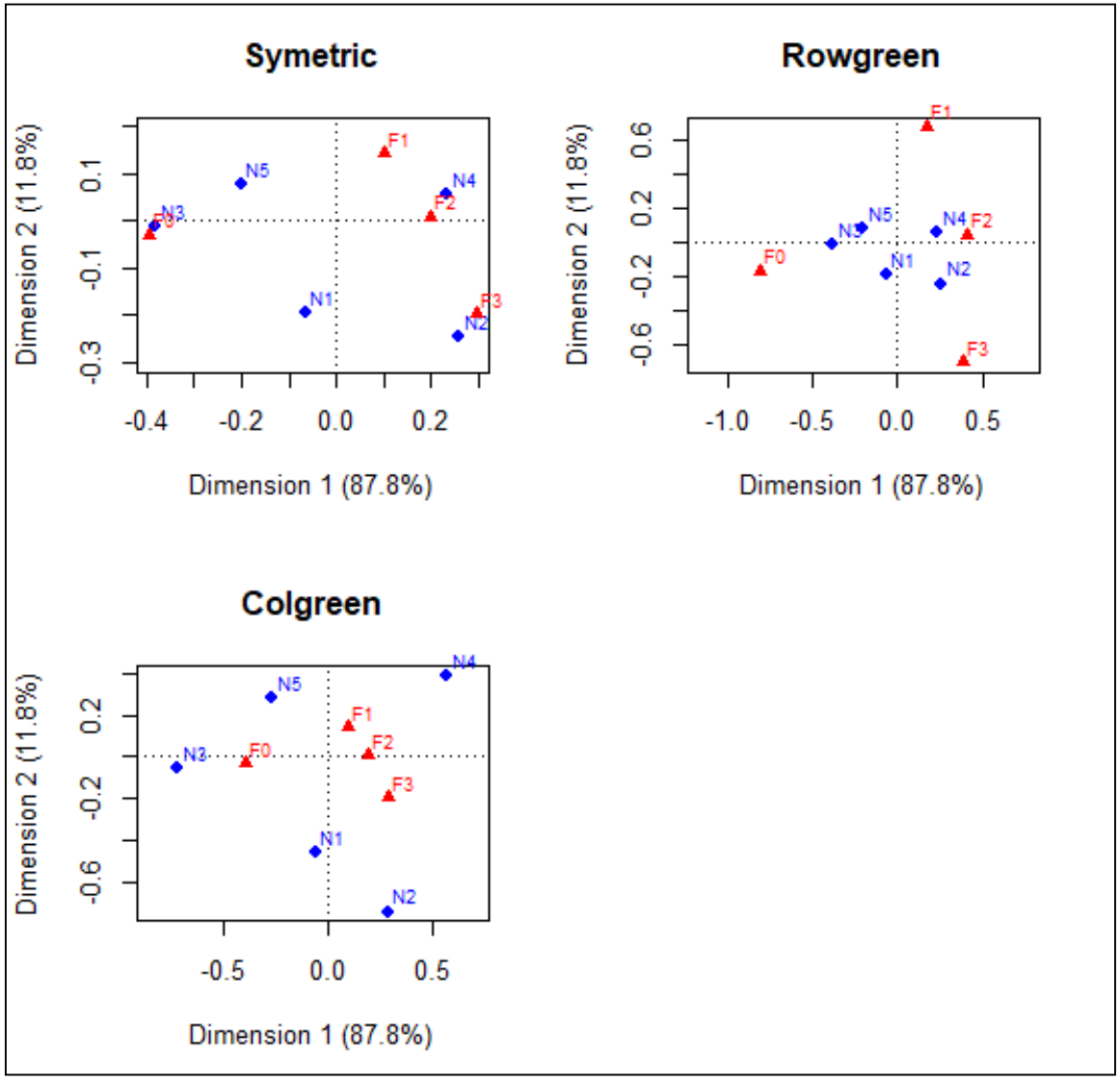
Rowprincipal: Linhas fixadas na tabela

Colprincipal: Colunas fixadas na tabela

Symetric: total geral da tabela fixado (Estudo transversal)

Análise de Correspondência - Escalonamento Multidimensional

Outras representações



Representação BiPlot: existem diferentes construções do BiPlot, visando diferentes padronizações dos eixos.

Mapa Simétrico: Linhas e Colunas em Coordenadas Principais

Mapas Assimétricos

-Rowgreen: Linhas em Coordenadas Principais e Colunas em Coordenadas Padronizadas (coord. Padrão*sqrt(massas))

-Colgreen: Colunas em Coordenadas Principais e Linhas em Coordenadas Padronizadas

Análise de Correspondência

Exemplo: (Greenacre, 2007)

Tabela A

	C1	C2	C3
L1	11	10	9
L2	10	11	9
L3	10	9	10
L4	9	9	12
L5	10	11	10

	C1	C2	C3
L1	0.37	0.33	0.30
L2	0.33	0.37	0.30
L3	0.34	0.31	0.34
L4	0.30	0.30	0.40
L5	0.32	0.35	0.32

X2=1.133, p=0.9973

Tabela C

	C1	C2	C3
L1	17	5	3
L2	3	20	4
L3	19	5	2
L4	6	8	35
L5	5	12	6

	C1	C2	C3
L1	0.68	0.20	0.12
L2	0.11	0.74	0.15
L3	0.73	0.19	0.08
L4	0.12	0.16	0.71
L5	0.22	0.52	0.26

X2=88.843, p=7.982e-16

Tabela B

	C1	C2	C3
L1	13	8	9
L2	6	14	10
L3	14	7	8
L4	7	9	18
L5	10	12	5

	C1	C2	C3
L1	0.43	0.27	0.30
L2	0.20	0.47	0.33
L3	0.48	0.24	0.28
L4	0.21	0.26	0.53
L5	0.37	0.44	0.19

X2=16.513, p=0.0356

Tabela D

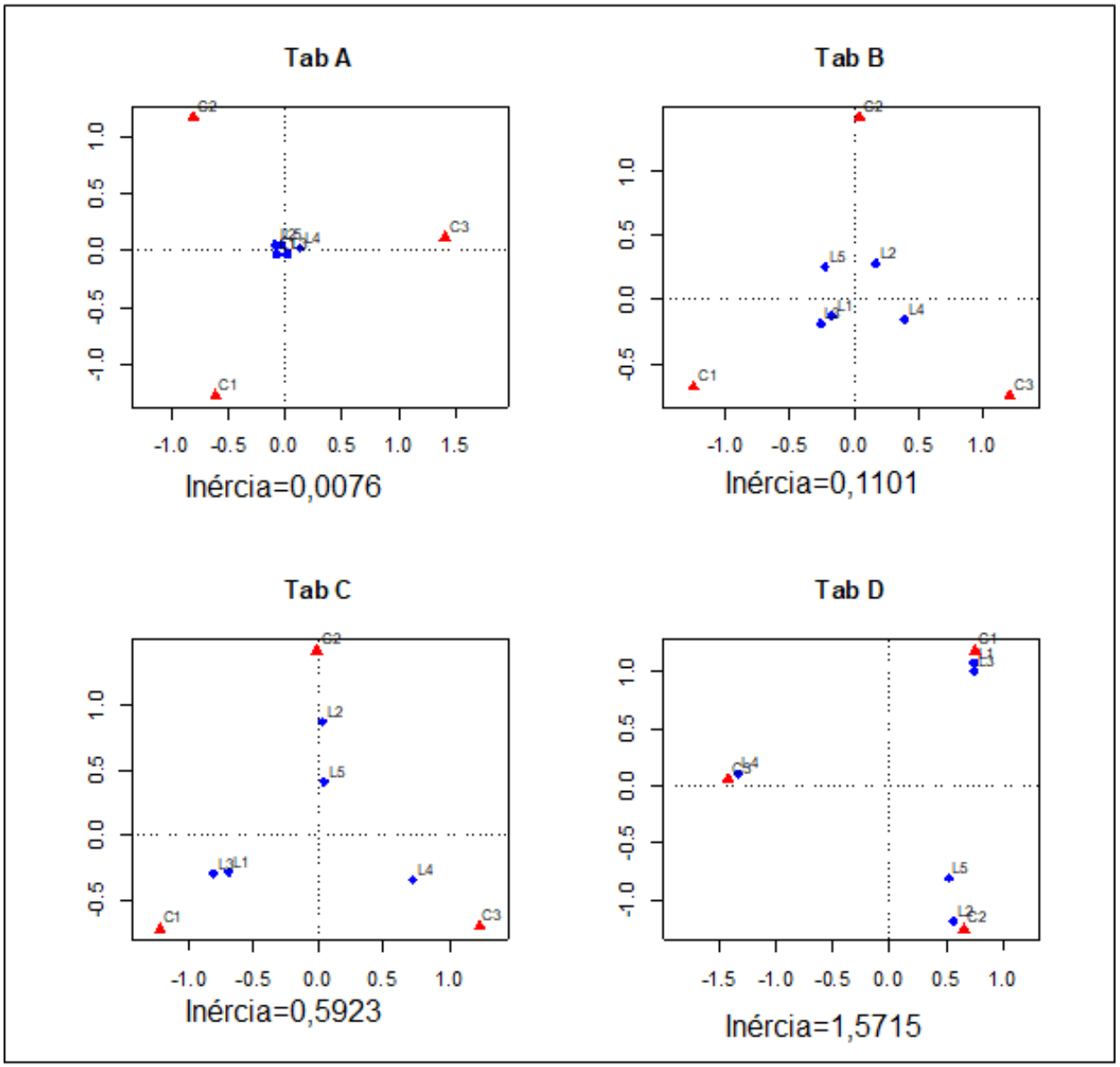
	C1	C2	C3
L1	20	1	0
L2	0	24	1
L3	24	2	0
L4	2	0	47
L5	4	23	2

	C1	C2	C3
L1	0.95	0.05	0.00
L2	0.00	0.96	0.04
L3	0.92	0.08	0.00
L4	0.04	0.00	0.96
L5	0.14	0.79	0.07

X2=235.731, p< 2.2e-16

Em cada caso calcule: vetores de proporções das ***trinomiais*** (linha), centróide, vetor de massas, vetor de pesos, distância Qui-Quadrado entre L1 e L2 e entre L1 (L2) e o centróide, inércia total. Obtenha a representação das 5 trinomiais no simplex correspondente. Interprete.

BiPlot e Inércias



Representação BiPlot
Mapa Assimétrico

- Linhas (trinomiais) em Coordenadas Principais

-Colunas em Coordenadas Padrão (vértices do simplex)