MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Sem/2022 - IME

Análise Multivariada

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

Já vimos ☺

- Estatísticas descritivas multivariadas, Episóides de Concentração, Boxplot Bivariado
- Distribuição N_p, Distribuições Amostrais (T^2 e W_p)
- $N_p(\mu_g;\Sigma_g)$: Inferências sobre μ_g (T^2, MANOVA, ICS, Correções para Múltiplos testes

Decomposições: Y_{nxp} , $SS_{T\ pxp}$,e D_{nxn}

Técnicas Multivariadas:

✓ Análise de Componentes Principais

- Escalonamento Multidimensional
- Análise de Correspondência

- Análise Fatorial
- Análise Discriminante (MANOVA)
- Análise de Agrupamento
- Análise de Correlação Canônica

Dados dos cães

	Y1	Y2	Y3	Y4	Y5	Y6
[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

Revisando ☺ Componentes Principais

$$Y_{n\times p} \rightarrow Z_{n\times p} = YV; \quad Z_{ik} = V_k' Y_i$$

$$\Sigma_{p\times p} = V\Lambda V' \rightarrow Cov(Z) = \Lambda = Diag(\lambda_j)$$

$$tr(\Sigma) = tr(\Lambda) = \sum_{j=1}^{p} \lambda_j$$

Matriz de Covariância S:

	Y1	Y2	Y3	Y4	Y5	Y6
Y1	**2.88**	5.25	4.85	1.93	6.53	7.74
Y2	5.25	**10.56**	8.90	3.59	11.46	15.58
Y3	4.85	8.90	**9.61**	3.51	13.43	16.31
Y4	1.93	3.59	3.51	**1.36**	4.86	5.92
Y5	6.53	11.46	13.43	4.86	**24.36**	24.68
Y6	7.74	15.58	16.31	5.92	24.68	**31.52**

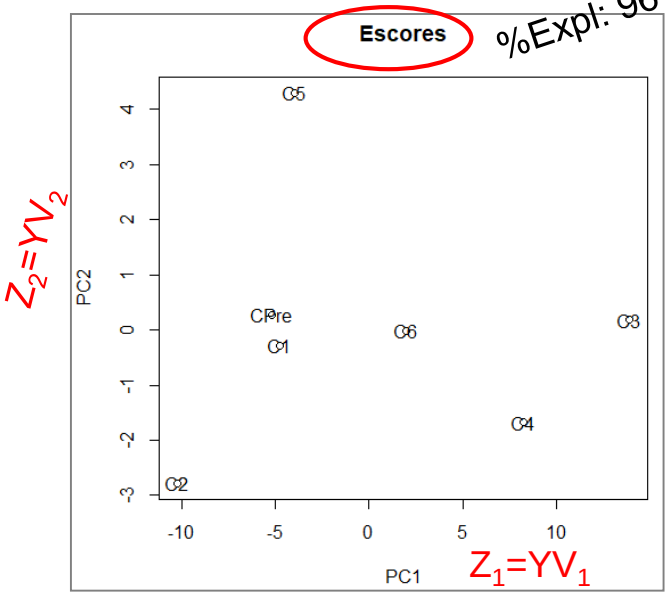
Autovalores de S (λ_j):

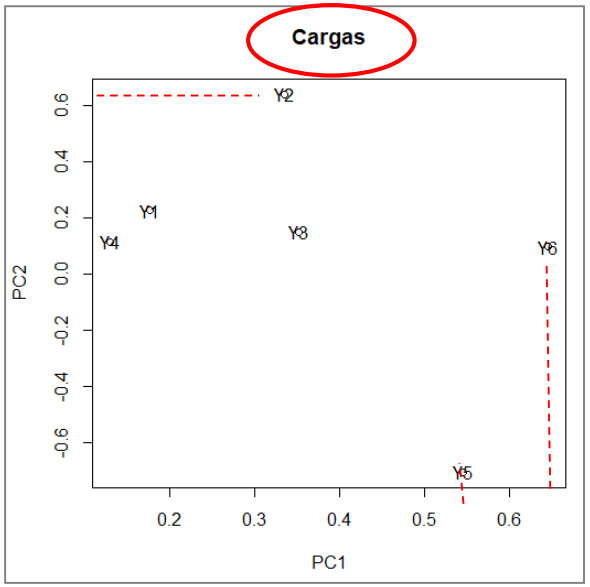
72.61 4.86 2.13 0.66 0.02 0.01

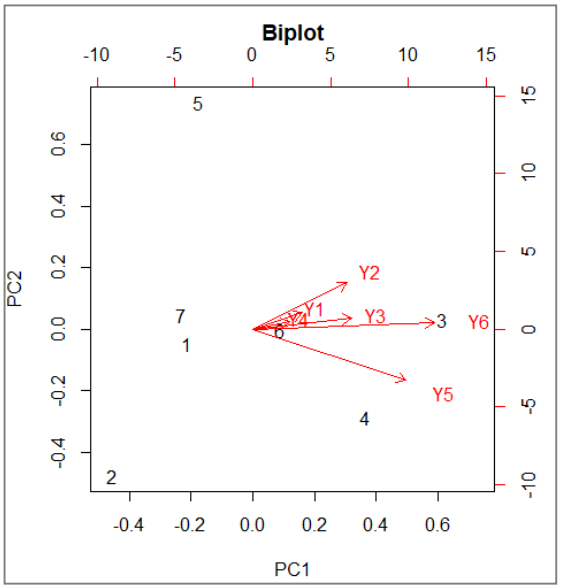
Autovetores de S (V_j):

	V1	V2	V3	V4	V5	V6
Y1	0.18	0.23	-0.41	0.10	0.65	-0.56
Y2	0.34	0.64	-0.33	-0.47	-0.37	0.09
Y3	0.35	0.15	-0.15	0.84	-0.36	0.01
Y4	0.13	0.11	-0.15	0.11	0.52	0.82
Y5	0.55	-0.70	-0.39	-0.21	-0.09	0.03
Y6	0.65	0.10	0.72	-0.08	0.18	-0.09

Cargas



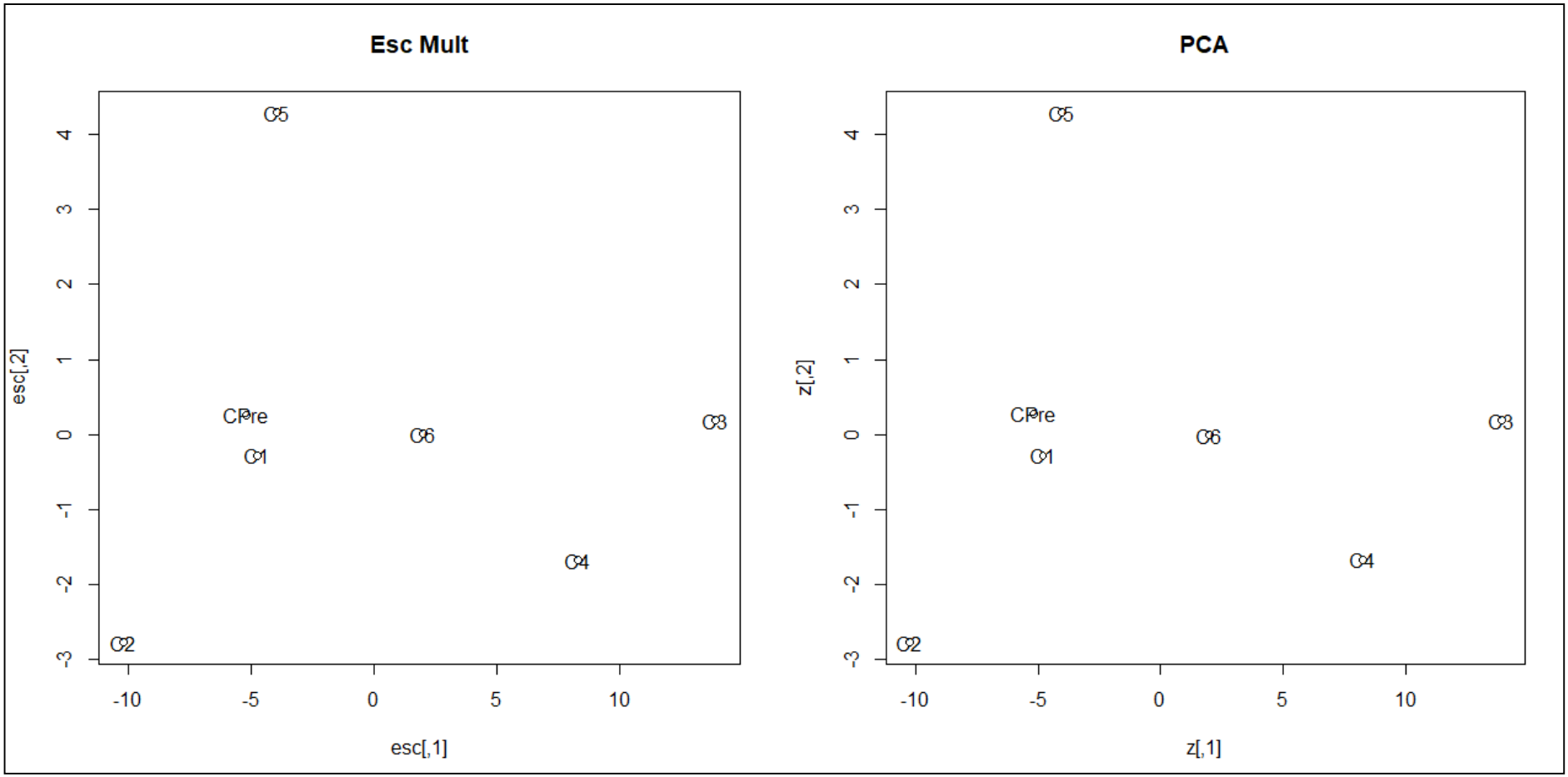




Componentes Principais e Coordenadas Principais
Equivalência

Já vimos

Coordenadas Principais obtidas de D_{nxn} $\Rightarrow$ Representação (em $\Re^2$) equivalente aos Componentes Principais obtidos de S_{pxp}



Análise de Correspondência

Análise de Correspondência

u.a. / Variável Linha	Variável Coluna 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}

$Y_{n\times p} \rightarrow$

Matriz de dados multivariados → Dados dispostos em "Tabelas de Contingência"

Objetivos:

$$Y_{n\times p} \rightarrow Y_{I\times J}$$

- Descrever graficamente os dados dispostos em tabelas de contingência, de tal forma a representar o padrão de associação entre variáveis ⇒ os vetores linha e os vetores coluna da tabela são visualizados como pontos em um espaço vetorial.
- Decompor a estatística χ^2 em tabelas de contingência.

TÉCNICA GRÁFICA MULTIDIMENSIONAL (similar ao Escalonamento!!)

(análise descritiva de dados categóricos dispostos em tabelas de contingência)

Análise de Correspondência
Representação Simplex

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

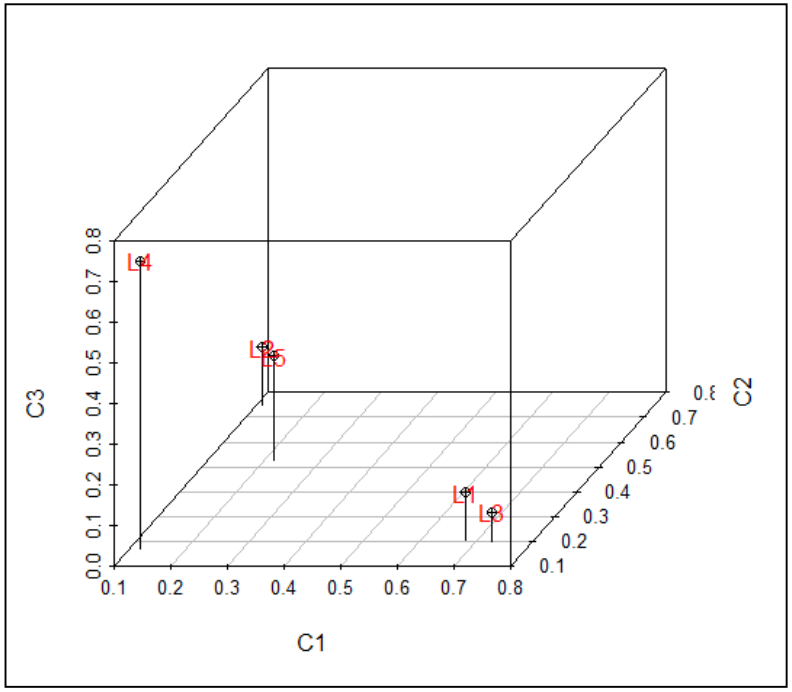
Exemplo 1

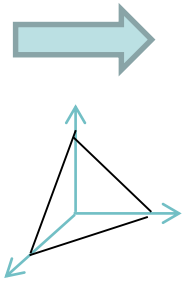
Tabela: Distribuição de 5 populações Trinomiais

	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)

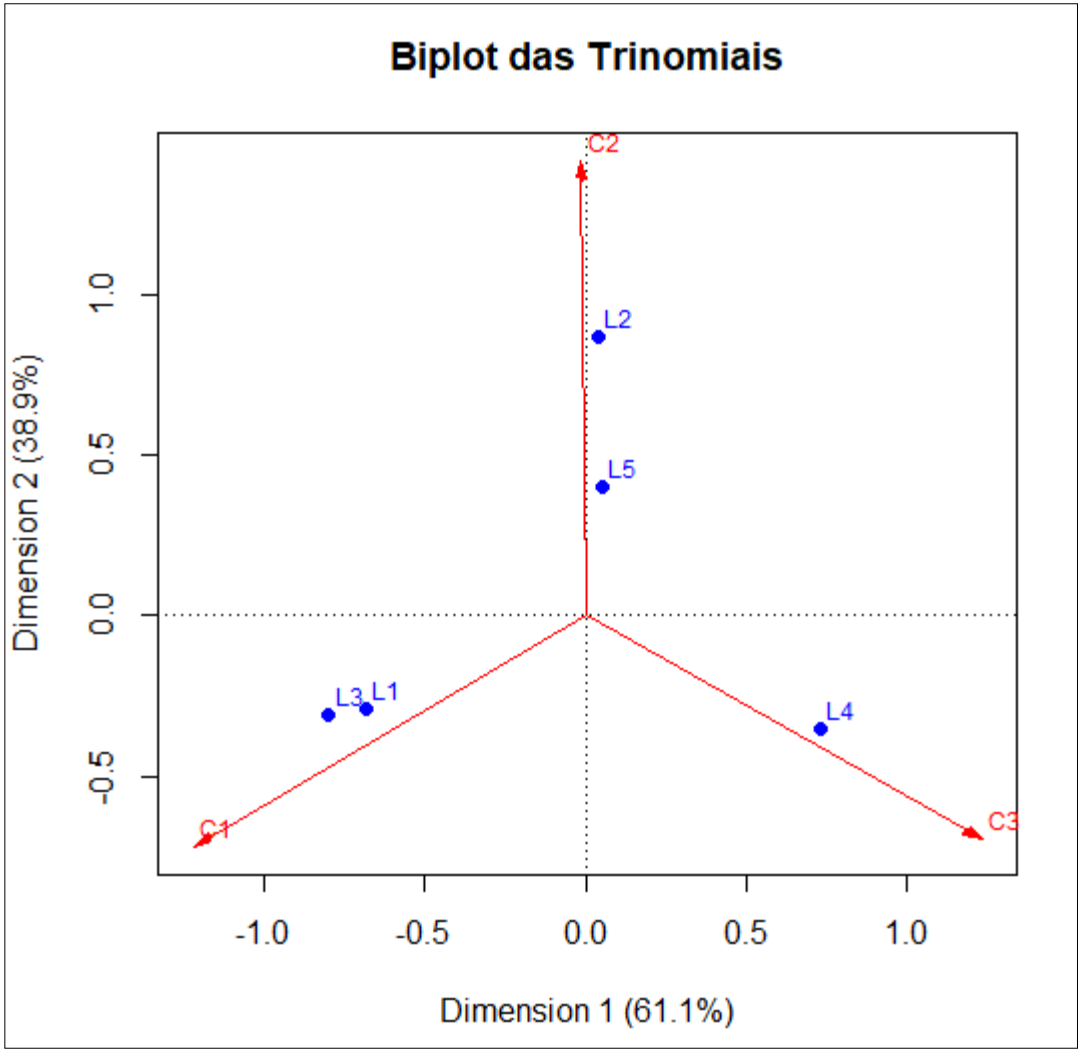
	C1	C2	C3
L1	0.68	0.20	0.12
L2	0.11	0.74	0.15
L3	0.73	0.19	0.08
L4	0.12	0.16	0.71
L5	0.22	0.52	0.26





Sob a restrição $p_1+p_2+p_3=1$, as Trinomiais podem ser representadas no plano $\Re^2$ (simplex), imposto por esta restrição, sem qualquer perda de informação. As categorias de resposta (C1, C2 e C3) definem os eixos (em coordenadas padrão).

Representação biplot das populações Trinomiais (L1 a L5)!



Este biplot representa um mapa assimétrico em que as linhas da tabela (trinomiais) estão em coordenadas principais e as colunas (classes de resposta da trinomial) estão em coordenadas padrão (eixos: (1,0,0),(0,1,0),(0,0,1))

Biplot (`map=rowprincipal`)

Representação ideal para tabelas com totais Linha fixos, em que a distribuição dos pontos (neste caso trinomiais) corresponde à informação da estatística Qui-Quadrado de homegeneidade entre multinomiais!!

Análise de Correspondência
Representação Simplex

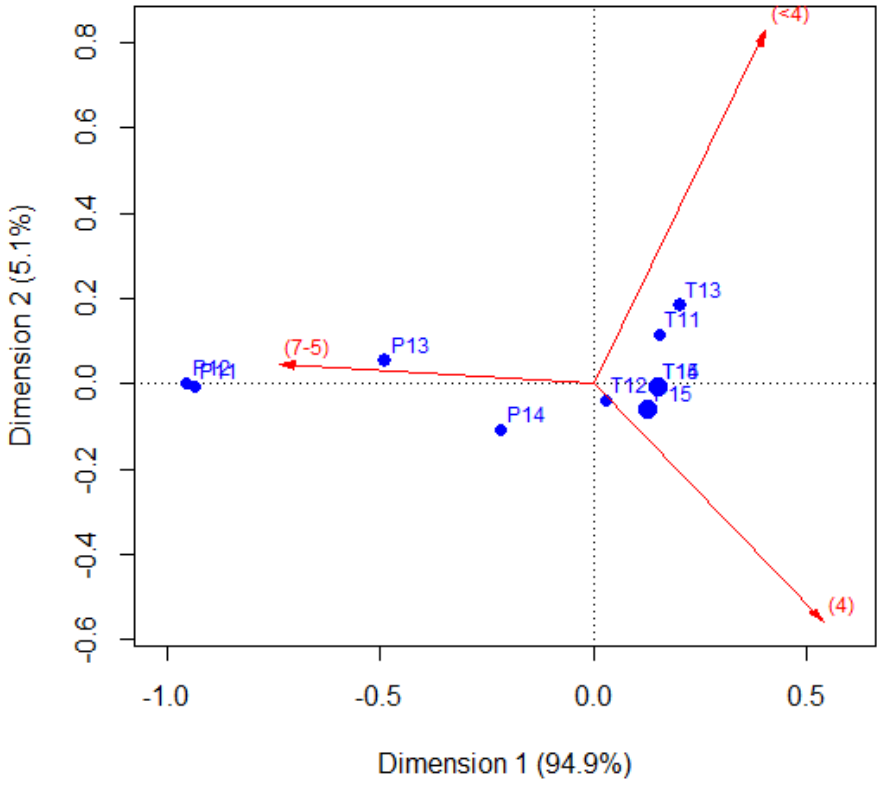
Exemplo 2: Distribuição do número de bulbilhos de alho de acordo com o tamanho (7-5, 4 e <4), tratamento e ano de plantio.

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

Como obter representações gráficas (Biplot) para casos gerais?

	Ano					
Jornal	**1976**	**1977**	**1978**	**1979**	**1980**	**Total**
A	64	58	67	59	60	308
B	18	18	23	20	17	96
C	12	10	9	12	9	52
D	36	25	34	31	27	153
E	29	21	25	20	20	115
F	133	115	116	107	89	560
G	34	28	30	26	29	147
H	178	143	180	150	148	799
I	8	8	5	6	6	33
J	101	113	143	112	107	576
K	66	56	60	58	53	293
L	87	69	79	68	69	372
M	23	19	17	19	17	95
N	34	24	29	26	23	136
O	70	56	60	55	50	291
P	29	20	25	19	18	111
Q	46	40	38	38	33	195
R	123	122	149	122	112	628
S	79	68	70	61	57	335
T	130	109	148	110	100	597
U	22	17	19	15	16	89
Total	1322	1139	1326	1134	1060	5981

Ao longo de 5 anos, em cada ano, cerca de 1000 pessoas de uma cidade foram amostradas e questionadas sobre quais jornais, dentre 21, eles liam regularmente.

Como representar o hábito de leitura de jornais dos cidadãos e sua variação ao longo do tempo?

Análise de Correspondência

Distribuição de 5.387 estudantes escoceses de acordo com a cor dos olhos e dos cabelos (Fisher, 1940)

	Cor do cabelo					
Cor olhos	**Claro**	**Ruivo**	**Médio**	**Escuro**	**Preto**	**Total**
Claros	688	116	584	188	4	1580
Azul	326	38	241	110	3	718
Médio	343	84	909	412	26	1774
Escuro	98	48	403	681	85	1315
Total	1455	286	2137	1391	118	5387

Como descrever graficamente o padrão de associação entre as variáveis cor dos olhos e dos cabelos dos estudantes escoceses ?

Análise de Correspondência

Distribuição dos funcionários de uma empresa (n=193) de acordo com o tabagismo (F0:não fuma, F1:fuma pouco, F2:fuma moderado e F3:fuma muito) e nível funcional (N1, N2, N3, N4 e N5).

	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

Para aderir a uma campanha nacional anti-tabagismo, o gerente de Recursos Humanos de uma empresa deseja conhecer o hábito de fumar dos funcionários. Os dados acima foram coletados para esta finalidade.

A representação gráfica dos dados é, em geral, de fácil entendimento. Como representar o padrão de associação entre o nível do funcionário e o hábito de fumar em um gráfico ?

Análise de Correspondência

(Everitt, 2004)

- A AC em tabelas de contingência é um método de decomposição da estatística Qui-Quadrado em componentes que correspondem a “eixos principais” que mais explicam a heterogeneidade entre as variáveis coluna (ou linhas) da tabela.
- Método que simultaneamente atribui uma escala às linhas e, separadamente, uma escala às colunas da tabela de tal forma a maximizar a correlação entre as duas escalas.
- Método de obtenção de coordenadas para representar as categorias das variáveis linha e das variáveis coluna da tabela. O padrão de associação entre as variáveis fica representado graficamente $\Rightarrow$ é uma análise de Escalonamento Multidimensional para uma medida de distância específica para dados categorizados, conhecida como distância Qui-Quadrado.

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 $D^L(D^C)$:

0.07	87.80%
0.01	11.76%
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

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)

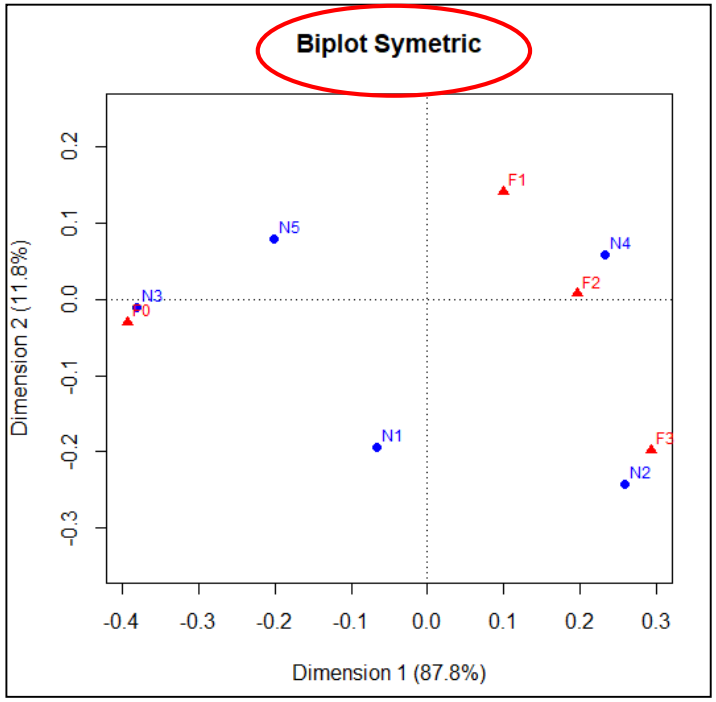
autovetor*sqrt(autovalor)

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)

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

χ^2=16.442 p=0.1718

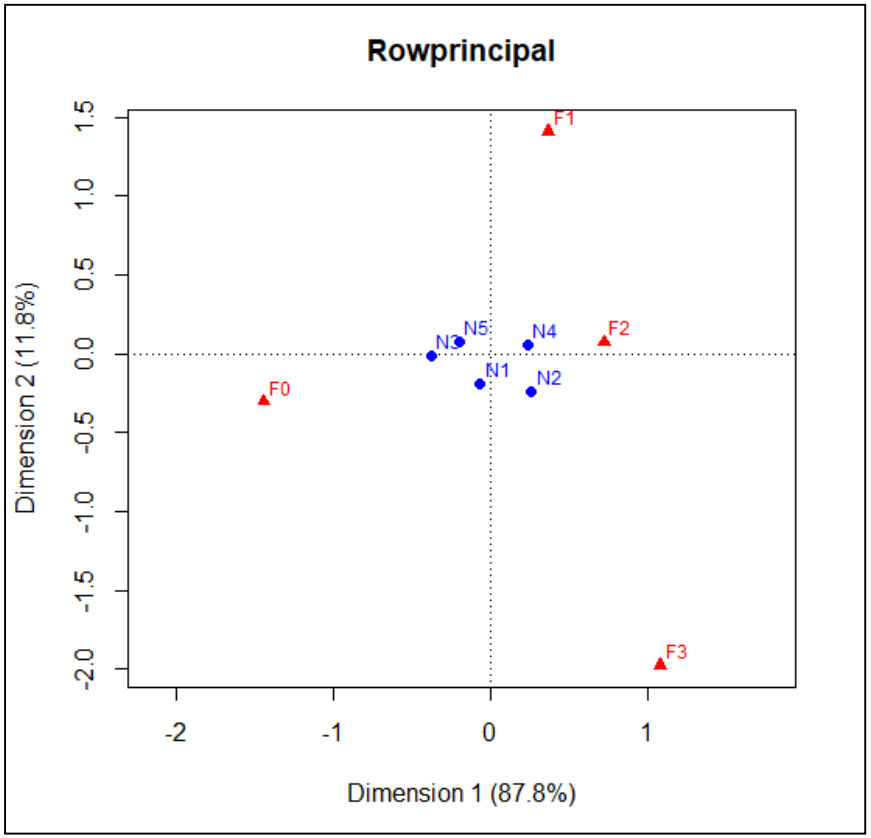
Coord. Principais de D^L:

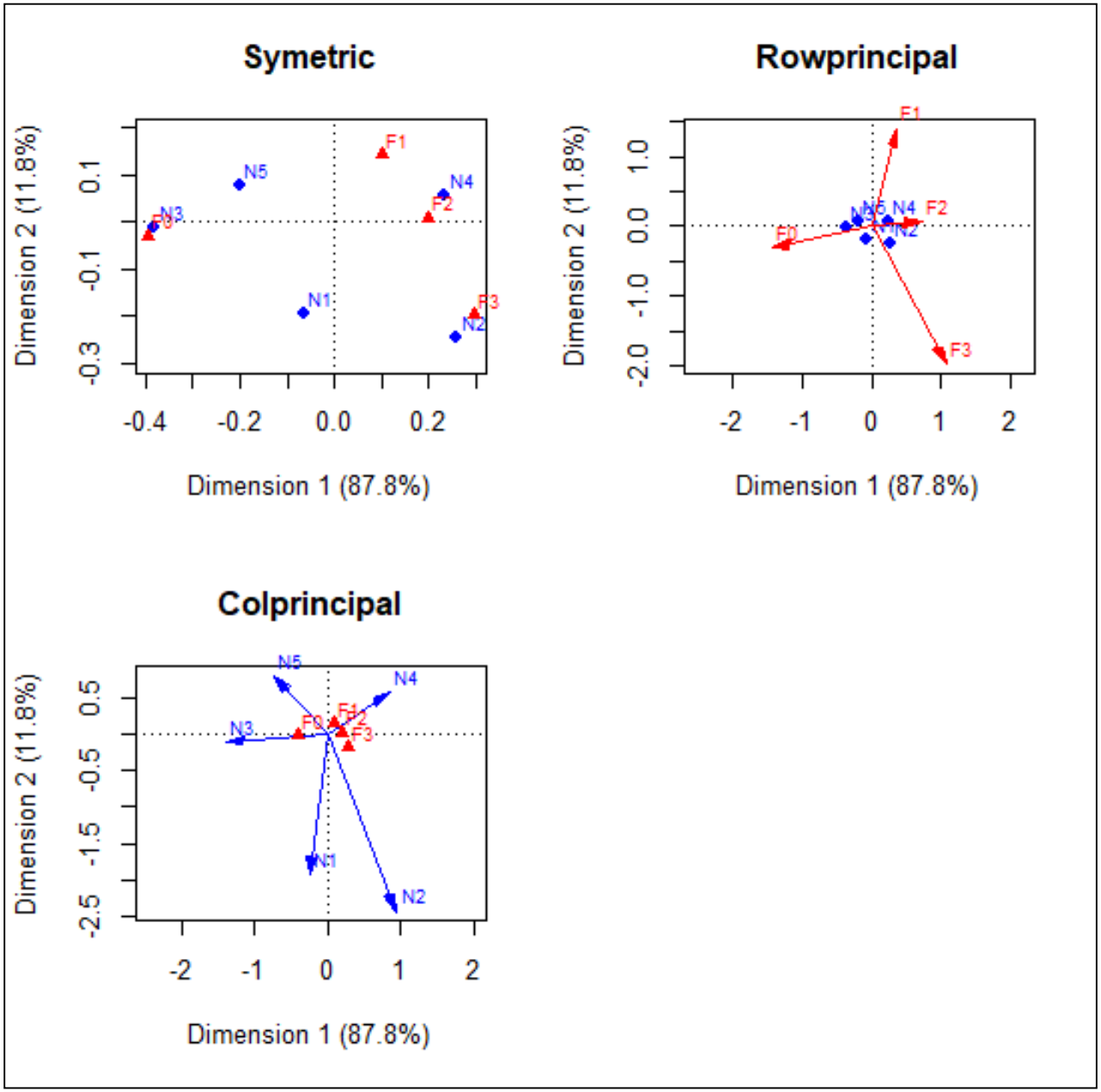
	Dim1	Dim2
N1	-0.07	-0.19
N2	0.26	-0.24
N3	-0.38	-0.01
N4	0.23	0.06
N5	-0.20	0.08

autovetor

Coordenadas Padrão de D^C:

	Dim1	Dim2	
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




Symetric
Dimension 2 (11.8%)
0.1
-0.1
-0.3
N5
F1
N4
N3
F2
F0
N1
F3
N2
-0.4
-0.2
0.0
0.2
Dimension 1 (87.8%)
Rowprincipal
Dimension 2 (11.8%)
1.0
0.0
-1.0
-2.0
F1
N5
N4
F2
F0
N2
F3
-2
-1
0
1
2
Dimension 1 (87.8%)
Colprincipal
Dimension 2 (11.8%)
0.5
-0.5
-1.5
-2.5
N5
N4
N3
F0
F1
F2
F3
N1
N2
-2
-1
0
1
2
Dimension 1 (87.8%)

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

χ^2=16.442 p=0.1718

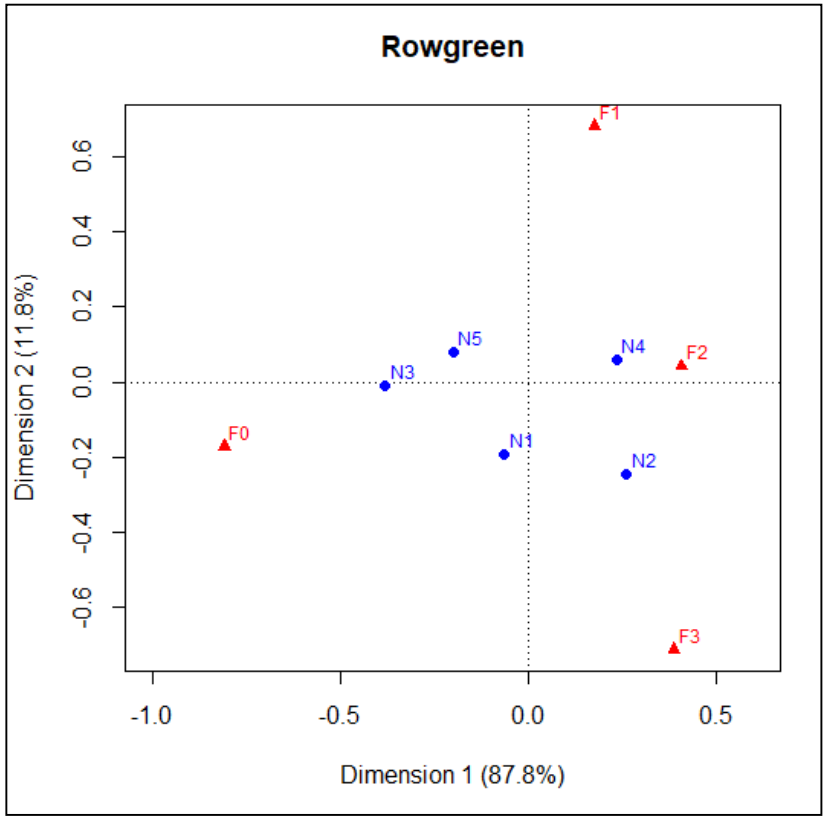
Coord. Principais de D^L:

	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

autovetor*sqrt(massa)

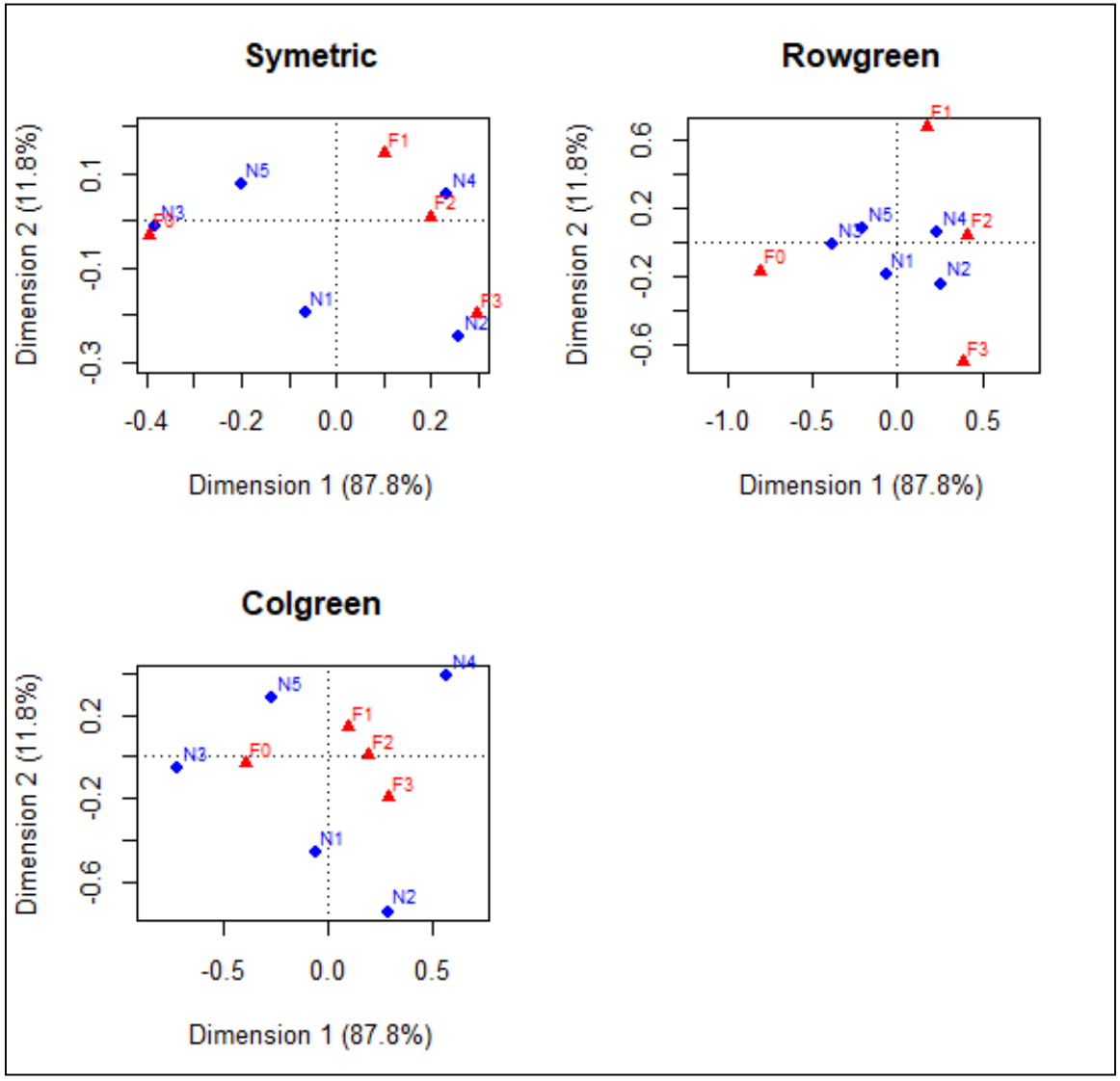
Coordenadas Padronizada de D^C:

	Dim1	Dim2
F0	-0.81	-0.17
F1	0.18	0.68
F2	0.41	0.04
F3	0.39	-0.71



Análise de Correspondência - Escalonamento Multidimensional

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



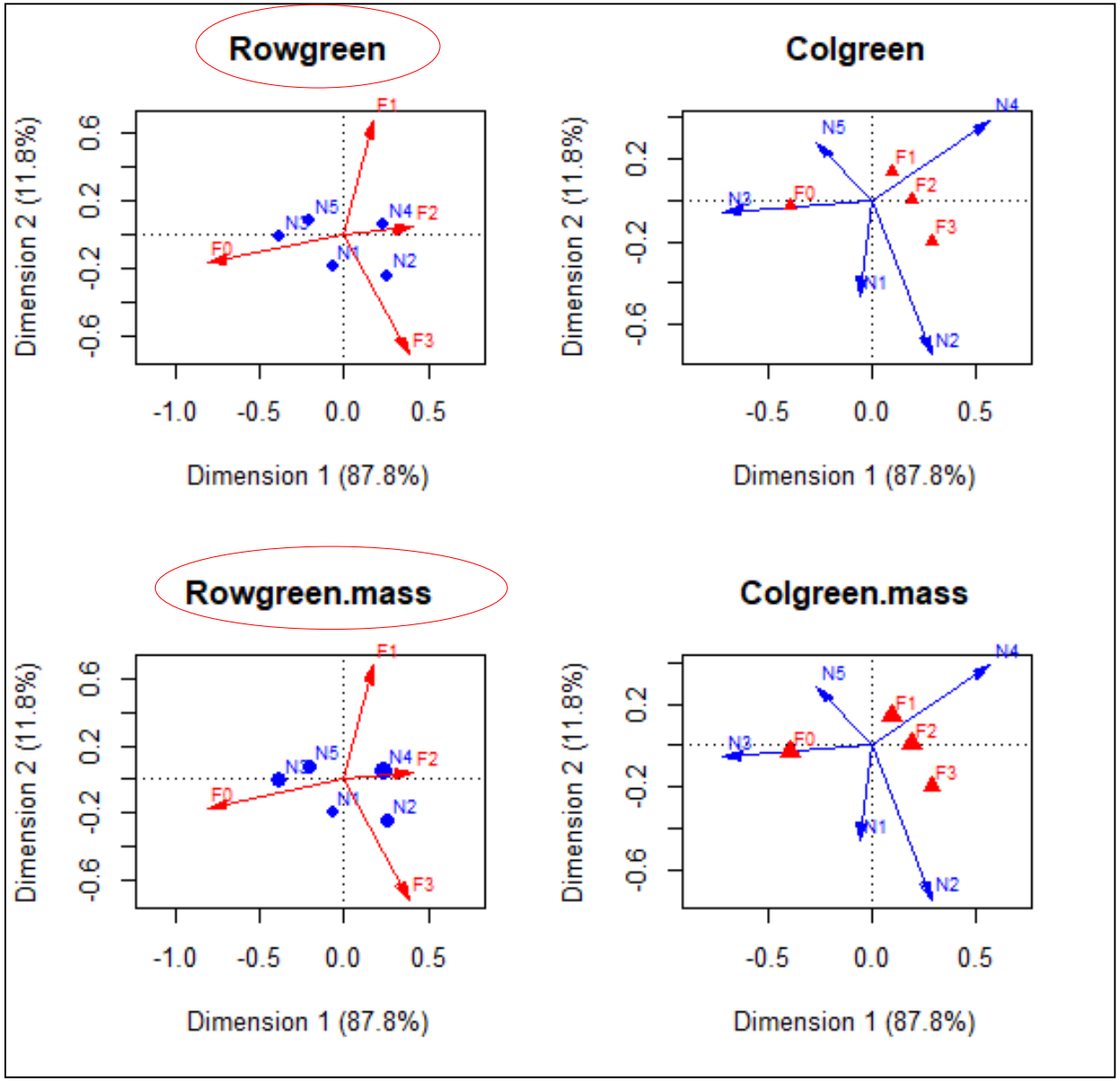
Representação BiPlot: existem diferentes construções do BiPlot, visando diferentes padronizações dos eixos.

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



A representação com os eixos auxilia no entendimento das classes de respostas das multinomiais.

Também é possível indicar as massas (tamanho amostral relative das multinomiais)

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

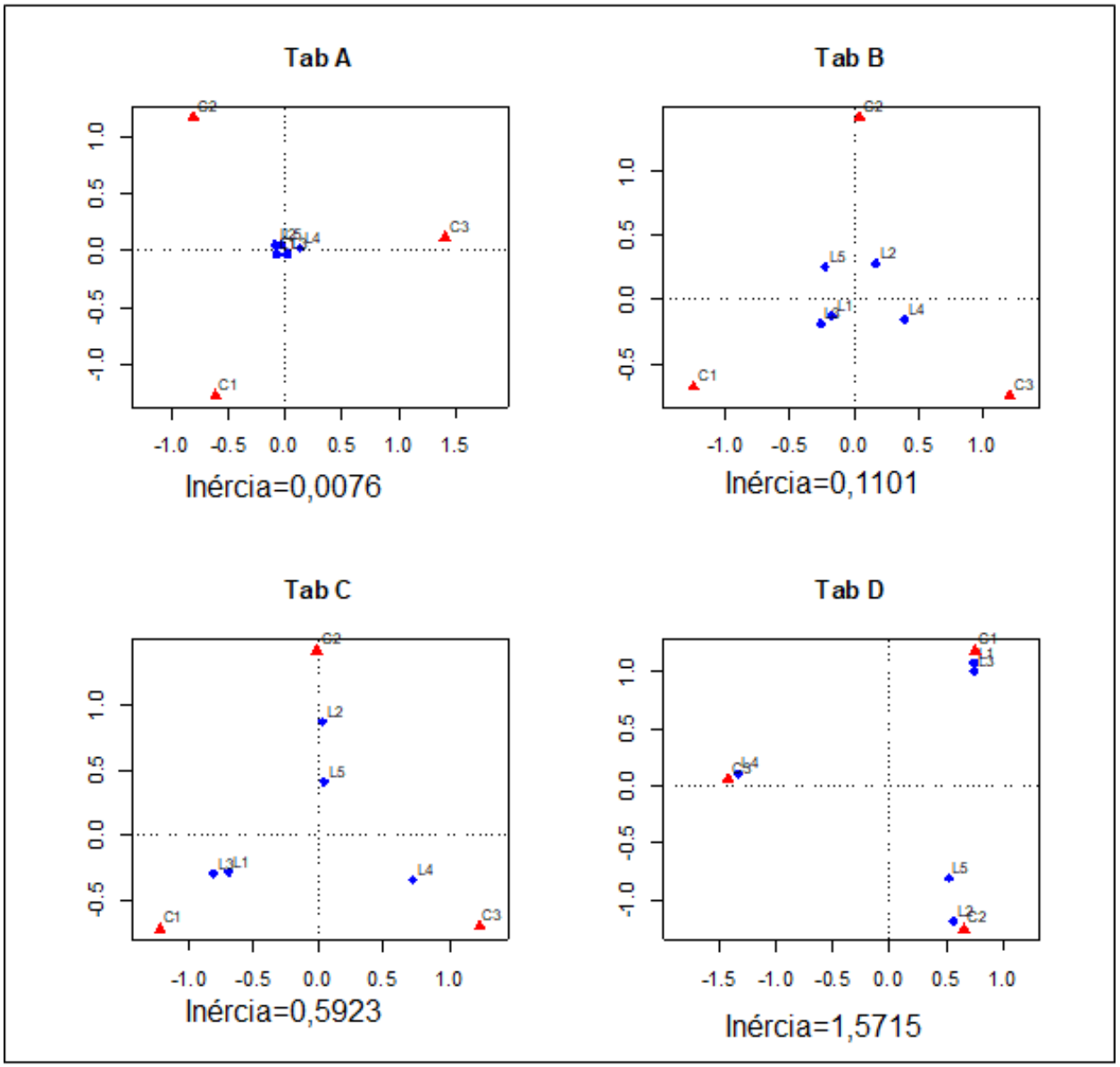
	C1	C2	C3
L1	20	1	0
L2	0	24	1
L3	24	2	0
L4	2	0	47
L5	4	23	2

	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)

Análise de Correspondência e Escalonamento Multidimensional

	Variável Coluna					
Variável Linha	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

• Em um Mapa Assimétrico (Rowprincipal ou Colprincipal) em que somente uma das variáveis (Linha ou Coluna) está representada em coordenadas principais (a outra está em coordenadas padrão), as distâncias entre os pontos são distâncias Euclidianas.

• MAS, em um gráfico Simétrico em que ambos os espaços (linha e coluna) estão representados simultaneamente, é preciso ter cuidado com a comparação entre categorias das linhas e colunas pois, neste caso, a medida de distância Euclidiana entre pontos pode não ser válida $\Rightarrow$ há assim diferentes propostas de padronização dos pontos (Mapas Rowgreen, Colgreen)