MAE5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Sem/2022

Análise Multivariada

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

Já vimos ☺

Matriz de Dados: Estatísticas descritivas multivariadas em $\Re^{p}$, $\Re^{p\times p}$ e $\Re^{n\times n}$

Episóides de Concentração (Diagnóstico de outliers), Boxplot Bivariado

Matriz Aleatória: Distribuição Normal Multivariada, Distribuições Amostrais

Testes de Hipóteses Multivariadas para μ e Σ:

Caso de Uma, Duas e Muitas Populações N_p (MANOVA)

Regiões de Confiança, I.C. Simultâneos, Correções para Múltiplos testes

Decomposições: $SS_T = SS_B + SS_W$

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

Observações *iid :* *Caso n>p* (soluções clássicas)

Caso n<<p (Big-p)

Caso n>>p (Big-n)

Observações Correlacionadas

Redução de Dimensionalidade

$\Re^p \rightarrow \Re^m$, m<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}

$$Y_{n\times p}; \quad n > p$$

$$Y_{i_{p\times 1}} \sim \left(\mu\,;\Sigma\right) \quad \Re^p \rightarrow \Re^m, m < p$$

Como veremos, a Redução de Dimensionalidade depende da **estrutura dos dados**!

Estrutura dos Dados:

$\mu_{g_{p\times 1}}$?	$\Sigma_{g_{p\times p}}$?	$i = 1,\ldots,n_g;\ \ g = 1,\ldots,G$?
$\mu_g = \mu$	$\Sigma_g = \Sigma$	iid ?

Redução de Dimensionalidade: $\Re^p \rightarrow \Re^m$, m<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}

$$Y_{n\times p} \approx \underbrace{\begin{bmatrix} - & - \\ - & - \\ - & - \\ - & - \end{bmatrix}_{n\times m}} \underbrace{\begin{bmatrix} | & | & | \\ | & | & | \end{bmatrix}_{m\times p}}$$

Matriz de "Escores" dos *n* indivíduos nas *m* novas variáveis

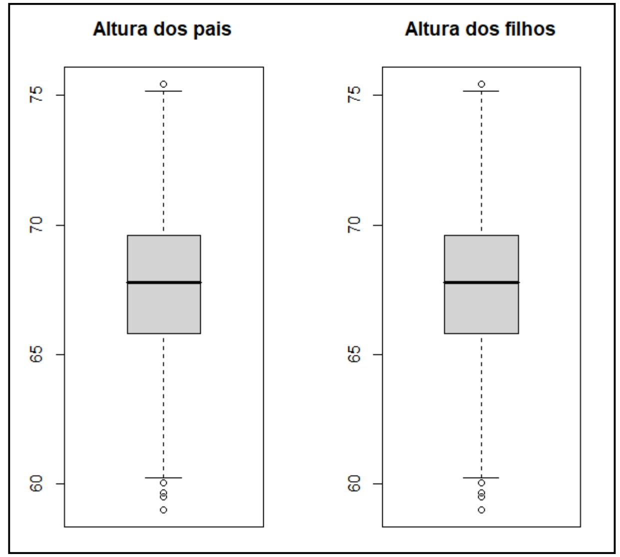
Matriz de "Cargas" das as *p* variáveis para as *m* novas variáveis

A matriz de dados Y_{nxp} é aproximada por uma Matriz de Escores na dimensão m (m<p).
Na redução de dimensionalidade as p variáveis contribuem com Cargas específicas.

Análise Multivariada

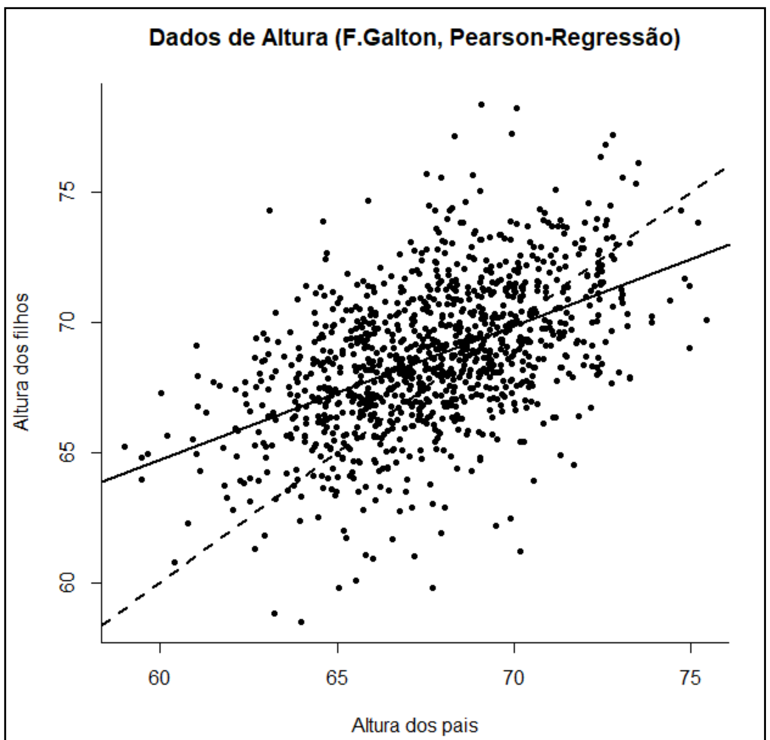
Dados "father.son" do R

```
        fheight  sheight
1      65.04851 59.77827
2      63.25094 63.21404
3      64.95532 63.34242
4       65.75250 62.79238
5    ...
1076 71.78314 69.30589
1077 70.73837 69.30199
1078 70.30609 67.01500
```



Matriz de dados:

$$Y_{1078\times 2}$$

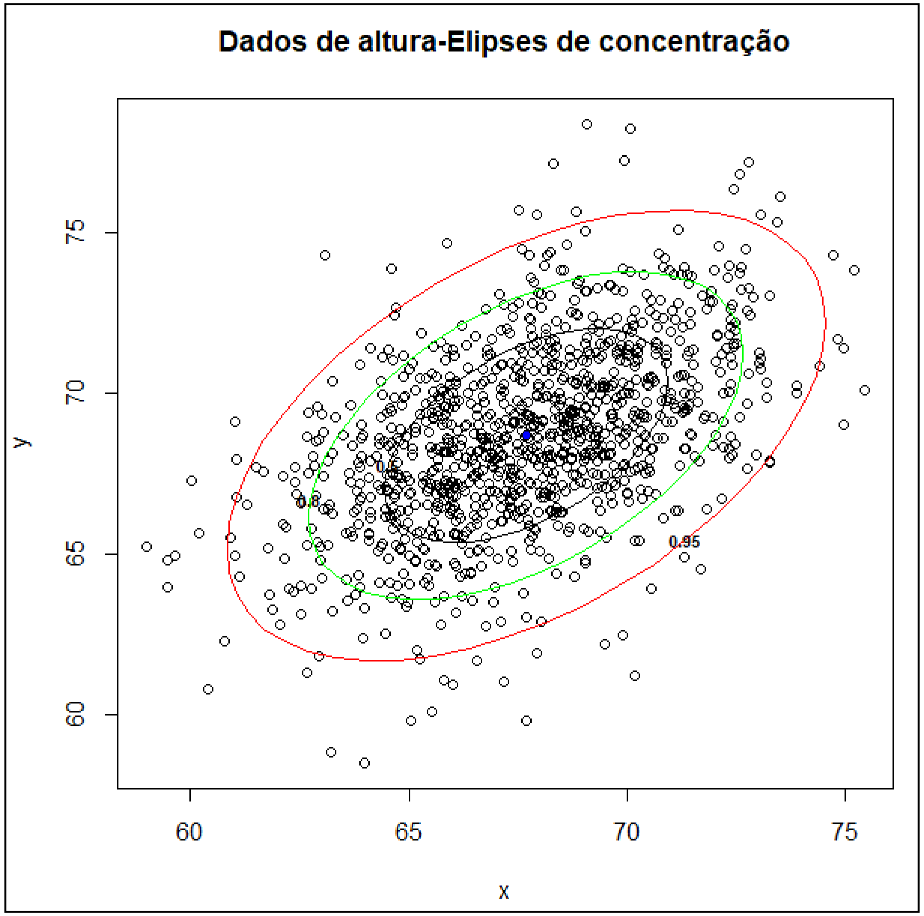


- Análise UNIVARIADA de cada variável
- Análise MULTIVARIADA: análise conjunta das variáveis. Leva em conta a correlação (r=0.50)

Indicação das retas y=x (tracejada) e da reta de regressão y=33.8866+0.5141x (linha contínua)

Note que há uma regressão da altura dos filhos quando os pais são mais altos que x=69.72628

Elipse de Concentração dos Dados



$$d_M^2 = \left(Y_i - \bar{Y}\right)' S^{-1} \left(Y_i - \bar{Y}\right) \le \chi^2_{p(1-\alpha)}$$

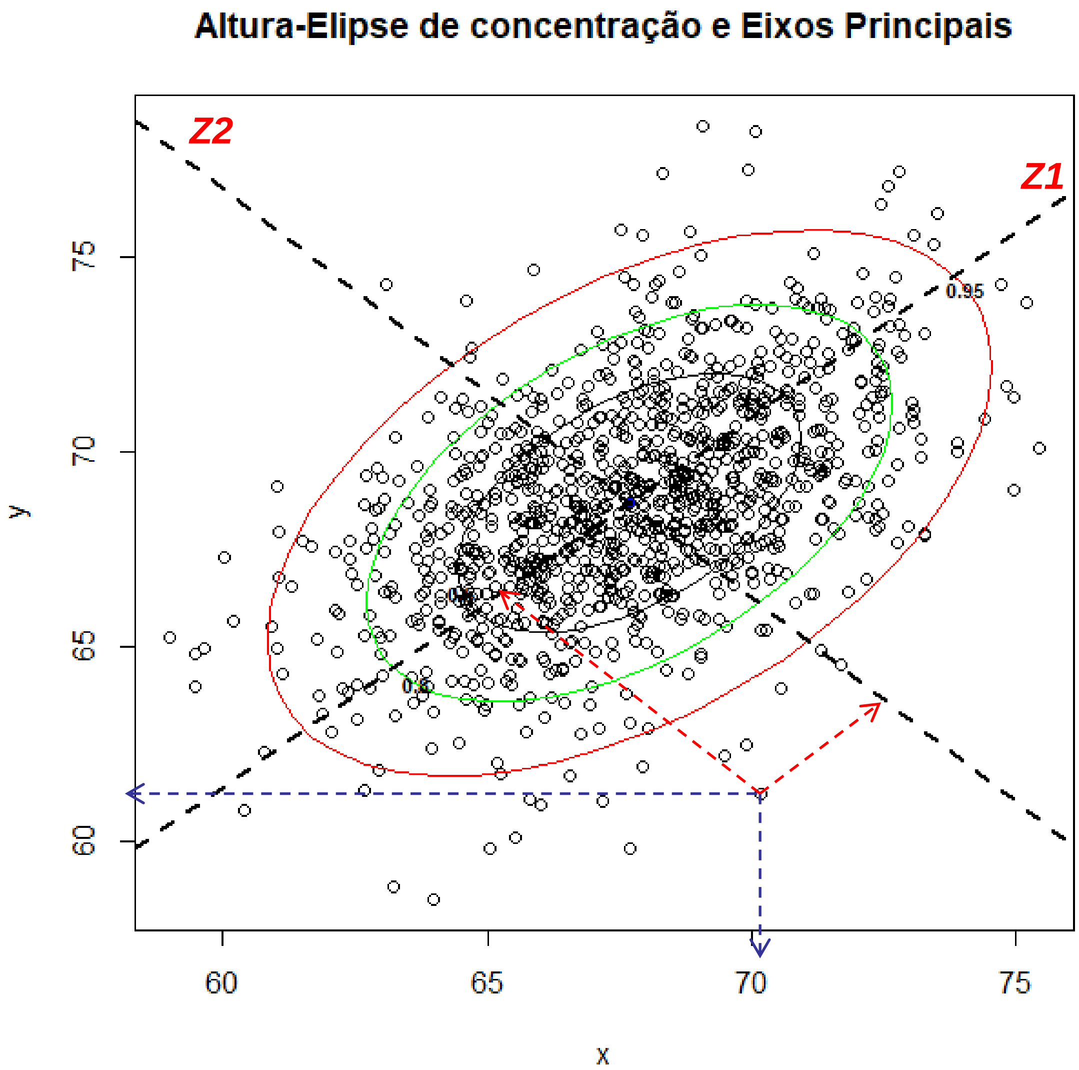
Elipse de 95% de concentração (p=2, α=5%)

Elipse de 80% de concentração (p=2, α=20%)

Elipse de 50% de concentração (p=2, α=50%)

Dados Y, representados como pontos (x,y) nos eixos X e Y, podem ser representados em **NOVOS EIXOS**, por exemplo, os dois **Eixos Principais da Elipse** de concentração, que correspondem às direções de maior e de menor variabilidade dos dados.

Elipse de Concentração dos Dados
Eixos Principais de Variação dos Dados



Os dados de altura dos pais e dos filhos, (x,y), podem ser representados por meio de coordenadas neste novo sistema de eixos principais:

$(x,y) \rightarrow (z1, z2)$

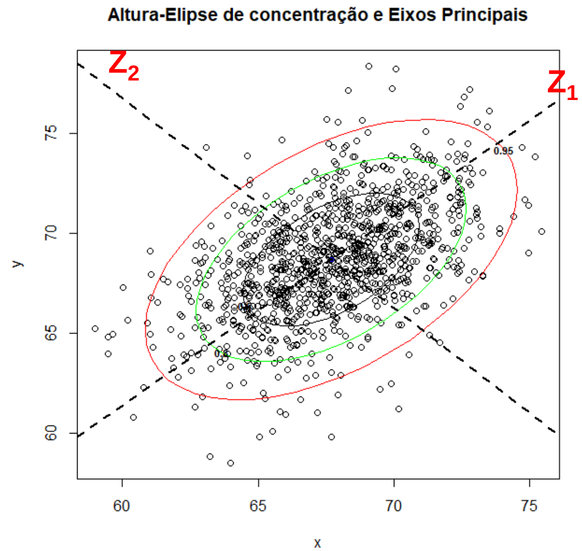
Os Eixos Principais da elipse são obtidos da Análise Espectral da Matriz de Covariâncias dos dados, o que equivale a realizar a **Análise de Componentes Principais**

Decomposição Espectral de Matrizes Quadradas

```
Dados "father.son" do R
     x:fheight y:sheight
1     65.04851  59.77827
2     63.25094  63.21404
...
1077  70.73837  69.30199
1078  70.30609  67.01500
```

p=2

$\Rightarrow$



(x,y) $\rightarrow$ (z1, z2)

(x,y)V_1 (x,y)V_2

Os vetores $\mathbf{V_1}$ e $\mathbf{V_2}$ são obtidos da

$$Y_{n\times p} \rightarrow S_{p\times p} \qquad S_{p\times p} = V\,\Lambda\,V'$$

Decomposição Espectral da Matriz de Covariância

$$|S - I\lambda| = 0 \quad \Rightarrow \quad \lambda_1, \lambda_2, \ldots, \lambda_p$$

$$S\lambda_j = \lambda_j V_j$$

V: matriz de autovetores de S (direções de variabilidade)

Λ: matriz diagonal de autovalores de S (medem a variabilidade dos dados)

$$V_{p\times p} = \left(V_1\; V_2\; \ldots V_p\right) = \begin{pmatrix} V_{11} & V_{12} & \ldots & V_{1p} \\ V_{21} & V_{22} & \ldots & V_{2p} \\ & & \ldots & \\ V_{p1} & V_{p2} & \ldots & V_{pp} \end{pmatrix}$$

$$\Lambda_{p\times p} = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_p \end{pmatrix}$$

Variância total e variância generalizada:

$$\lambda_1 + \ldots + \lambda_p = traço S$$

$$\lambda_1 * \ldots * \lambda_p = \det S$$

Decomposição Espectral de Matrizes Quadradas

Dados "father.son" do R

```
      fheight  sheight
1     65.04851 59.77827
2     63.25094 63.21404
3     64.95532 63.34242
4     65.75250 62.79238
5   ...
1076 71.78314 69.30589
1077 70.73837 69.30199
1078 70.30609 67.01500
```

Matriz de Covariância: S

```
         x        y
x 7.534303 3.873333
y 3.873333 7.922545
```

Var(x)=7.53 Var(y)=7.92

Cor(x,y)=$\frac{3.87}{\sqrt{7.53}\sqrt{7.92}} = 0.50$

Decomposição Espectral de S
Uso do comando eigen() do pacote R

Autovalores: 11.60662 3.85023

Autovetores:

```
          [,1]       [,2]
[1,] 0.6891827 -0.7245877
[2,] 0.7245877  0.6891827
```

$$Y_{n\times p} \rightarrow S_{p\times p}$$

$$S_{p\times p} = V\,\Lambda\,V'$$

$$V_{2\times 2} = \begin{array}{cc} V_1 & V_2 \\ \begin{pmatrix} 0.689 & -0.7245 \\ 0.7245 & 0.6891 \end{pmatrix} \end{array}$$

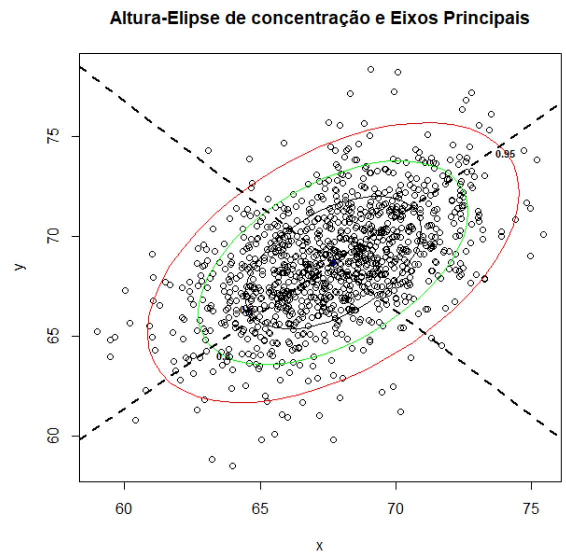
$$\Lambda_{2\times 2} = \begin{pmatrix} 11.606 & 0 \\ 0 & 3.850 \end{pmatrix}$$

Para finalidade de redução de dimensionalidade:
V é Matriz de CARGAS
YV é Matriz de Escores

Transformação de Coordenadas: $\Re^2 \to \Re^2$

Alturas: Coordenadas Originais

	fheight	sheight
1	65.04851	59.77827
2	63.25094	63.21404
3	64.95532	63.34242
4	65.75250	62.79238
5	...	
1076	71.78314	69.30589
1077	70.73837	69.30199
1078	70.30609	67.01500



Cov(alt)

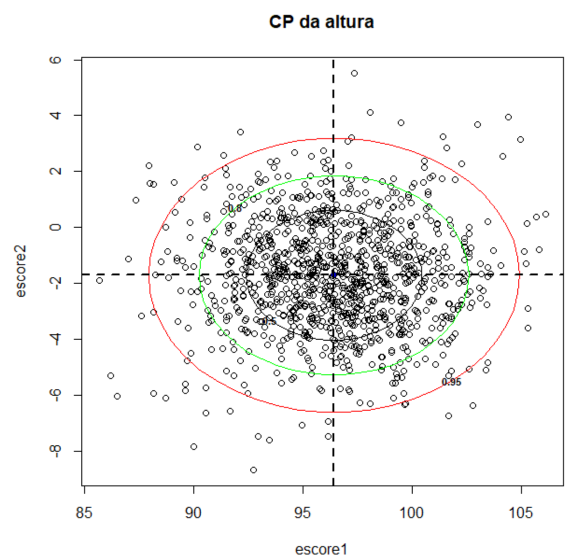
	x	y
x	7.534303	3.873333
y	3.873333	7.922545

Variância total:
15.45685

Variância generalizada:
44.68815

Alturas em Coordenadas nos Eixos Principais

	Escore1	Escore2
1	88.14490	-5.935200758
2	89.39556	-2.264830643
3	90.66322	-3.411326085
4	90.81407	-4.368030899
5	...	
1076	99.68989	-4.248760107
1077	98.96702	-3.494420473
1078	97.01198	-4.757349555



Cov(cp.alt)

	Escore1	Escore2
Escore1	11.6066	0
Escore2	0	3.8502

Variância total:
15.45685

Variância generalizada:
44.68815

Ténicas Multivariadas de Redução de Dimensionalidade

Como obter vetores reducionistas dos dados?

- Análise de Componentes Principais: $Y_{n\times p} \Rightarrow \Sigma_{p\times p}$
- Escalonamento Multidimensional: $Y_{n\times p} \Rightarrow D_{n\times n}$
- Análise de Correspondência: $Y_{I\times J} \in [0,1]^{I\times J};\ D_{I\times I};\ D_{J\times J}$
- Análise Fatorial: $Y_{n\times p} \Rightarrow \Sigma_{p\times p}$
- Análise Discriminante (MANOVA): $Y_{n\times(p+1)} \Rightarrow \Sigma_{T\ p\times p} = \Sigma_{B\ p\times p} + \Sigma_{W\ p\times p}$
- Análise de Agrupamento: $Y_{n\times p} \Rightarrow D_{n\times n}$
- Análise de Correlação Canônica: $Y_{n\times(p+q)} \Rightarrow \Sigma = \begin{pmatrix} \Sigma_{p\times p} & \Sigma_{p\times q} \\ \Sigma_{q\times p} & \Sigma_{q\times q} \end{pmatrix}$

PLS

✓ Objetivo da análise
✓ Estrutura dos Dados
✓ Soluções (e Restrições impostas)
✓ Representação Gráfica dos dados: BiPlot, Dendrograma, HeatMap

Análise de Componentes Principais

Análise Clássica

$n > p$
Observações iid
(respostas quantitativas)

Análise de Componentes Principais

(Pearson, 1901)

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}

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

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

- A variável Y_j , para algum j=1,...,p, pode ser eliminada da análise?
- Como as p variáveis podem ser ordenadas segundo sua "importância" na análise?

Considerar a estrutura de Σ

Análise de Componentes Principais
Estruturas de Σ e R

Como proceder com a redução de dimensionalidade nos seguintes casos?

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

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

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 será uma média ponderada que atribui pesos iguais à todas as variáveis.

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

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

Caracterização nutricional de 27 produtos alimentícios (Everitt, 2007)

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	-270.67	-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.
Como obter um ESCORE resumindo o padrão nutricional destes produtos?

Análise de Componentes Principais
Dados dos Cães

$Y_{7\times 6}$

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 7 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

$$Y_i \in \Re^p \quad \rightarrow \quad Z_i = A_{p\times p} \, Y_{i_{p\times 1}} \in \Re^p$$

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

$$tr\,\Sigma = tr\Lambda$$ Preservar a variância total

1. Realizar uma transformação linear de Y que preserve a variância total
2. Redução de dimensionalidade: transformar Y em Z, reduzindo de p para m variáveis (m<p), mas preservando ao máximo a variância total

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

em termos de autovalores de Σ

$$Y_{i\,p\times 1} \in \Re^p \quad \Rightarrow \quad Z_{i\,m\times 1} \in \Re^m$$

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

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

...

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

$$Z_{ij} = a_j{}'Y_i$$

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

$$A_{p\times m} = (a_{jk})$$

Como obter a matriz ***A*** de cargas (que atribui pesos às variáveis) ?

Análise de Componentes Principais

- Formulação de CP como um problema de otimização:

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

$$Z_{ik} = a_k{}' Y_i; \quad \mathrm{var}(a_k{}' Y) = a_k{}' \Sigma a_k$$

$$\boxed{\arg\max_{\|a\|=1} \frac{a\ '\Sigma a}{a\ 'a} = V_1; \quad \max \frac{V_1\ '\Sigma V_1}{V_1\ 'V_1} = \lambda_1}$$

$V_1 \ e \ \lambda_1$: primeiro autovetor e primeiro autovalor da decomposição spectral de Σ. Os m primeiros autovetores podem ser usados na redução de dimensionalidade

$$\Rightarrow Z_{ik} = V_k' \, Y_{ip\times 1}; \qquad Var\left(Z_{ik}\right) = \lambda_k \quad k = 1, 2, \ldots, m, \quad i = 1, 2, \ldots, n$$

$$Y_i \in \Re^p \Rightarrow Z_i = V'_{m\times p} \, Y_{i_{p\times 1}} \in \Re^m \qquad V_{p\times m} = \left(V_1 \quad \cdots \quad V_m\right) = \begin{pmatrix} V_{11} & & V_{1m} \\ \cdots & \cdots & \cdots \\ V_{p1} & & V_{pm} \end{pmatrix}$$

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

Análise de Componentes Principais

Decomposição spectral de Σ (posto completo, *p*):

$$\Sigma_{p\times p} = V_{p\times p}\Lambda_{p\times p}V_{p\times p}' \quad ; \quad VV' = V'V = I_p \qquad \Lambda = diag\left(\lambda_j\right)$$

$$\left.\begin{array}{l} \lambda_j\ ?; \quad \left|\Sigma - \lambda I_p\right| = 0 \\ \\ V_j\ ?; \quad \Sigma V_j = \lambda_j V_j \end{array}\right\} \quad \begin{array}{ccc} \lambda_1 & \rightarrow & V_1 \\ \lambda_2 & \rightarrow & V_2 \\ \ldots & & \\ \lambda_m & \rightarrow & V_m \\ \ldots & & \\ \lambda_p & \rightarrow & V_p \end{array}$$

$$tr\,\Sigma = tr\left(V\Lambda V'\right) = \sum_{j=1}^{p}\lambda_j V_j{}'V_j = \sum_{j=1}^{p}\lambda_j = tr\Lambda$$

Aproximação para Σ (∈ $\Re^{pxp}$) em $\Re^{mxm}$ (p<m)

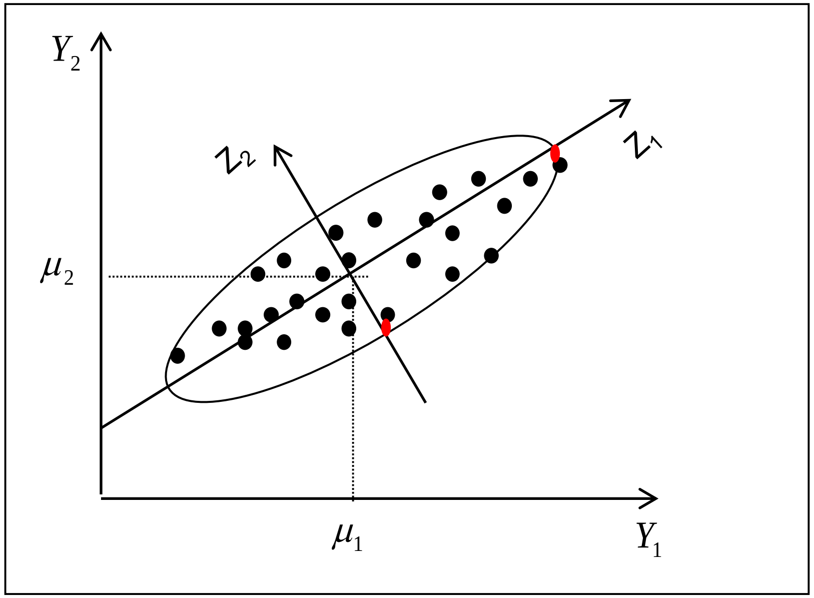
$$\Sigma = \sum_{j=1}^{p}\lambda_j V_j V_j{}' \cong \sum_{j=1}^{m}\lambda_j V_j V_j{}'$$

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 elipsóides



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

$$Y \quad \Rightarrow Z = YV$$

$$(y_1, y_2) \Rightarrow (z_1, z_2)$$

Z_1 : primeiro componente principal (escore 1)

Z_2 : segundo componente principal (escore 2)

$$Z_1 = V_1'Y \; ; \quad Var(Z_1) = V_1'\Sigma V_1 = \lambda_1$$

$$Z_2 = V_2'Y \; ; \quad Var(Z_2) = V_2'\Sigma V_2 = \lambda_2$$

$$Var(Z_1) \geq Var(Z_2)$$

$$Cov(Z_1, Z_2) = V_1'\Sigma V_2 = 0$$

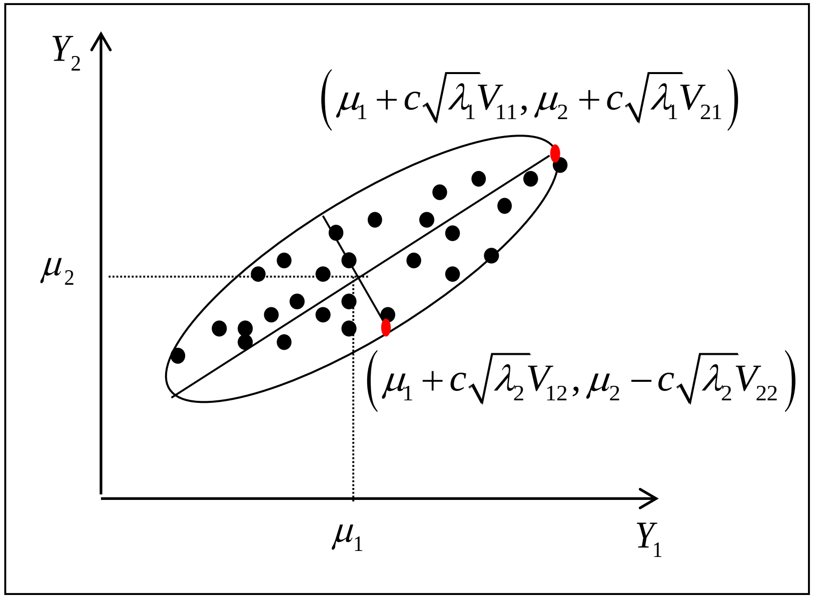
Decomposição espectral de Σ (autovalores e autovetores) permite uma representação dos dados em eixos ortogonais e nas direções de máxima variação (total) dos dados.

Decomposição Espectral de Σ e a Elipse de Concentração de Observações

Exemplo da Normal bivariada:

$$\mathbf{Y}_{2\times 1} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N_2\left(\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \boldsymbol{\Sigma}_{2\times 2} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right); \qquad \sigma_{11} = \sigma_{22}$$

Elipse de Concentração de observações



$$d_M^2 \sim \chi_2^2 \quad \Rightarrow \quad P\left(d_M^2 \le c^2\right) \le (1-\alpha)$$

Obter *c* para a inclusão de 90%, 95% e 98% dos pontos amostrais.

Os eixos (maior e menor) da elipse de concentração são calculados pela decomposição espectral de Σ:

$$\left|\Sigma - \lambda I_2\right| = 0$$ autovalores

$$\Rightarrow \quad \lambda_1 = \sigma_{11} + \sigma_{12} \qquad \lambda_2 = \sigma_{11} - \sigma_{12}$$

$$\Sigma V_j = \lambda_j V_j$$ autovetores

$$\Rightarrow \quad V_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \qquad V_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

$$\Rightarrow \quad V_j' V_j = 1 \qquad V_j' V_{j'} = 0$$

Revise os seguintes resultados:

Análise de Componentes Principais

Exemplo 1: $\Sigma = \sigma^2 I \quad ; \quad \Sigma = V\Lambda V' \Rightarrow V = I; \quad \Sigma V_j = \sigma^2 V_j$

σ^2 é autovalor com multiplicidade p.

$$Z_{ji} = V_j' Y_i = Y_{ij}$$

Não é possível reduzir nem ordenar as variáveis.

Exemplo 2: $\Sigma = diag\left(\sigma_{jj}\right); \quad \Sigma = V\Lambda V' \Rightarrow V = I; \quad \Sigma V_j = \sigma_{jj} V_j$

$$Z_{ji} = V_j' Y_i = Y_{i(j)}; \quad \left(\sigma_{jj}; V_j\right)$$

Os CP são as variáveis originais ordenadas pelas variâncias.

Exemplo 3: $\Sigma = \left(1-\rho\right) I + \rho 1 1'; \rho > 0 \quad ; \quad \Sigma = V\Lambda V' \Rightarrow \lambda_1 = 1 + \left(p-1\right)\rho; \quad V_1 = 1/\sqrt{p} 1_p$

$$\lambda_2 = \ldots = \lambda_p = 1 - \rho$$

$$Z_{1i} = V_1' Y_i = \sum_{j=1}^{p} \frac{Y_{ij}}{\sqrt{p}}$$

CP1 é um “índice” com pesos iguais, e de norma 1, para todas as variáveis

$$\%VarExpl = \frac{\lambda_1}{p} = \frac{1+(p-1)\rho}{p} = \rho + \frac{1-\rho}{p} \cong \rho \text{ se } \rho \to 1 \text{ ou p} \to \infty$$

Componentes Principais
Quantos Componentes Reter na Análise?

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

- Preservar "grande" parte da variância total dos dados:

Para variáveis padronizadas: $\lambda_j \geq 1$

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

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

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

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:

$$r_{jk} = Cor\left(Y_j, Z_k\right) = \frac{a_{jk}\sqrt{\lambda_k}}{\sqrt{\sigma_{jj}}}$$

a_{jk} é a coordenada *j* do autovetor *k*

- Garantir que "grande" parte da variabilidade de cada variável original seja explicada pelos *m* CP

r_{jk}^2 é a proporção da variância de Yj que é explicada pela CP Zk

Análise de Componentes Principais

Obtenção dos Componentes Principais dos Dados Nutricionais

Decomposição Espectral da Matriz de Covariância Su:

Autovalores de Su

```
11552.53  4903.92    20.43      2.07      0.35
```

Autovetores de Su

Cargas (pesos) das variáveis no PC1

	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

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

$$PC1 = Z_1 = YV_1$$

$$PC2 = Z_2 = YV_2$$

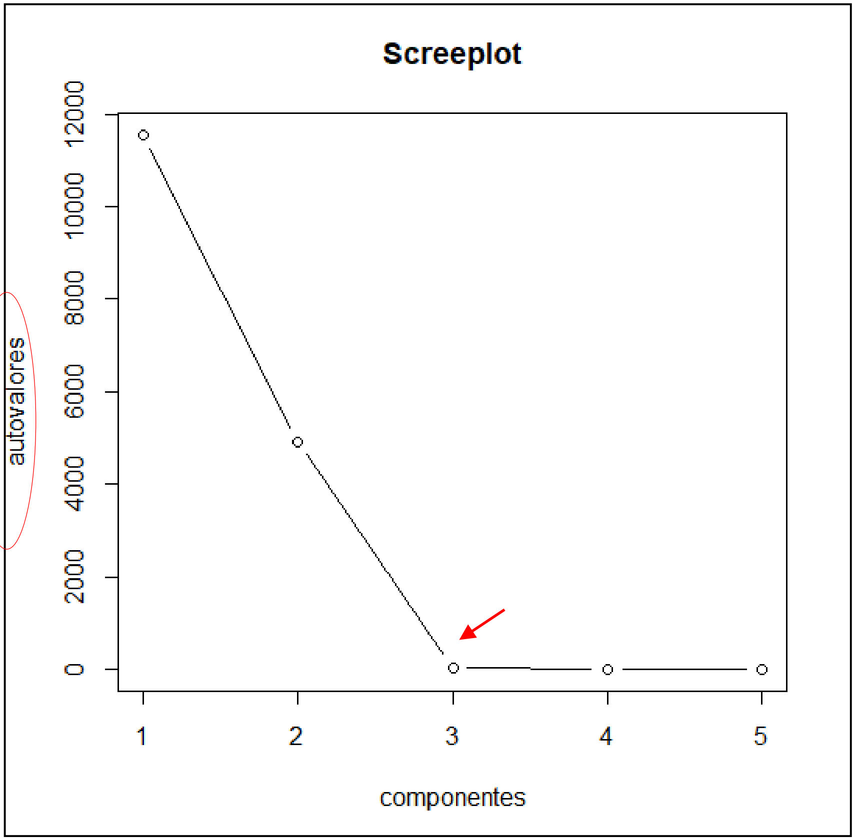
Importância dos Componentes Principais:

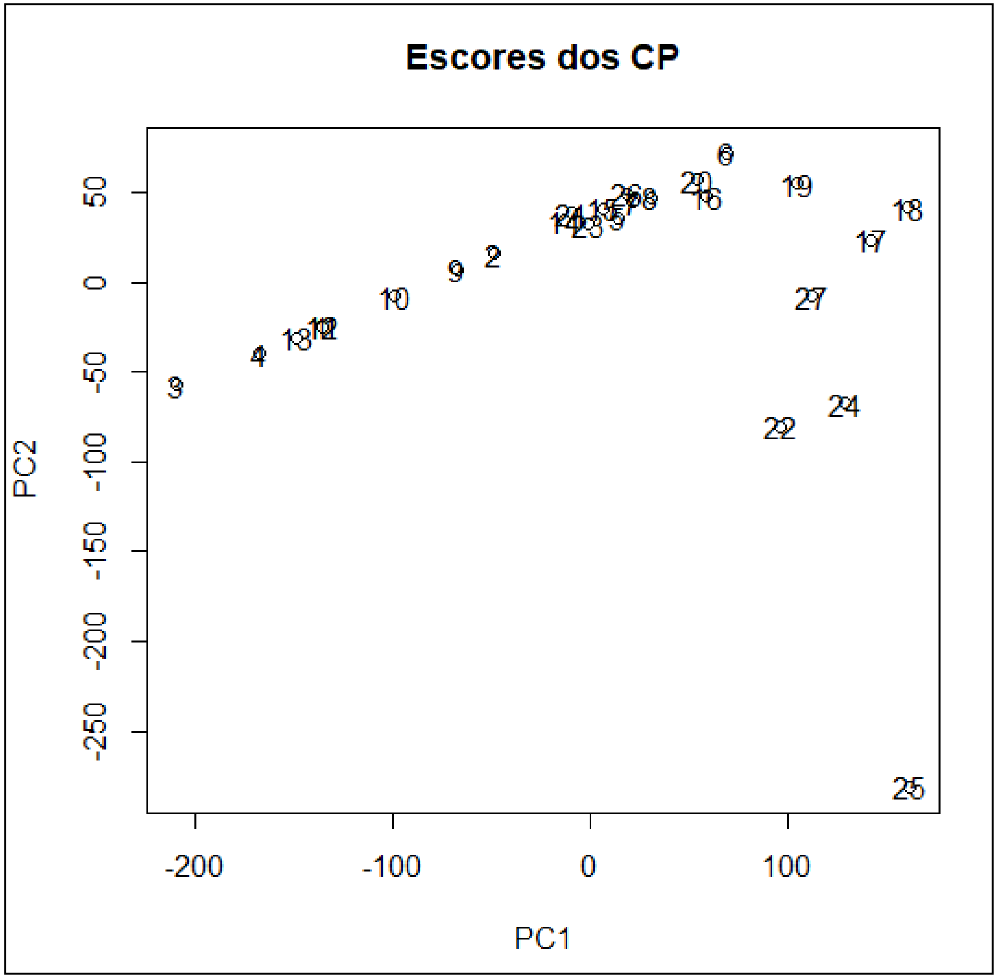
$\sqrt{11552.53}$

	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

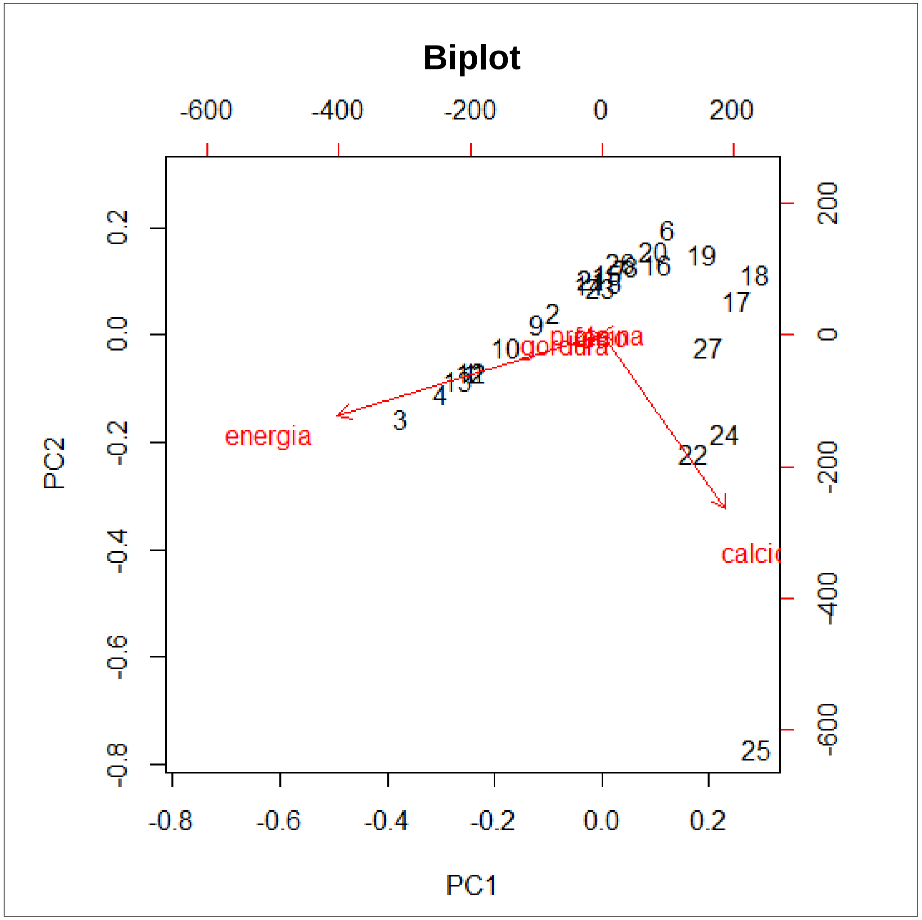
Dados Nutricionais (n=27; p=5)





Análise de Componentes Principais

Dados Nutricionais (n=27; p=5)



Biplot: Representação simultânea dos escores dos CP e dos 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

Calcular a correlação entre as variáveis originais e CP1 e CP2.

Calcular a variância de CP1 e CP2 (mostre que é os autovalores correspondentes).

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 (r)

	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(r^2)

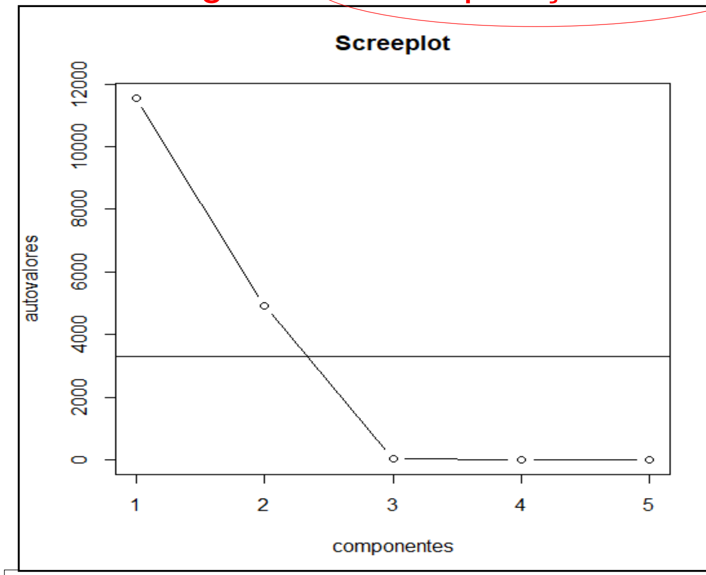
	PC1	PC2	variância
energia	0.915715932	0.0842893318	10243.019943
proteina	0.030317116	0.0003893978	18.076923
gordura	0.887918489	0.0869946033	126.720798
calcio	0.338254192	0.6617320111	6089.344729
ferro	0.009787159	0.0002638710	2.134103

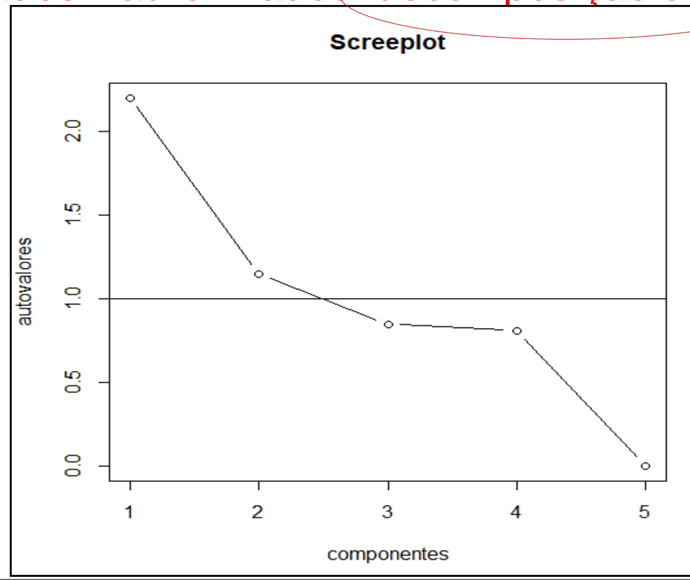
Análise de Componentes Principais

NÃO é invariante por padronização dos dados

Dados Originais: decomposição de S

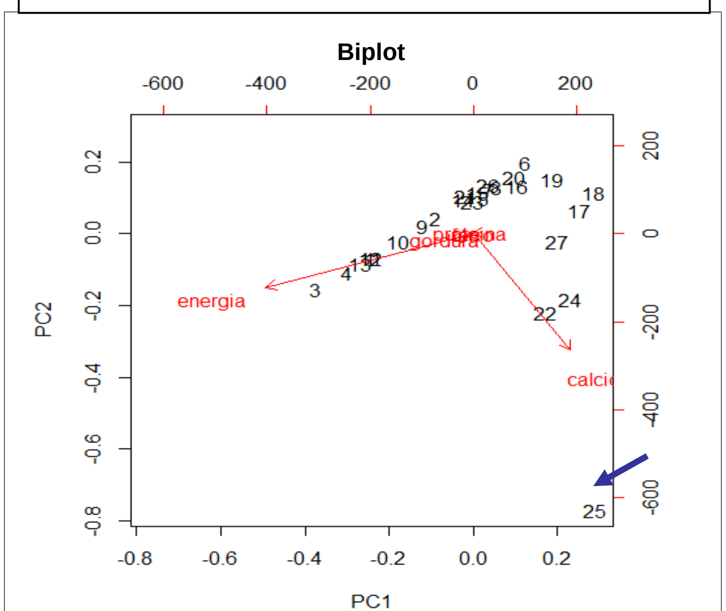
Dados Padronizados: decomposição de R

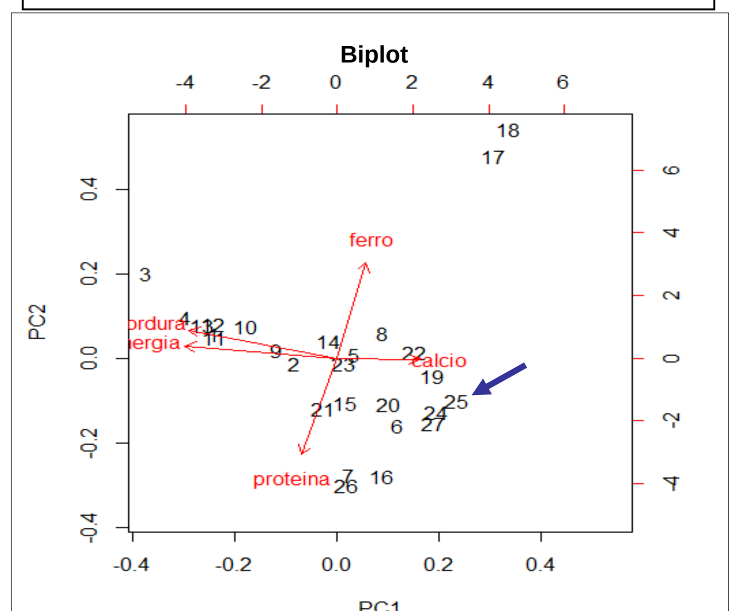




Prop.Ac.Expl.;
0.44 0.67 0.84

%Expl:
99,86





%Expl:
67%

Análise de Componentes Principais

Na prática, Σ e R não são conhecidas e estimativas (MVS ou estimadores robustos) são utilizadas na decomposição espectral.

- Variáveis originais (Y) em escalas diferentes (com heterocedasticidade) podem ser padronizadas, o que equivale aos CP via R. Os resultados via Σ 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 (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 (coordenadas dos autovetores V_j) e as correlações (r_{YjZk}) das variáveis originais com os CP são úteis na interpretação dos componentes principais.

- A estrutura de Σ é decisiva na análise de CP. Sob a estrutura uniforme, as variáveis originais têm o mesmo "peso" na construção do CP1.

Análise de Componentes Principais

Pesquisar:

- Algoritmo NIPALS (Nonlinear Iterative Partial Least-Squares Algorithm)

  Wold (1966)

  Permite calcular a matriz de escores e os vetores de cargas de uma matriz Y

  No R: função `nipals` da biblioteca `chemometrics`
- PCA Não Linear ou Autoencoder (rede neural para obter as Cargas): Scholz et al., 2005, `nlpca no R`
- PCA robusto : `library(pcaPP) do R`
- Independent Component Analysis (ICA) – Comom, P, 1994; Hyvärinen and Oja, 2002, `library(mixomics) do R`
- PCA para dados "heterogêneos" (uma mistura de dados quantitativos e categóricos): Chavent et al., 2014, `PCAmixdata do R`

  Udell, M., 2015, `GLRM implementado usando Julia`
- PCA Multinomial – GLM-PCA (Collins et al., 2002)

Redução de Dimensionalidade em $\Re^p$
Quociente de Rayleigh

Veja também:

Seja M uma matriz simétrica em $\Re^{pxp}$, com autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p$ e os correspondentes autovetores $V_1, V_2, \ldots, V_p$. Então:

$$\max_{\|a\|=1} a'Ma = \max_{a\neq 0} \boxed{\frac{a'Ma}{a'a}} = \lambda_1; \quad a = V_1 \in \Re^p$$

$$\min_{\|a\|=1} a'Ma = \min_{a\neq 0} \boxed{\frac{a'Ma}{a'a}} = \lambda_p; \quad a = V_p \in \Re^p$$

A redução de dimensionalidade pode ser formulada como um problema de otimização de formas quadráticas, cuja solução está na teoria de decomposição espectral de matrizes simétricas $\Re^{pxp}$.

Veremos equivalências de soluções nos espaços Duais: $\Re^{pxp}$, $\Re^{nxn}$ e $\Re^{nxp}$.