

MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler
pavan@ime.usp.br

1º Sem/2022 - IME

Análise Multivariada

$$Y_{n \times p} = (Y_{ij}) \in \mathbb{R}^{n \times p}$$

Já vimos 😊


- ✓ Estatísticas descritivas multivariadas, Episódios de Concentração, Boxplot Bivariado
- ✓ Distribuição N_p , Distribuições Amostrais (T^2 e W_p)
- Decomposições: SS_T e $Y_{n \times p}$
- ✓ $N_p(\mu_g; \Sigma_g)$: Inferências sobre μ_g (T^2 , MANOVA, ICS, Correções para Múltiplos testes)



Técnicas Multivariadas:

Redução de dimensionalidade
Vizualização de dados

- ✓ Análise de Componentes Principais (CP) $Y_{n \times p} \Rightarrow S_{p \times p}, R_{p \times p}$
- ✓ Escalonamento Multidimensional (CoP) $Y_{n \times p} \Rightarrow D_{n \times n}$
- ✓ Análise de Correspondência $Y_{n \times p} \Rightarrow [0, 1]^{I \times J}$

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

Análise Multivariada

REVISANDO ☺

Dados (hipotéticos): Variáveis avaliadas em atletas

V0 (no basal), V1 e V2 (antes e depois de uma intervenção) e o Genótipo de um Gene $n = 95$ $p = (5+1)$

Dados

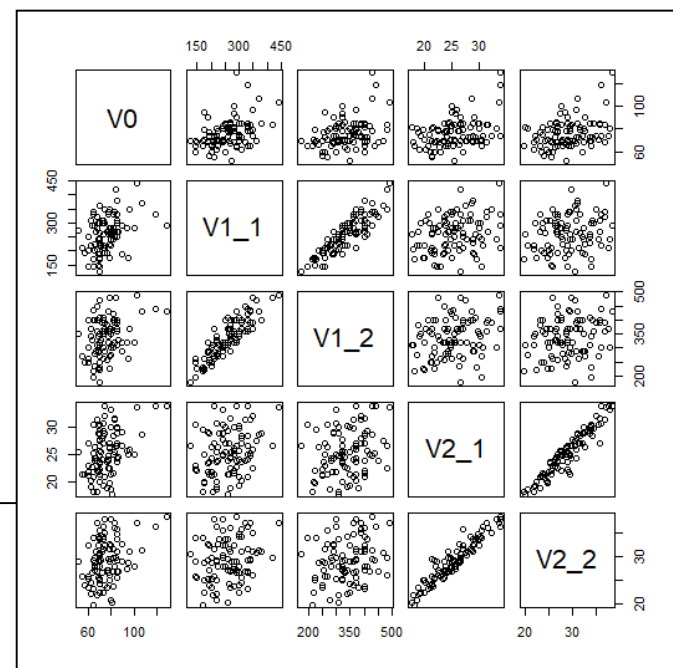
	V0	Gene	V1_1	V1_2	V2_1	V2_2
1	73.0	1	210	260	27.133	31.398
2	78.0	1	260	320	23.841	26.950
3	70.0	3	220	320	23.755	28.937
...						
94	95.0	2	175	260	25.120	28.605
95	71.0	2	220	330	25.452	29.029

Vetor centróide

	V0	V1_1	V1_2	V2_1	V2_2
	75.97	261.79	337.58	25.48	29.13

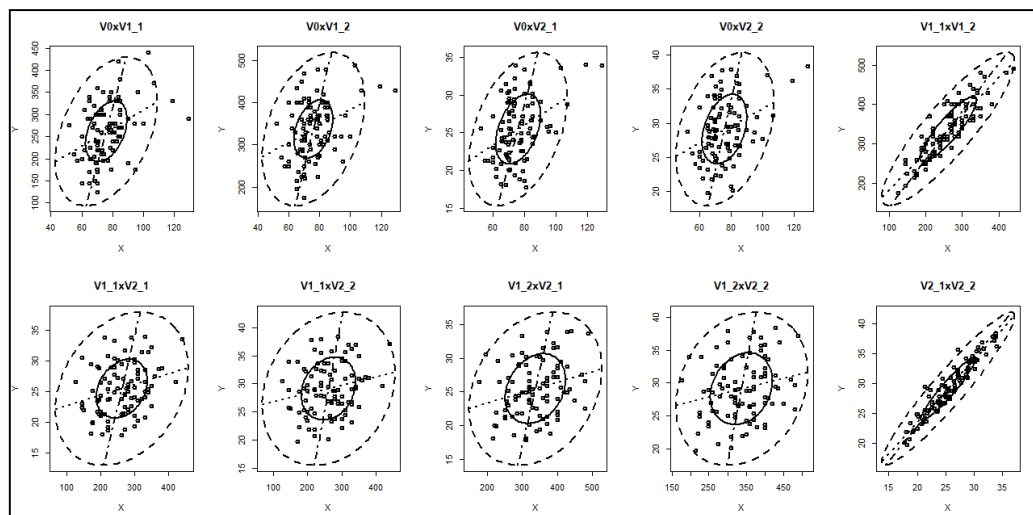
Matriz de Covariância (superior) e Correlação (inferior)

	V0	V1_1	V1_2	V2_1	V2_2
V0	155.24	317.53	303.47	20.80	19.37
V1_1	0.40	4032.93	3815.02	70.70	59.41
V1_2	0.36	0.88	4640.35	71.05	65.73
V2_1	0.41	0.27	0.26	16.40	17.10
V2_2	0.35	0.21	0.22	0.95	19.85

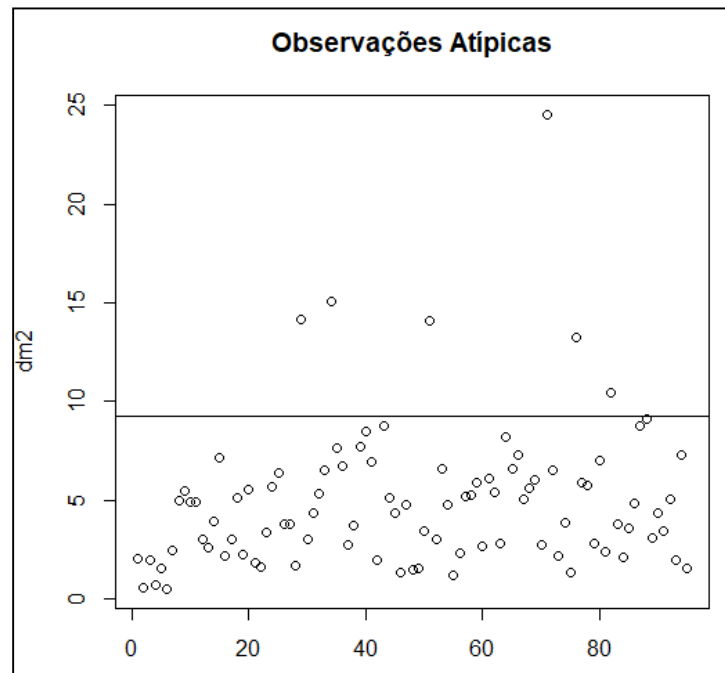


Análise Multivariada – Observações Atípicas

Box-plot bivariado: observações atípicas (p=2)



Distância de Mahalanobis (p=5)



$$\chi^2(p=5, 90\%) = 9.24$$

A distância de Mahalanobis é invariante por padronização dos dados!

Escore Z (Atletas atípicos Segundo dm2)

	dm2	V0	V1_1	V1_2	V2_1	V2_2
82	10.42	2.17	2.81	2.24	2.01	1.80
76	13.25	-0.14	-0.82	0.48	2.08	1.95
51	14.06	0.73	0.76	0.77	0.41	1.48
29	14.16	0.32	-0.03	0.48	-1.03	0.07
34	15.06	3.45	1.07	1.50	2.11	1.60
71	24.54	4.26	0.44	1.36	2.07	2.08






Análise Multivariada – Observações e Variáveis

Exemplo Hipotético

Indivíduos

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95					

Variáveis

V0	V1_1	V1_2	V2_1	V2_2
				

Componentes Principais $S_{5 \times 5}$

Exemplo Hipotético: Análise dos Dados originais (S)

Autovalores (importância dos CP):

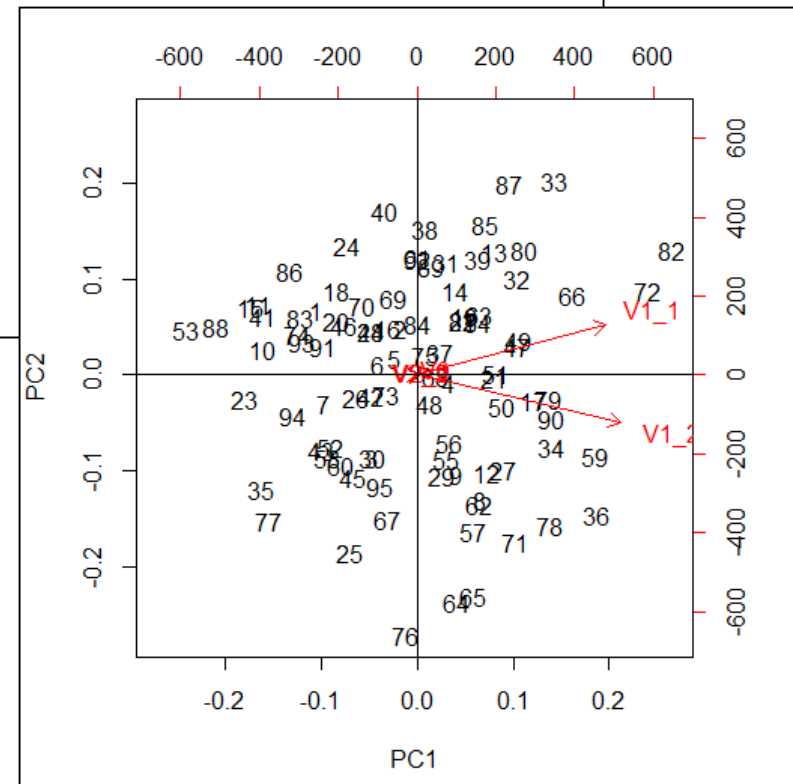
	PC1	PC2	PC3	PC4	PC5
Standard deviation	90.4980	22.61788	11.56229	5.36349	0.9289
Proportion of Variance	0.9239	0.05771	0.01508	0.00325	0.0001
Cumulative Proportion	0.9239	0.98158	0.99666	0.99990	1.0000

Autovetores (cargas)

	PC1	PC2	PC3	PC4	PC5
V0	0.05	0.07	0.97	-0.20	0.01
V1_1	0.68	0.73	-0.09	0.00	0.01
V1_2	0.73	-0.68	0.01	-0.01	0.00
V2_1	0.01	0.01	0.14	0.64	-0.75
V2_2	0.01	0.00	0.14	0.74	0.66

Biplot: representação simultânea dos atletas e das variáveis

As variáveis V1_1 e V1_2 são as que mais contribuem para a redução de dimensionalidade.



Componentes Principais

$R_{5 \times 5}$

Exemplo Hipotético: Análise dos Dados Normalizados (R)

Autovalores (importância dos CP):

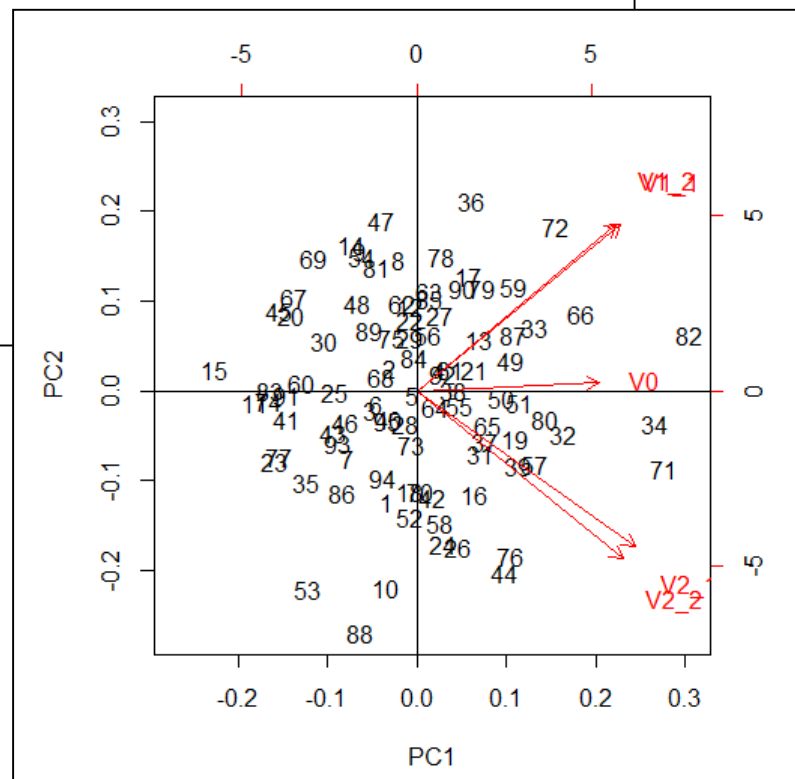
	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.6524	1.1983	0.8173	0.34393	0.21785
Proportion of Variance	0.5461	0.2872	0.1336	0.02366	0.00949
Cumulative Proportion	0.5461	0.8333	0.9668	0.99051	1.00000

Autovetores (cargas)

	PC1	PC2	PC3	PC4	PC5
V0	0.40	0.02	0.91	-0.05	0.04
V1_1	0.45	0.51	-0.18	0.70	0.12
V1_2	0.44	0.51	-0.25	-0.69	-0.09
V2_1	0.48	-0.47	-0.16	0.11	-0.71
V2_2	0.46	-0.51	-0.22	-0.10	0.69

Biplot: representação dos atletas e das variáveis

Todas as variáveis contribuem para a redução de dimensionalidade



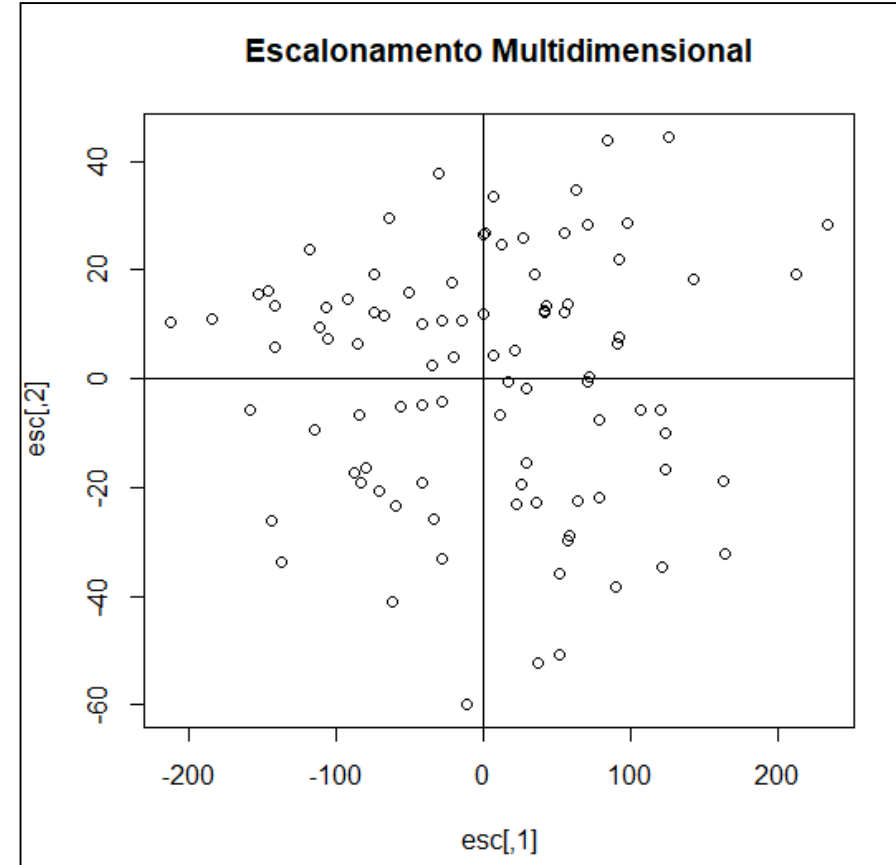
Coordenadas Principais—Escalonamento Multidimensional

Exemplo Hipotético: Matriz de Distância Euclidiana ($De_{95 \times 95}$)

de	1	2	3	4	...	95
1	0.00	78.46	61.04	122.22		70.80
2	78.46	0.00	40.84	44.83		41.90
3	61.04	40.84	0.00	72.65		10.19
4	122.22	44.83	72.65	0.00		67.51
...						
95	70.80	41.90	10.19	67.51		0.00

Coordenadas Principais

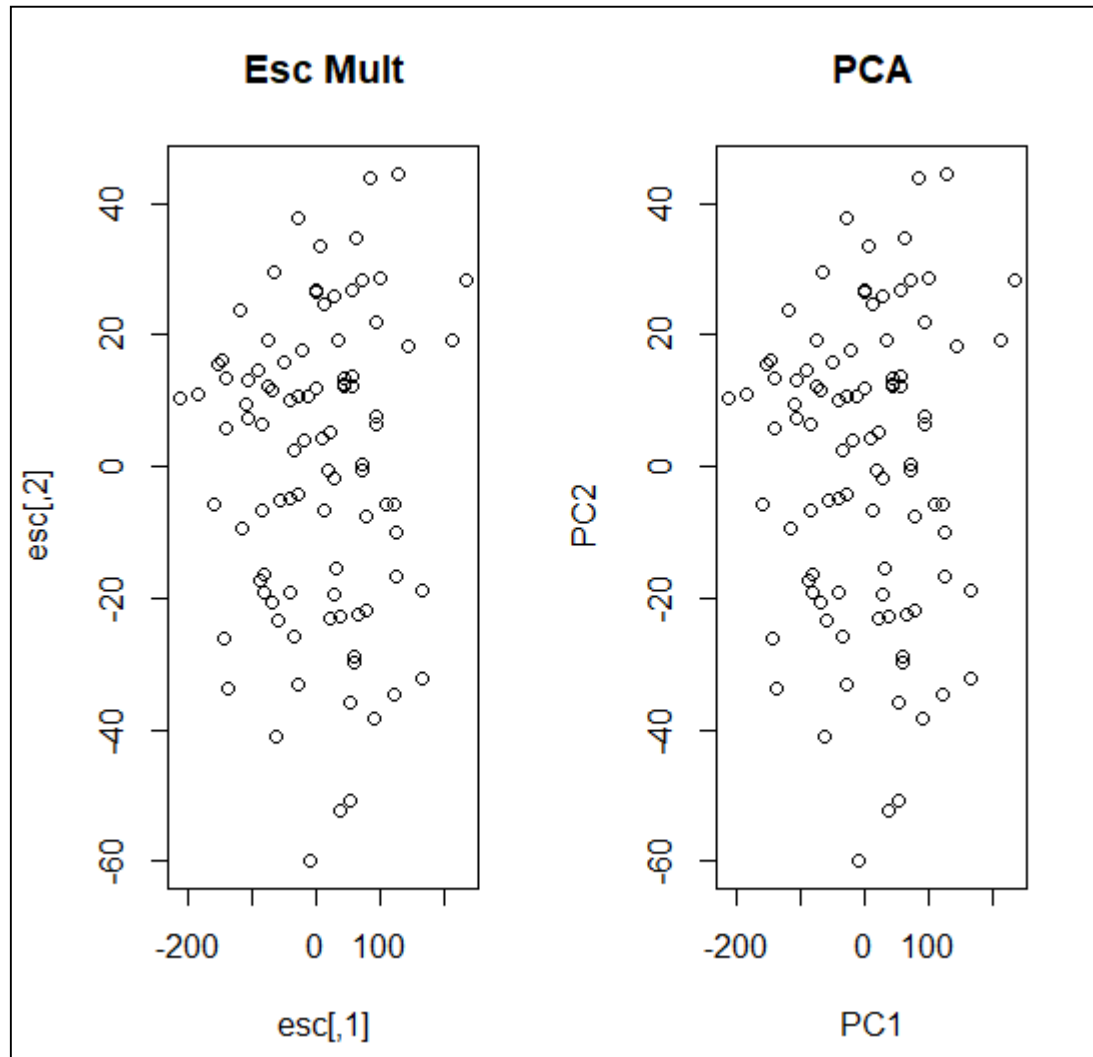
	[,1]	[,2]
1	-92.09	14.75
2	-14.04	10.77
3	-41.55	-18.99
4	28.92	-1.76
...		
94	-114.66	-9.23
95	-34.14	-25.70



%Expl: 0.9815757

Coordenadas Principais–Componentes Principais

Exemplo Hipotético – Análise dos Dados Originais



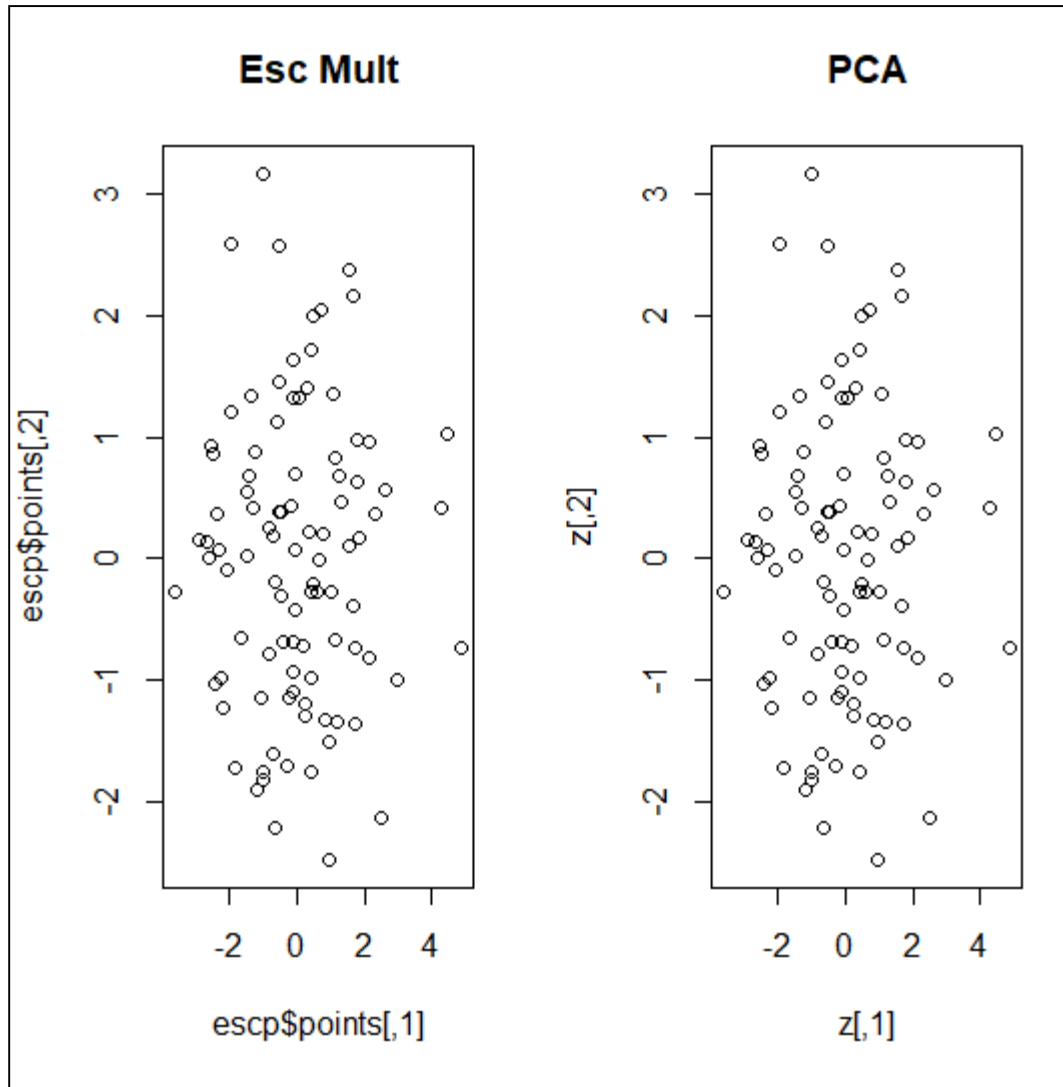
Análises Equivalentes
(Solução Métrica)
Mesma representação

Coordenadas Principais (nxn):
Matriz de Distância: **D**_{95x95}

Componentes Principais (pxp):
Matriz de Distância: **S**_{5x5}

Coordenadas Principais–Componentes Principais

Exemplo Hipotético – Análise dos Dados Padronizados



Análises Equivalentes
(Solução Métrica)
Mesma representação

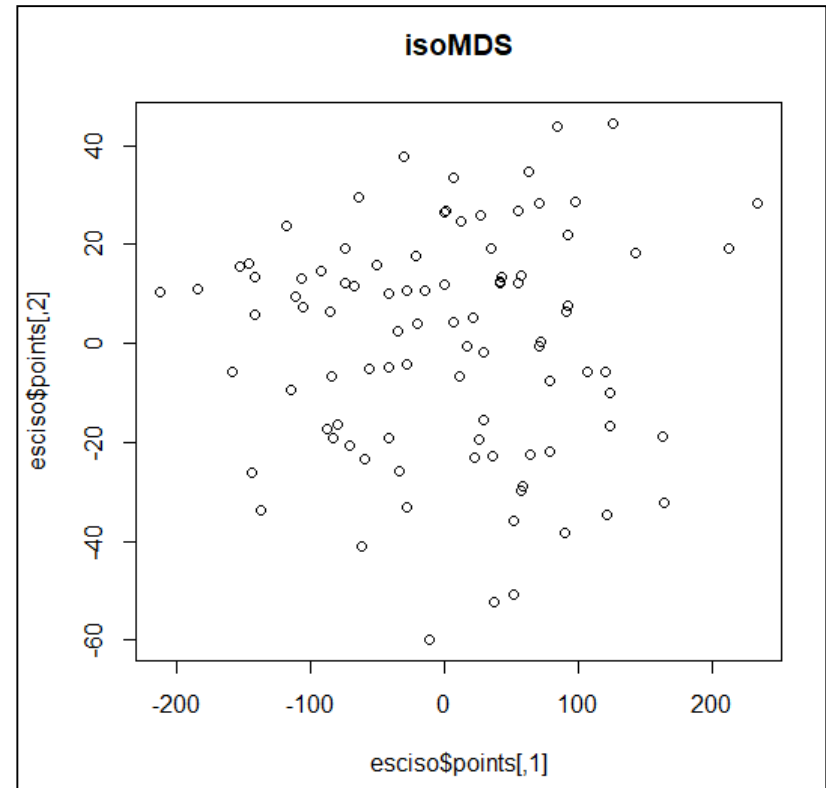
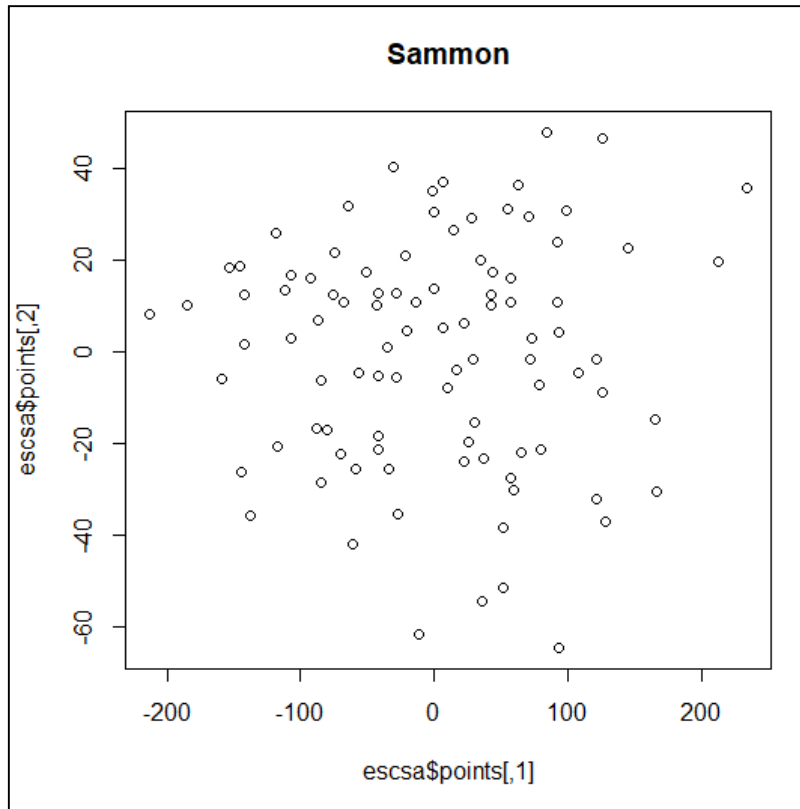
Coordenadas Principais (nxn):
Matriz de Distância: $\mathbf{D}_{95 \times 95}$

Componentes Principais (pxp):
Matriz de Distância: $\mathbf{R}_{5 \times 5}$

Coordenadas Principais–Escalonamento Multidimensional

Exemplo Hipotético – Análise dos Dados Originais

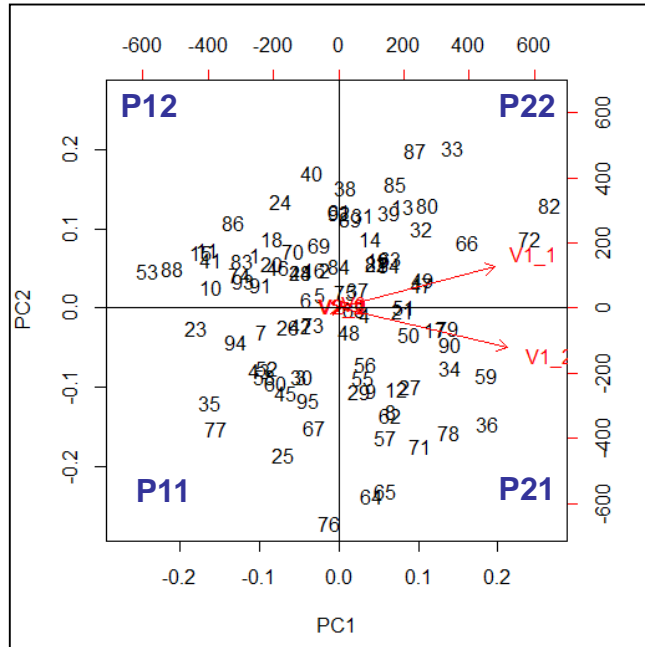
Soluções Não-Métricas



Análise de Correspondência

Representação gráfica das variáveis em uma tabela de contingência

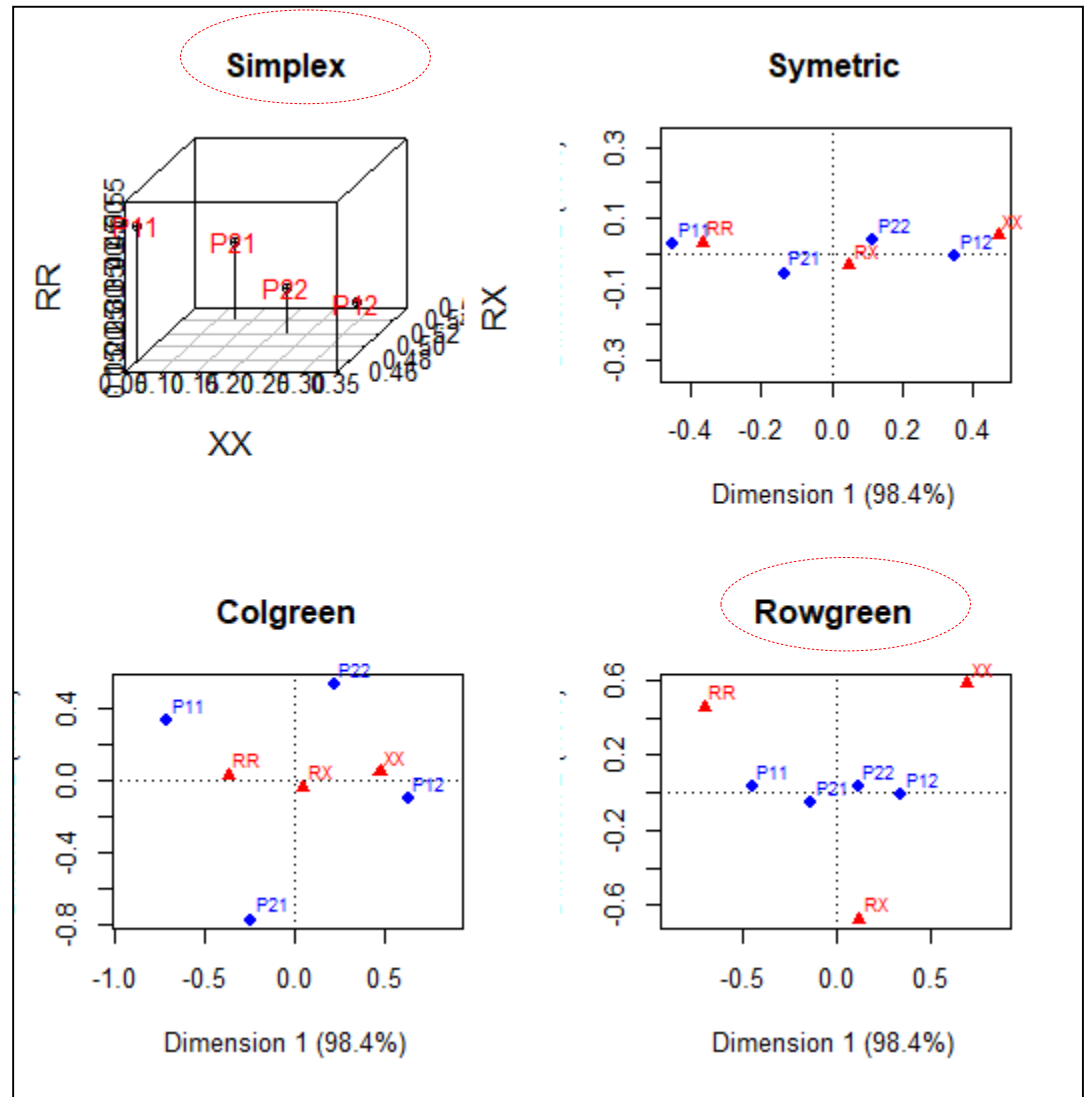
Considere o Biplot (de S) e a Categorização das observações em cada quadrante de acordo com o Genótipo do Gene sob estudo



Distribuição dos atletas

1=XX 2=RX 3=RR

P11	1	9	9
P12	7	14	4
P21	3	13	8
P22	6	14	7

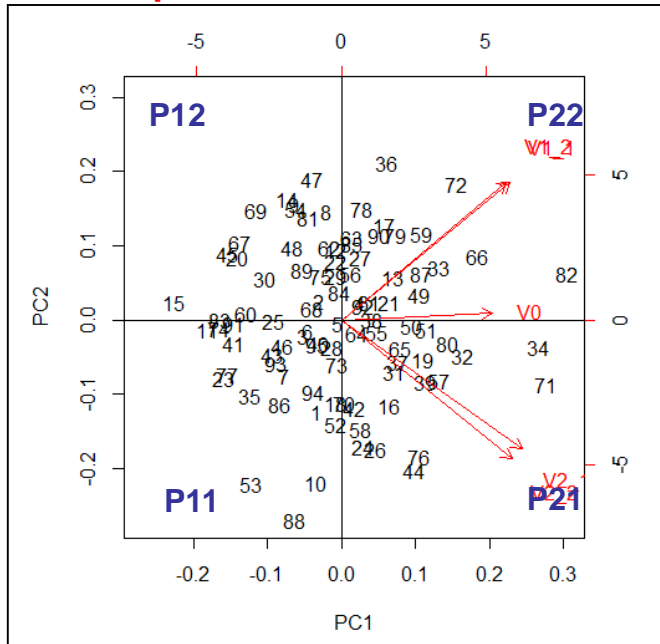


X-squared = 7.8021, df = 6, p-value = 0.253

Análise de Correspondência

Representação gráfica das variáveis em uma tabela de contingência

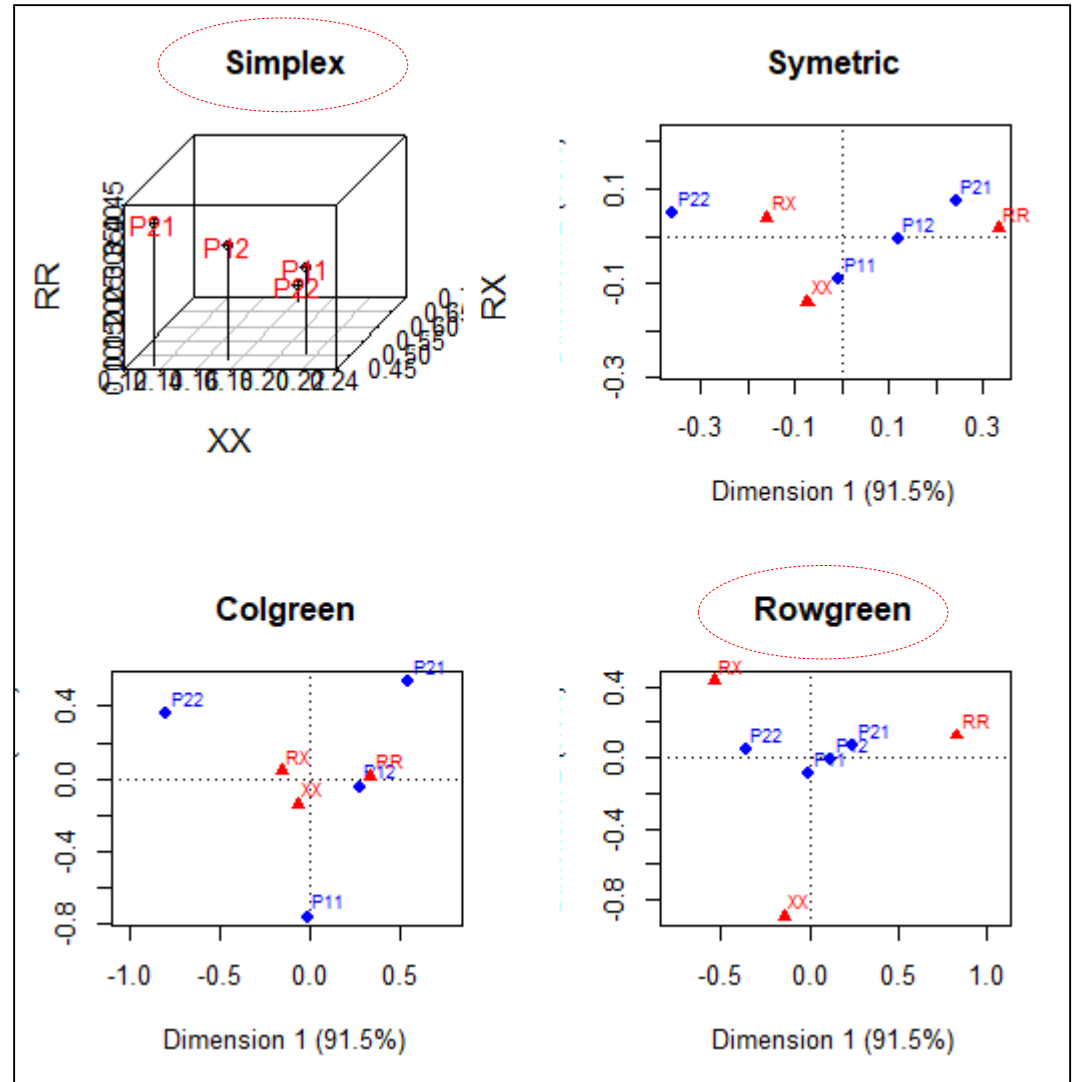
Considere o Biplot (de R) e a Categorização das observações em cada quadrante de acordo com o Genótipo do Gene sob estudo



Distribuição dos atletas

1=XX 2=RX 3=RR

P11	6	14	8
P12	4	11	8
P21	3	10	9
P22	4	15	3



X-squared=4.86, df=6, p-value=0.5616

Análise de Fatores (Análise Fatorial)

(Análise de Fatores Comuns e Específicos)

Revisando

INTRODUZINDO

Análise de CP e Análise Fatorial (Exploratória)

Como obter as variáveis originais Y a partir das componentes principais Z ?

$$Y_{n \times p} ; Y_i \stackrel{iid}{\sim} (\mu; \Sigma); \Sigma_{p \times p} = V \Lambda V'; V = (v_{jk}); V' = (v_{kj})$$

$$Z_{ik} = V_k' Y_i = v_{1k} Y_{i1} + v_{2k} Y_{i2} + \dots + v_{pk} Y_{ip} \quad \begin{array}{l} \text{k-ésimo CP} \\ \text{(k-ésima coluna de V)} \end{array}$$

$$V_{p \times p} = \begin{matrix} \text{autovetores} \\ \begin{matrix} V_1 \downarrow & V_2 & & V_p \\ \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1p} \\ v_{21} & v_{22} & \dots & v_{2p} \\ & & \dots & \\ v_{p1} & v_{p2} & \dots & v_{pp} \end{pmatrix} \end{matrix} \end{matrix}$$

$$Z_{i \times p} = V' Y_i \stackrel{=1}{\Leftrightarrow} V Z_i = V V' Y_i = Y_i \Rightarrow Y_{i \times p} = V Z_i$$

$$Y_{ij} = v_{j1} Z_{i1} + v_{j2} Z_{i2} + \dots + v_{jp} Z_{ip} \quad \text{j-ésima linha de V}$$

$$Y_{ij} \cong \underbrace{v_{j1} \sqrt{\lambda_1}}_{F_1} \underbrace{\frac{Z_{i1}}{\sqrt{\lambda_1}}}_{F_1} + v_{j2} \sqrt{\lambda_2} \underbrace{\frac{Z_{i2}}{\sqrt{\lambda_2}}}_{F_2} + \dots + v_{jp} \sqrt{\lambda_m} \underbrace{\frac{Z_{im}}{\sqrt{\lambda_m}}}_{F_m}$$

Sistema de equações das p variáveis Y_{ij} ($j=1,2,\dots,p$) descritas em função de m "fatores comuns" Z_{ik} ($k=1,2,\dots,m$; $m < p$). Este é um dos objetivos da **Análise Fatorial Exploratória**.

Análise de Fatores

$$Y_{n \times p} ; Y_i \stackrel{iid}{\sim} (\mu; \Sigma)$$

Unidades Amostrais	Variáveis					
	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}

Objetivos:

- Obter **Fatores Comuns** às “p” variáveis que expliquem a covariância entre elas. Isso equivale a **decompor a matriz de covariância (ou correlação)** tal que:

$$\Sigma \cong \text{fatores comuns} + \text{fatores específicos}$$



Análise Fatorial Exploratória \Rightarrow obter fatores comuns (constructos, variáveis latentes) e específicos na modelagem de Σ

Análise Fatorial Confirmatória \Rightarrow verificar se uma estrutura conhecida (diagrama de caminhos ou grafo) se ajusta aos dados (à matriz de correlações)

Motivação

Exemplo 1: É possível explicar o **desempenho escolar** por meio de 2 constructos (habilidades dos estudantes)?

	Geografia	Inglês	História	Aritmética	Algebra	Geometria
R =	1	0,439	0,41	0,288	0,329	0,248
	0,439	1	0,351	0,354	0,32	0,329
	0,41	0,351	1	0,164	0,19	0,181
	0,288	0,354	0,164	1	0,595	0,47
	0,329	0,32	0,19	0,595	1	0,464
	0,248	0,329	0,181	0,47	0,464	1

Exemplo 2: A **Síndrome Metabólica** é uma doença multifatorial caracterizada pela ocorrência de pelo menos três fatores de risco dentre hipertensão, alteração nos componentes do colesterol, hiperglicemia, alto IMC, elevada circunferência abdominal. Estas variáveis foram avaliadas em 80 pacientes. É possível obter variáveis latentes para mensurar esta doença multifatorial?

Notação!

Análise Fatorial Exploratória

Variáveis observadas Y podem ser modeladas em função de um conjunto de variáveis latentes F (não observáveis, constructos)?

$$Y_{i \times 1} \stackrel{iid}{\sim} (\mu; \Sigma) \Rightarrow Y_{ij} = \mu_j + \phi_{j1}F_{1i} + \phi_{j2}F_{2i} + \dots + \phi_{jm}F_{mi} + e_{ij}$$

Diferentemente de CP e CoP:

Modelo de equações estruturais é adotado aos dados Y :

$$\left\{ \begin{array}{l} Y_{i1} - \mu_1 = \phi_{11}F_{1i} + \phi_{12}F_{2i} + \dots + \phi_{1m}F_{mi} + e_{1i} \\ Y_{i2} - \mu_2 = \phi_{21}F_{1i} + \phi_{22}F_{2i} + \dots + \phi_{2m}F_{mi} + e_{2i} \\ \dots \\ Y_{ip} - \mu_p = \phi_{p1}F_{1i} + \phi_{p2}F_{2i} + \dots + \phi_{pm}F_{mi} + e_{pi} \end{array} \right.$$

Modelo de Fatores
Notação Matricial:

$$Y_i - \mu = \Phi_{p \times m} \mathbf{f}_{i \times m \times 1} + e_{i \times p \times 1}$$

$$\mathbb{R}^p \rightarrow \mathbb{R}^m$$

$\mathbf{f} = (F_1, \dots, F_m)'$: fatores comuns (var. latentes)

$e = (e_1, \dots, e_p)'$: fatores específicos (erros)

$\Phi = (\phi_{ij})$: cargas fatoriais

Como obter $\left\{ \begin{array}{l} \Phi_{p \times m} \\ \mathbf{f}_{i \times m \times 1} \end{array} \right. ?$

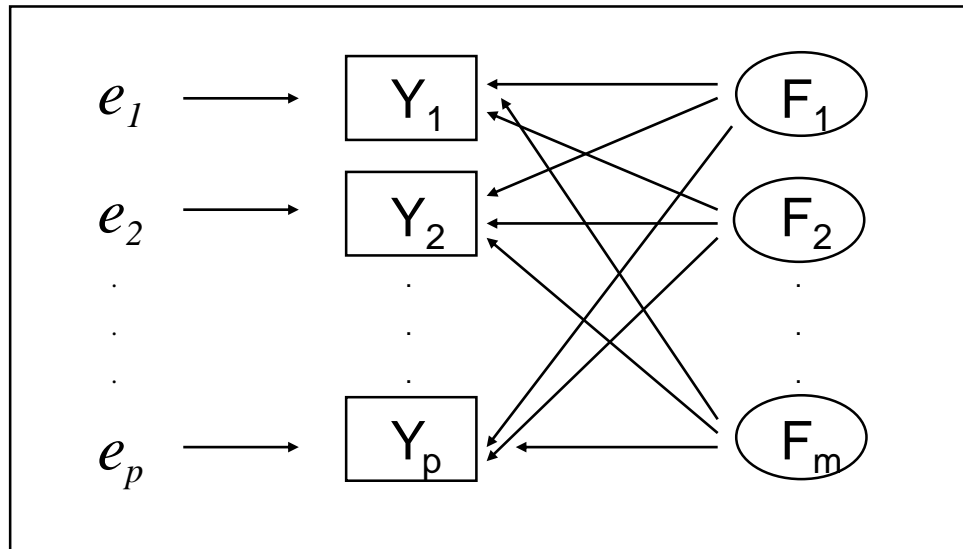
Pacote computacional:
factanal-R

Análise Fatorial Confirmatória

$$Y_i \Rightarrow Y_i - \mu = \Phi \mathbf{f}_i + e_i$$

Equações de mensuração

Diagrama de Caminhos (Grafo) de um modelo de Análise Fatorial ortogonal



Var. Observadas: retângulos

Var. Latentes (constructo): círculos

Erros: sem representação gráfica

As setas (direcionadas) partem de uma variável independente e atingem uma variável dependente

Se existirem correlações (entre os fatores comuns F ou entre os específicos e), estas devem ser representadas por arcos


Um particular Grafo (que define um Modelo de Equações Estruturais - MEE), se ajusta aos dados?

- MEE (Bollen, 1989): modelos gerais envolvendo equações estruturais para as variáveis observadas (equações de mensuração) bem como equações estruturais para as variáveis latentes

Pacotes computacionais:
LISREL, Lavaan-R, (fa-R)

Análise Fatorial


- Modelo estrutural: $Y_i - \mu = \Phi \mathbf{f}_i + e_i$
- Suposições do modelo fatorial (\mathbf{f} e e são variáveis aleatórias):



$$\mathbf{f}_{i_{m \times 1}} = \begin{bmatrix} F_{1i} \\ \dots \\ F_{mi} \end{bmatrix} \stackrel{iid}{\sim} (0; I_m); \quad e_{i_{p \times 1}} \stackrel{iid}{\sim} (0; \Psi = \text{diag}(\psi_1, \dots, \psi_p)); \quad \text{Cov}(\mathbf{f}, e) = 0$$

Matriz de Covariância (marginal) de Y:

Decomposição muito particular da matriz Σ !

$$\text{Cov}(Y_i) = \Sigma_{p \times p} = \text{Cov}(\Phi \mathbf{f}_i + e_i) \Rightarrow \boxed{\Sigma = \Phi \Phi' + \Psi}$$


componente de covariâncias devido ao fator comum
componente de variâncias devido ao fator específico (matriz diagonal)

$$\text{Var}(Y_{ij}) = \boxed{\phi_{j1}^2 + \phi_{j2}^2 + \dots + \phi_{jm}^2} + \psi_j = h_j^2 + \psi_j$$

comunalidade da variável Y_{ij}
especificidade de Y_{ij}

Análise Fatorial

Alguns resultados!

$$Y_i \in \mathbb{R}^p; \quad Y_i - \mu = \Phi \mathbf{f}_i + e_i \quad \Leftrightarrow \quad \Sigma = \Phi \Phi' + \Psi$$

$$Var(Y_{ij}) = \phi_{j1}^2 + \phi_{j2}^2 + \dots + \phi_{jm}^2 + \psi_j = h_j^2 + \psi_j$$

$$\bar{h}_j^2 = \frac{h_j^2}{Var(Y_{ij})}$$

% da $Var(Y_{ij})$ explicada
pelo conjunto dos m
fatores comuns

$$H^2 = \frac{\sum_{j=1}^p h_j^2}{\sum_{j=1}^p Var(Y_{ij})}$$

% da Variância Total de Y
explicada pelo conjunto dos
m fatores comuns

$$H_{F_k}^2 = \frac{\sum_{j=1}^p \phi_{jk}^2}{\sum_{j=1}^p Var(Y_{ij})}$$

% da Variância Total
explicada pelo fator
comum Fk

$$Cov(Y_{ij}, Y_{ij'}) = \phi_{j1}\phi_{j'1} + \phi_{j2}\phi_{j'2} + \dots + \phi_{jm}\phi_{j'm} \Rightarrow \text{depende somente de fatores comuns}$$

$$Cov(Y_{ij}, F_{ki}) = Cov(\phi_{j1}F_{1i} + \dots + \phi_{jk}F_{ki} + \dots + \phi_{jm}F_{mi}; F_{ki}) = Cov(\phi_{jk}F_{ki}; F_{ki}) = \phi_{jk}$$

$$Corr(Y_{ij}, F_{ki}) = \phi_{jk} / \sqrt{Var(Y_{ij})} = \phi_{jk} / \sqrt{h_j^2 + \psi_j}$$

Análise Fatorial Exploratória

$$Y_i \in \mathbb{R}^p; \quad Y_i - \mu = \Phi \mathbf{f}_i + e_i$$

$$\Leftrightarrow \Sigma = \Phi\Phi' + \Psi$$

$$\left\{ \begin{array}{l} Y_{1i} - \mu_1 = \phi_{11}F_{1i} + \phi_{12}F_{2i} + \dots + \phi_{1m}F_{mi} + e_{1i} \\ Y_{2i} - \mu_2 = \phi_{21}F_{1i} + \phi_{22}F_{2i} + \dots + \phi_{2m}F_{mi} + e_{2i} \\ \dots \\ Y_{pi} - \mu_p = \phi_{p1}F_{1i} + \phi_{p2}F_{2i} + \dots + \phi_{pm}F_{mi} + e_{pi} \end{array} \right.$$

Como obter:

- Matriz de Coeficientes ou Cargas (Φ)
- Componentes Específicos (Ψ)
- Escores Fatoriais ($\mathbf{f}=(F_{ki})$)

Soluções:

- Via Componentes Principais
- Via Máxima Verossimilhança

E ainda, para qualquer matriz orthogonal Γ , tem-se:

$$\Phi^* = \Phi_{p \times m} \Gamma_{m \times m}; \quad \Gamma\Gamma' = I_m \quad \Rightarrow \quad \Phi^* \Phi^{*'} + \Psi = \Phi\Gamma\Gamma'\Phi + \Psi = \Sigma$$

Não unicidade das cargas (ϕ) e possibilidade de rotacionar soluções

Análise Fatorial via Componentes Principais

$$Y_i \in \mathbb{R}^p; \quad Y_i - \mu = \Phi \mathbf{f}_i + e_i \quad \Leftrightarrow \quad \Sigma = \Phi \Phi' + \Psi$$

$$\Sigma = V \Lambda V' \Rightarrow \Sigma = \lambda_1 V_1 V_1' + \dots + \lambda_m V_m V_m' + \dots + \lambda_p V_p V_p'$$

$$\mathbb{R}^p \rightarrow \mathbb{R}^m \Rightarrow \Sigma \approx \lambda_1 V_1 V_1' + \dots + \lambda_m V_m V_m' = \Phi \Phi'$$

Aproximação de Σ usando m **CP** \Rightarrow
define as cargas dos fatores comuns!

■ $\Phi = (\phi_1, \dots, \phi_m) = (\sqrt{\lambda_1} V_1, \dots, \sqrt{\lambda_m} V_m) \Rightarrow \phi_{jk} = \sqrt{\lambda_k} v_{jk}$ Carga do fator F_k à variável Y_j

■ $\Psi = \text{diag}(\sigma_1^2 - h_1^2, \dots, \sigma_p^2 - h_p^2) \Rightarrow \psi_j = \sigma_{jj} - \sum_{k=1}^m \phi_{jk}^2$ Componente específico da variável Y_j

■ Qual o valor do escore fatorial?

$$Y_{i(p \times 1)}; \quad Y_i - \mu = \Phi \mathbf{f}_i + e_i \Rightarrow \hat{\mathbf{f}}_{i(p \times 1)} = Z_i D_{\lambda_j}^{-1/2}$$

$$F_{i1} = \frac{Z_{i1}}{\sqrt{\lambda_1}}; \dots; F_{im} = \frac{Z_{im}}{\sqrt{\lambda_m}}$$


Os escores fatoriais são os componentes principais padronizados

Análise Fatorial via Componentes Principais

Obtenção do modelo de fatores comuns e específicos

$$Y_i \in \mathbb{R}^p; \quad Y_i - \mu = \Phi \mathbf{f}_i + e_i \quad \Leftrightarrow \quad \Sigma_{p \times p} = \Phi_{p \times m} \Phi'_{m \times p} + \Psi_{p \times p}$$

$$\Sigma \approx \lambda_1 V_1 V_1' + \dots + \lambda_m V_m V_m' = \Phi \Phi'$$


$$\Phi = (\phi_{jk}) = (\sqrt{\lambda_j} v_{jk}) \quad \Psi = \text{diag}(\sigma_{jj} - h_j^2)$$

$$R_{res} = \Sigma - (\Phi \Phi' + \Psi) \quad \text{Matriz residual}$$

Sob a solução via CP os elementos da diagonal de Σ estão bem aproximados (por construção), já para os elementos fora da diagonal de Σ a aproximação pode não ser boa!!

Critério de bondade de ajuste do modelo fatorial:

$$\text{SQ dos componentes de } R_{res} \leq (\lambda_{m+1}^2 + \lambda_{m+2}^2 + \dots + \lambda_p^2)$$

Análise Fatorial

Dados de Nutrição (n=27 e p=5)

Matriz de Correlação (R)

	energia	proteina	gordura	calcio	ferro
energia	1.00	0.17	0.99	-0.32	-0.10
proteina	0.17	1.00	0.02	-0.09	-0.17
gordura	0.99	0.02	1.00	-0.31	-0.06
calcio	-0.32	-0.09	-0.31	1.00	0.04
ferro	-0.10	-0.17	-0.06	0.04	1.00

Análise com m=2:

$$R_{5 \times 5} \approx \Phi_{5 \times 2} \Phi'_{2 \times 5} + \Psi_{5 \times 5}$$

Decomposição Espectral de R:

Autovalores (Λ :diag)

2.20 1.14 0.85 0.81 0.00
67%

Autovetores (V)

	V1	V2	V3	V4	V5
Y1	-0.65	0.09	-0.15	0.20	0.71
Y2	-0.15	-0.69	0.46	0.52	-0.10
Y3	-0.64	0.20	-0.22	0.13	-0.70
Y4	0.35	-0.01	-0.65	0.67	0.00
Y5	0.12	0.69	0.54	0.47	0.01

$$\Phi = (\phi_{jk}) = (\sqrt{\lambda_j} v_{jk})$$

Matriz de Cargas: Φ

	V1n	V2n
Y1	-0.97	0.09
Y2	-0.22	-0.74
Y3	-0.95	0.22
Y4	0.53	-0.01
Y5	0.18	0.74

Matriz de especificidades:

$\Psi = \text{diag}(0.052 \ 0.404 \ 0.055 \ 0.724 \ 0.424)$

$$1 - (-0.97^2 + 0.09^2)$$

	energia	proteina	gordura	calcio	ferro	z1	z2	F1	F2
[1,]	340	20	28	9	2.6	-239.75	23.25	-161.72	21.73
[2,]	245	21	17	9	2.7	-170.73	12.11	-115.17	11.32
[3,]	420	15	39	7	2.0	-299.12	35.50	-201.77	33.19
[4,]	375	19	32	9	2.5	-265.06	27.73	-178.79	25.92
[5,]	180	22	10	17	3.7	-120.94	4.98	-81.58	4.65
[6,]	115	20	3	8	1.4	-77.13	-2.26	-52.03	-2.11
[7,]	170	25	7	12	1.5	-114.98	-0.06	-77.56	-0.05
[8,]	160	26	5	14	5.9	-106.07	0.99	-71.55	0.93
[9,]	265	20	20	9	2.6	-185.59	15.09	-125.19	14.10
[10,]	300	18	25	9	2.3	-211.41	20.32	-142.60	19.00
[11,]	340	20	28	9	2.5	-239.76	23.18	-161.73	21.67
[12,]	340	19	29	9	2.5	-240.25	24.07	-162.06	22.50
[13,]	355	19	30	9	2.4	-250.71	25.51	-169.11	23.85
[14,]	205	18	14	7	2.5	-142.94	9.96	-96.42	9.32
[15,]	185	23	9	9	2.7	-126.69	3.88	-85.45	3.63
[16,]	135	22	4	25	0.6	-85.22	-2.35	-57.49	-2.20
[17,]	70	11	1	82	6.0	-18.26	2.33	-12.32	2.18
[18,]	45	7	1	74	5.4	-4.22	2.55	-2.85	2.38
[19,]	90	14	2	38	0.8	-48.67	-1.10	-32.83	-1.03
[20,]	135	16	5	15	0.5	-88.51	1.99	-59.71	1.86
[21,]	200	19	13	5	1.0	-140.07	7.61	-94.48	7.12
[22,]	155	16	9	157	1.8	-53.63	4.54	-36.18	4.24
[23,]	195	16	11	14	1.3	-131.84	9.00	-88.93	8.41
[24,]	120	17	5	159	0.7	-27.76	-0.78	-18.73	-0.73
[25,]	180	22	9	367	2.5	3.68	1.73	2.48	1.62
[26,]	170	25	7	7	1.2	-116.79	-0.23	-78.78	-0.22
[27,]	110	23	1	98	2.6	-40.97	-4.91	-27.64	-4.59

Os
escores
fatoriais
são os
escores
dos CP
padroniza-
dos!

Análise dos dados nutricionais com 2 Fatores Comuns

Matriz Residual: $\text{Res} = \Sigma - (\Phi\Phi' + \Psi)$

	energia	proteina	gordura	calcio	ferro
energia	0.000	0.026	0.048	0.190	0.007
proteina	0.026	0.000	-0.028	0.028	0.410
gordura	0.048	-0.028	0.000	0.192	-0.048
calcio	0.190	0.028	0.192	0.000	-0.046
ferro	0.007	0.410	-0.048	-0.046	0.000

Análise dos dados nutricionais com 3 Fatores Comuns

Matriz Residual: $\text{Res} = \Sigma - (\Phi\Phi' + \Psi)$

	energia	proteina	gordura	calcio	ferro
energia	0.000	0.084	0.021	0.107	0.075
proteina	0.084	0.000	0.057	0.284	0.198
gordura	0.021	0.057	0.000	0.072	0.050
calcio	0.107	0.284	0.072	0.000	0.253
ferro	0.075	0.198	0.050	0.253	0.000

Análise Fatorial via Máxima Verossimilhança

Suponha que os **fatores comuns F** e os **específicos e** seguem **distribuição Normal**, tal que, a distribuição marginal de Y_i é :

$$Y_{ip \times 1} \stackrel{iid}{\sim} N_p(\mu_{p \times 1}; \Sigma_{p \times p} = \Phi\Phi' + \Psi)$$

Para uma amostra de n vetores independentes de Y_i a função de verossimilhança é:

$$L(\mu, \Phi, \Psi | \mathbf{Y}) = \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} e^{-\frac{1}{2} \sum_{i=1}^n (Y_i - \mu)' \Sigma^{-1} (Y_i - \mu)}$$

sendo $\Sigma = \Phi\Phi' + \Psi$. Para $\hat{\mu} = \bar{Y}$ temos (exceto por constantes):

$$\ln L(\Phi, \Psi | S, \hat{\mu}) = -\frac{n}{2} \left(\ln |\Sigma| + \text{tr}(\Sigma^{-1} S) \right)$$

(Everitt, 2004;
Johnson and Wichern, 1992)

Maximizar $\ln L$ em Φ e Ψ é equivalente a minimizar a função D (com S e p , constantes):

$$D(\Phi, \Psi | S; \hat{\mu}) = \ln |\Sigma| - \ln |S| + \text{tr}(\Sigma^{-1} S) - p = \text{tr}(\Sigma^{-1} S) - \ln |\Sigma^{-1} S| - p$$

É necessário usar métodos numéricos para obter os estimadores de Φ e Ψ que minimizem a função D . Na solução da Análise Fatorial via MVS é preciso avaliar também a **identificabilidade** do *modelo reduzido* proposto.

Análise Fatorial via Máxima Verossimilhança

É interessante avaliar situações em que o modelo fatorial (um modelo reduzido) oferece uma interpretação mais simplificada (em uma dimensão mais baixa) para Y .

Como o vetor de parâmetros de locação μ não é de interesse na análise, usamos uma estimativa e podemos avaliar a aproximação da matriz de covariância amostral por meio do seguinte sistema de equações:

$$S_{p \times p} \cong \hat{\Phi}_{p \times m} \hat{\Phi}'_{m \times p} + \hat{\Psi}_{p \times p} \Rightarrow \text{Re } s = S - (\hat{\Phi} \hat{\Phi}' + \hat{\Psi})$$

com a restrição de unicidade: $(\Phi' \Psi^{-1} \Phi)$ é matriz diagonal.

Além disso, a diferença no número de parâmetros envolvidos é:

$$\delta = \frac{p(p+1)}{2} - \left[\underset{\substack{\uparrow \\ \text{parâmetros} \\ \text{em } S}}{pm} + \underset{\substack{\uparrow \\ \text{em } \Phi}}{p} - \underset{\substack{\uparrow \\ \text{em } \Psi}}{\left(\frac{m(m-1)}{2} \right)} \right] \underset{\substack{\uparrow \\ \text{sob a restrição} \\ \text{de unicidade}}}{\quad}$$

}

$\delta < 0$: sistema com mais parâmetros do que equações e o modelo fatorial não está bem definido aos dados

$\delta = 0$: soluções exatas são possíveis mas o modelo fatorial não oferece simplificação

$\delta > 0$: a simplificação/redução é possível por meio do modelo fatorial

Análise Fatorial – Escores Fatoriais

Os coeficientes na matriz Φ e os elementos da diagonal da matriz Ψ podem ser obtidos via CP ou MVS! Mas, e o escore dos fatores comuns?

Escore Fatorial: valor de cada indivíduo nos fatores comuns

Para o indivíduo i :

$$Y_{i(p \times 1)} \Rightarrow Y_i - \mu = \Phi \mathbf{f}_i + e_i \quad \left\{ \begin{array}{l} Y_{1i} - \mu_1 = \phi_{11}F_{1i} + \phi_{12}F_{2i} + \dots + \phi_{1m}F_{mi} + e_{1i} \\ Y_{2i} - \mu_2 = \phi_{21}F_{1i} + \phi_{22}F_{2i} + \dots + \phi_{2m}F_{mi} + e_{2i} \\ \dots \\ Y_{pi} - \mu_p = \phi_{p1}F_{1i} + \phi_{p2}F_{2i} + \dots + \phi_{pm}F_{mi} + e_{pi} \end{array} \right.$$

Como obter o valor de \mathbf{f}_i ? $i = 1, 2, \dots, n$

Análise Fatorial – Escores Fatoriais

Qual o valor de \mathbf{f}_i , ? $i = 1, 2, \dots, n$

✓ Método de Componentes Principais:

m primeiros componentes principais padronizados

$$Y_{i(p \times 1)} \Rightarrow Y_i - \mu = \Phi \mathbf{f}_i + e_i \Rightarrow \mathbf{f}_i = Z_i D_{\lambda_j}^{-1/2}$$

▪ **Método de Mínimos Quadrados Ponderados (Bartlett):**

Supondo μ , Φ e Ψ conhecidos \Rightarrow o modelo fatorial pode ser formulado como um *modelo de regressão linear heterocedástico* nas variáveis preditoras Φ .

O estimador (preditor) de \mathbf{f}_i é dado por:

$$\hat{\mathbf{f}}_i = \underbrace{(\Phi' \Psi^{-1} \Phi)^{-1}}_{(m \times p)} \underbrace{\Phi' \Psi^{-1} (Y_i - \mu)}_{(p \times 1)}$$

(m x p)

(p x 1)

Coefficiente do fator

Análise Fatorial – Escores Fatoriais

$$Y_{i(p \times 1)} \Rightarrow Y_i - \mu = \Phi \mathbf{f}_i + e_i \quad \text{Qual o valor de } \mathbf{f}_i, ? \quad i = 1, 2, \dots, n$$

- Método da Regressão: (μ , Φ e Ψ são assumidos conhecidos)

$$\mathbf{f}_i \sim N_m(0, I_m) \quad e_i \sim N_p(0, \Psi) \Rightarrow \begin{pmatrix} e_i \\ \mathbf{f}_i \end{pmatrix} \sim N_{p+m} \left(0, \begin{pmatrix} \Psi & 0 \\ 0 & I_m \end{pmatrix} \right)$$

$$Y_i - \mu = \Phi \mathbf{f}_i + e_i \sim N_p(0, \Sigma = \Phi\Phi' + \Psi) \Rightarrow \begin{pmatrix} Y_i - \mu \\ \mathbf{f}_i \end{pmatrix} \sim N_{p+m} \left(0, \begin{pmatrix} \Sigma & \Phi \\ \Phi' & I_m \end{pmatrix} \right)$$

$$\mathbf{f}_i / Y_i \sim N_m \left(\boxed{\Phi' \Sigma^{-1} (Y_i - \mu)}; I_m - \Phi' \Sigma^{-1} \Phi \right)$$

Esperança condicional dos fatores

O preditor de \mathbf{f}_i é dado por:

$$\hat{\mathbf{f}}_i = \Phi' \Sigma^{-1} (Y_i - \mu) = \underbrace{\Phi' (\Phi\Phi' + \Psi)^{-1}}_{(m \times p)} (Y_i - \mu)$$

Coeficiente do fator

Análise Fatorial – Rotação dos Fatores

Para qualquer matriz orthogonal Γ , tem-se:

$$\Phi^* = \Phi\Gamma; \quad \Gamma\Gamma' = I \quad \Rightarrow \quad \Phi^* \Phi^{*'} + \Psi = \Phi\Gamma\Gamma'\Phi + \Psi = \Sigma$$

Logo, não existe uma solução única para representar as cargas dos fatores comuns Φ .
Soluções rotacionadas podem ser mais interpretáveis.

*As communalidades
são invariantes!*

Métodos de Rotação Ortogonal:

- Rotação Varimax: simplifica as colunas da matriz de cargas Φ
- Rotação Quartimax: simplifica as linhas da matriz de cargas Φ
- Rotação Equimax: é um compromisso entre as duas outras técnicas

Existem ainda as rotações oblíquas. Neste caso, as communalidades variam.

Análise Fatorial via MVS

Dados de Nutrição (n=27 e p=5)

Matriz de Correlação (R)

	energia	proteina	gordura	calcio	ferro
energia	1.00	0.17	0.99	-0.32	-0.10
proteina	0.17	1.00	0.02	-0.09	-0.17
gordura	0.99	0.02	1.00	-0.31	-0.06
calcio	-0.32	-0.09	-0.31	1.00	0.04
ferro	-0.10	-0.17	-0.06	0.04	1.00

$$R_{5 \times 5} \approx \Phi_{5 \times 2} \Phi'_{2 \times 5} + \Psi_{5 \times 5}$$

Especificidades (Diagonal da matriz Ψ)

energia	proteina	gordura	calcio	ferro
0.005	0.005	0.005	0.897	0.965

Coefficientes (cargas) dos Fatores comuns (matriz Φ)

	Factor1	Factor2
energia	0.998	
proteina	0.197	0.978
gordura	0.983	-0.172
calcio	-0.319	
ferro		-0.160

F1: explica 0.998² da var. de energia e somente (-0.319)² da var. de cálcio

	Factor1	Factor2
Var	2.113	1.013
Proportion Var	0.423	0.203
Cumulative Var	0.423	0.625

R não está bem estruturada por 2 fatores comuns!

Test of the hypothesis that 2 factors are sufficient:

The chi square statistic is 10.67 on 1 degree of freedom (p-value=0.00109)

Análise Fatorial

Matriz de Correlação Observada (R)

	energia	proteina	gordura	calcio	ferro
energia	1.0000	0.1738	0.9871	-0.3204	-0.0998
proteina	0.1738	1.0000	0.0249	-0.0851	-0.1746
gordura	0.9871	0.0249	1.0000	-0.3081	-0.0606
calcio	-0.3204	-0.0851	-0.3081	1.0000	0.0443
ferro	-0.0998	-0.1746	-0.0606	0.0443	1.0000

Modelo ajustado: a matriz R está estruturada por 2 fatores comuns além da especificidade:

$$R_{5 \times 5} \approx \Phi_{5 \times 2} \Phi'_{2 \times 5} + \Psi_{5 \times 5}$$

Matriz de Correlação Ajustada via AF-MVS

	energia	proteina	gordura	calcio	ferro
energia	1.0017	0.1736	0.9854	-0.3183	-0.0928
proteina	0.1736	1.0000	0.0251	-0.0854	-0.1757
gordura	0.9854	0.0251	1.0016	-0.3102	-0.0675
calcio	-0.3183	-0.0854	-0.3102	1.0000	0.0346
ferro	-0.0928	-0.1757	-0.0675	0.0346	1.0000

Matriz Residual

	energia	proteina	gordura	calcio	ferro
energia	-0.0017	0.0002	0.0016	-0.0021	-0.0070
proteina	0.0002	0.0000	-0.0002	0.0003	0.0011
gordura	0.0016	-0.0002	-0.0016	0.0020	0.0069
calcio	-0.0021	0.0003	0.0020	0.0000	0.0097
ferro	-0.0070	0.0011	0.0069	0.0097	0.0000

Com base na matriz residual, de forma descritiva, 2 fatores comuns sugerem uma boa aproximação!

Dados de Nutrição e Escores (calculados por Bartlet) dos 2 primeiros fatores comuns:

	energia	proteina	gordura	calcio	ferro	Factor1	Factor2
[1,]	340	20	28	9	2.6	1.31	-0.02
[2,]	245	21	17	9	2.7	0.38	0.40
[3,]	420	15	39	7	2.0	2.06	-1.37
[4,]	375	19	32	9	2.5	1.63	-0.32
[5,]	180	22	10	17	3.7	-0.22	0.76
[6,]	115	20	3	8	1.4	-0.89	0.42
[7,]	170	25	7	12	1.5	-0.33	1.50
[8,]	160	26	5	14	5.9	-0.44	1.76
[9,]	265	20	20	9	2.6	0.59	0.12
[10,]	300	18	25	9	2.3	0.93	-0.43
[11,]	340	20	28	9	2.5	1.31	-0.02
[12,]	340	19	29	9	2.5	1.33	-0.27
[13,]	355	19	30	9	2.4	1.44	-0.29
[14,]	205	18	14	7	2.5	-0.01	-0.24
[15,]	185	23	9	9	2.7	-0.21	1.00
[16,]	135	22	4	25	0.6	-0.70	0.86
[17,]	70	11	1	82	6.0	-1.40	-1.63
[18,]	45	7	1	74	5.4	-1.62	-2.55
[19,]	90	14	2	38	0.8	-1.19	-0.95
[20,]	135	16	5	15	0.5	-0.79	-0.55
[21,]	200	19	13	5	1.0	-0.06	0.01
[22,]	155	16	9	157	1.8	-0.52	-0.61
[23,]	195	16	11	14	1.3	-0.24	-0.65
[24,]	120	17	5	159	0.7	-0.85	-0.31
[25,]	180	22	9	367	2.5	-0.26	0.77
[26,]	170	25	7	7	1.2	-0.33	1.50
[27,]	110	23	1	98	2.6	-0.93	1.14

Escores de Bartlet

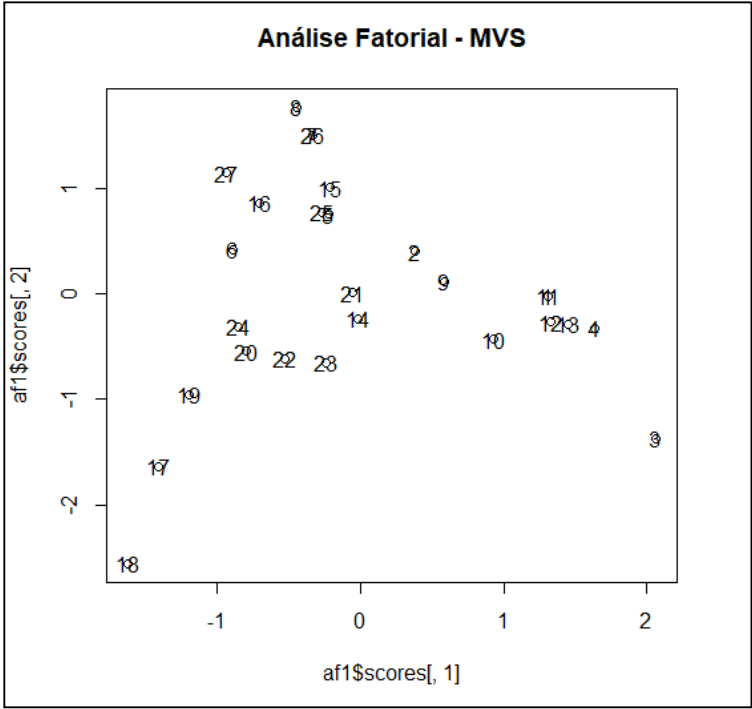


Gráfico dos escores das observações em \mathbb{R}^2 . Os escores fatoriais podem ser usados para análise de agrupamento de observações bem como para análise de diagnóstico de observações atípicas (o que também pode ser feito com os escores dos Componentes Principais e com a distância de Mahalanobis).

Dados de Nutrição: Matrizes de cargas (Φ) rotacionadas

Matriz de cargas (loadings) para estruturar a matriz R:

	Factor1	Factor2
energia	0.998	
proteina	0.197	0.978
gordura	0.983	-0.172
calcio	-0.319	
ferro		-0.160

	Factor1	Factor2
SS loadings	2.113	1.013
Proportion Var	0.423	0.203
Cumulative Var	0.423	0.625

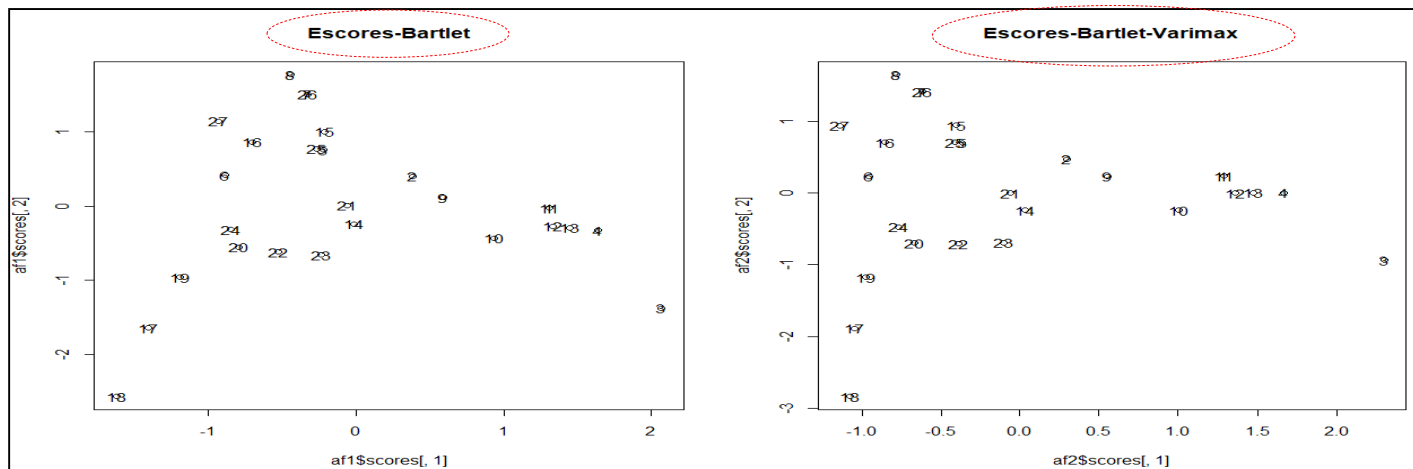
Cargas mudam mas
%Var permanence
constante!

Matriz de cargas rotacionada (Varimax):

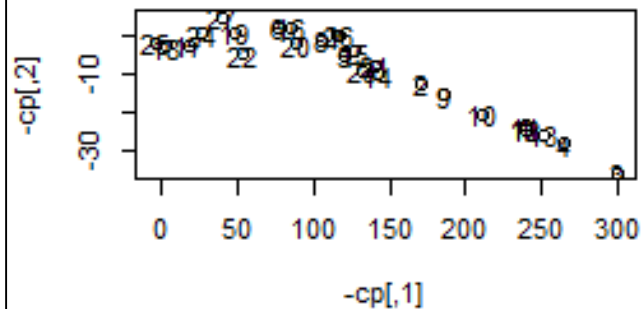
	Factor1	Factor2
energia	0.983	0.176
proteina		0.998
gordura	0.998	
calcio	-0.308	
ferro		-0.176

Busca por soluções
canônicas

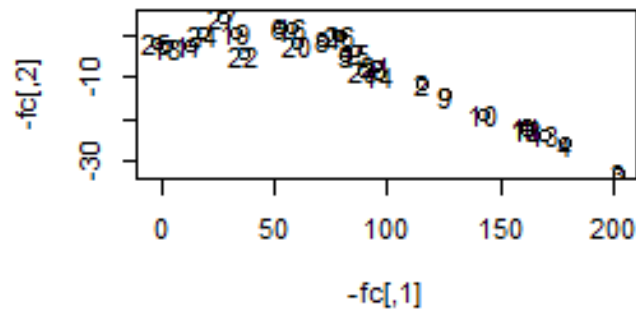
	Factor1	Factor2
SS loadings	2.061	1.065
Proportion Var	0.412	0.213
Cumulative Var	0.412	0.625



Escores - CP



Fatores Comuns via CP

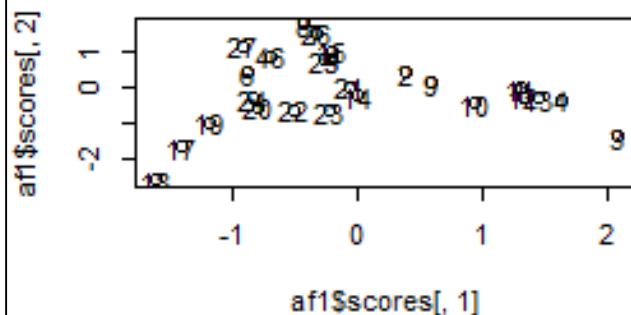


Dados de Nutrição:
Diferentes Escores
para a redução de
dimensionalidade:

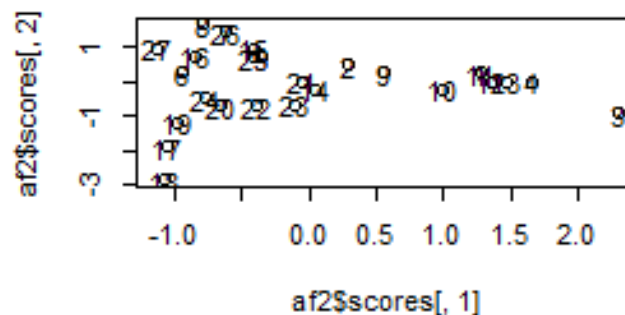
$$\mathbb{R}^5 \rightarrow \mathbb{R}^2$$

Em cada caso, qual é
o critério de
otimalidade adotado?

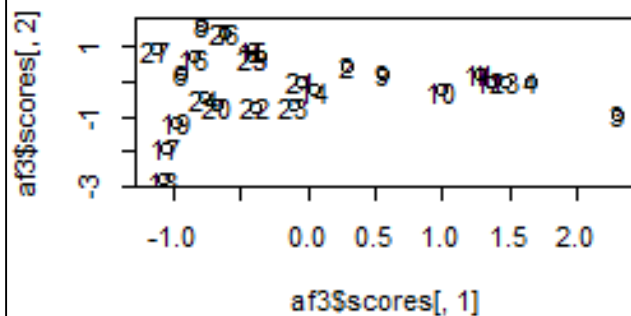
Escores-Bartlet



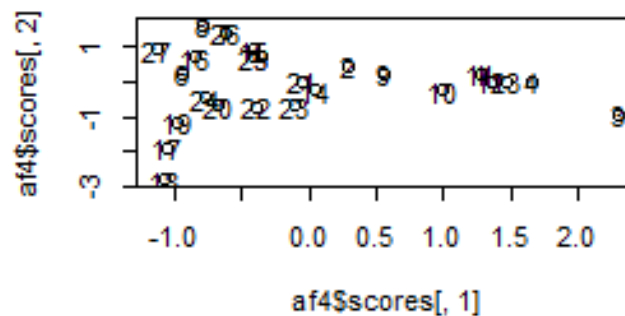
Escore-Bartlet-Varimax



Escore-Regressão-Varimax



Escore-Regressão-Varimax



Componentes Principais x Análise Fatorial

- Ambas buscam uma Redução de Dimensionalidade
- CONTUDO, os critérios de otimização usados em cada caso são diferentes:
 - ⇒ Análise Fatorial é ótima no sentido de obter uma boa aproximação das covariâncias (e correlações) por meio de fatores latentes comuns e específicos.
 - ⇒ Análise de CP tem o compromisso de maximizar a variância total das variáveis.
 - ⇒ Na análise de CP se o número de componentes retidos aumenta, isto NÃO altera os anteriores, mas isto pode não acontecer na Análise Fatorial sob a solução de MVS.
 - ⇒ Cálculo dos escores das observações nos CPs tem solução única. No caso de Análise fatorial (via MVS) existem diferentes procedimentos inferenciais propostos.

Componentes Principais x Análise Fatorial

⇒ As análises de CPs via matriz de covariância ($\Sigma = \text{Cov}(Y)$) ou de correlação ($R = \text{Cov}(Y^*)$) são **diferentes**, e não é possível relacioná-las. Na Análise Fatorial via MVS, a solução obtida é para a modelagem da matriz de correlação, mas a correspondente solução modelando a matriz de covariância é dada por:

$$\Phi = D_{s_{jj}}^{1/2} \Phi^*, \quad \Psi = D_{s_{jj}}^{1/2} \Psi^* D_{s_{jj}}^{1/2}$$

⇒ Teste (assintótico) da adequação do Modelo Fatorial:

$$H_0 : \Sigma = \Phi\Phi' + \Psi \quad H_1 : \Sigma \text{ com estrutura geral}$$

A estatística da razão de verossimilhanças (sob normalidade) é:

$$-2 \ln \frac{L_0}{L_1} = n \ln \left(\frac{|\hat{\Phi}\hat{\Phi}' + \hat{\Psi}|}{|S_n|} \right)$$

Usando a correção de Bartlett, rejeita-se H_0 a um nível de significância α se:

$$(n-1-(2p+4m+5)/6) \ln \left(\frac{|\hat{\Phi}\hat{\Phi}' + \hat{\Psi}|}{|S_n|} \right) > \chi^2_{[(p-m)^2-p-m]/2}(\alpha)$$