MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Sem/2019 - IME

Análise Multivariada

Já vimos ☺

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

- Revisão de Metodologias Clássicas: Foco na obtenção de vetores reducionistas

$$\Re^{p} \rightarrow \Re^{m}$$

Combinações lineares de Y

- Componentes Principais ($m \leq min(n,p)$)
- Análise Fatorial Exploratória (via CP: $m \leq min(n,p)$)
- Coordenadas Principais – Escalonamento Multidimensional ($m \leq posto(D_{nxn})$)
- Análise de Correspondência
- Análise Discriminante ($m \leq min(n,p, G-1)$)
- Correlação Canônica ($m \leq min(n,p,q)$)

$n > p$
Observações Independentes

$\Rightarrow$ Problema Dual de redução de dimensionalidade ($\Re^{nxp}$, $\Re^{pxp}$, $\Re^{nxn}$)
$\Rightarrow$ Representações Bi-Plot
$\Rightarrow$ Problema n << p (*Big-n*)

Componentes Principais (Pearson, 1901)

$$Y_{i\,p\times 1} \overset{iid}{\sim} (\mu\,;\Sigma\,)$$

Unidades Amostrais	Variáveis 1	2	…	j	…	p
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
…	…	…	…	…	…	…
n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

Redução de dimensionalidade:

$$\boxed{\Re^p \rightarrow \Re^m; \quad m \le \min(n, p)}$$

$$Y_{n\times p} \rightarrow Z_{n\times m};\ Z_{ji} = a_j\,'Y_i$$

$$a_j\ ; \quad \arg\max_{\|a_j\|=1} \sum_{j=1}^{m} Var\left(Z_{ji}\right), \quad \text{tr}(\Sigma)$$

$$Cov\left(Z_j^i; Z_{j'}^i\right) = 0$$

Solução: Decomposição Espectral em $\Re^{pxp}$

$$\Sigma = P\Lambda P'; \quad \Lambda = D_{\lambda_j} \quad PP' = P'P = I; \quad \left|\Sigma - \lambda I_p\right| = 0; \quad \Sigma P_j = \lambda_j P_j$$

$$\Sigma_Y = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22}^2 & \dots & \sigma_{2p} \\ & & \dots & \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp}^2 \end{pmatrix} = \boxed{\lambda_1 P_1 P_1' + \dots + \lambda_m P_m P_m'} + \dots + \lambda_p P_p P_p'$$

$$\arg\max_{\|a\|=1} f(\Sigma; a) = f(\Sigma; P_1) = \frac{P_1'\Sigma P_1}{P_1' P_1} = \lambda_1;$$

$$\boxed{f(\Sigma; P_1) \ge \dots \ge f(\Sigma; P_m)} \ge \dots \ge f(\Sigma; P_p)$$

Análise Fatorial Exploratória (Spearman, 1904)

Redução de dimensionalidade $\Rightarrow$ Descrever a estrutura de dependência de "p" variáveis por meio de "m" fatores comuns além de termos específicos:

$$\Re^p \rightarrow \Re^m; \quad m \leq \min(n, p)$$ (Solução via CP)

$$Y_{n\times p}; \begin{cases} Y_1^i - \mu_1 = \phi_{11}F_{i1} + \phi_{12}F_{i2} + \ldots + \phi_{1m}F_{im} + e_{i1} \\ Y_2^i - \mu_2 = \phi_{21}F_{i1} + \phi_{22}F_{i2} + \ldots + \phi_{2m}F_{im} + e_{i2} \\ \ldots \\ Y_p^i - \mu_p = \phi_{p1}F_{i1} + \phi_{p2}F_{i2} + \ldots + \phi_{pm}F_{im} + e_{ip} \end{cases}$$

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

$$\mathbf{f}_{i\,m\times 1} \overset{iid}{\sim} (0; I_m) \perp e_{i\,p\times 1} \overset{iid}{\sim} (0; \Psi)$$

$$\Leftrightarrow \boxed{\Sigma_{p\times p} \cong \underbrace{\Phi_{p\times m}\Phi'_{m\times p}}_{\text{Comunalidade}} + \underbrace{\Psi_{p\times p}}_{\text{Especificidade (matriz diagonal)}}}$$

Solução via Componentes Principais

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

Ψ : Matriz (diagonal) de fatores específicos

$\mathbf{f} = (F_1, \ldots, F_m)'$: Vetor de fatores comuns/ variáveis latentes

$$\Sigma = P\Lambda P' = \left(P\Lambda^{1/2}\right)\left(P\Lambda^{1/2}\right)'$$

$$\Rightarrow Z_{ip\times 1} = P'Y_i \Rightarrow Y_i = PZ_i$$

$$\Phi = P\Lambda^{1/2}, \quad \Psi = Diag\left(\sigma_j^2 - \sum_{k=1}^{m}\varphi_{jk}\right)$$

$$\boxed{\mathbf{f}_{i\,m\times 1} = D_{\lambda_j}^{-1/2} Z_{i\,m\times 1}}$$ Fatores comuns: CP padronizados

Análise Discriminante (Fisher, 1938)

$$\Re^{(p+1)} \rightarrow \Re^{m};\ m \leq \min(n, p; G-1)$$

Obter combinações lineares para a máxima separação dos grupos:

$$Y_{n\times(p+1)}$$

p variáveis mais grupo

	Unidades Amostrais	Variáveis 1	2	…	j	…	p
Grupo1	1	Y_{111}	Y_{112}		Y_{11j}		Y_{11p}
	2	Y_{121}	Y_{122}		Y_{12j}		Y_{12p}
	…	…	…	…	…		…
	n_1	Y_{1n11}	Y_{1n12}		Y_{1n1j}		Y_{1n1p}
Grupo2	1	Y_{211}	Y_{212}		Y_{21j}		Y_{21p}
	2	Y_{221}	Y_{222}		Y_{22j}		Y_{22p}
	…	…	…	…	…		…
	n_2	Y_{2n21}	Y_{2n22}		Y_{2n2j}		Y_{2n2p}

... G grupos

$$Y_{i\ p\times 1} \mid \tau_g \overset{iid}{\sim} \left(\mu_g; \Sigma_g\right) \quad \Rightarrow X_i = l' Y_i \overset{iid}{\sim} \left(l'\mu_g; l'\Sigma l\right)$$

Suposição $\Rightarrow \quad \Sigma_g = \Sigma$

Solução (linear) de Fisher:

$$\frac{l'\sum_{g=1}^{G}\left(\mu_g - \mu_.\right)\left(\mu_g - \mu_.\right)' l}{l'\Sigma l} = \boxed{\frac{l'\Sigma_b l}{l'\Sigma l}}$$

Situação ideal para discriminação: variáveis com covariâncias ENTRE e DENTRO de sinais contrários!

$$\Sigma_T = \Sigma_b + \Sigma_w = \Sigma$$

Correlação Canônica (Hotelling, 1935)

Obter combinações lineares de Y_1 e de Y_2 com máxima correlação

	Variáveis Y_1				Variáveis Y_2		
Unidades Amostrais	Y_{11}	…	Y_{1p}		Y_{21}	…	Y_{2q}
1	Y_{111}		Y_{1p1}		Y_{211}		Y_{2q1}
…	…	…	…		…	…	…
n	Y_{11n1}		Y_{1pnp}		Y_{21n1}		Y_{2qnq}

$$\Re^{p+q} \rightarrow \Re^{m};$$

$$m \leq \min(n, p, q)$$

$$\boldsymbol{\Sigma}_{11\,p\times p} \qquad \boldsymbol{\Sigma}_{22\,p\times p}$$

$$\boldsymbol{\Sigma}_{12\,p\times q}$$

$$Y_{i\,(p+q)\times 1} \overset{iid}{\sim} \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

Matriz de covariância entre os dois conjuntos de variáveis

$$U = a' Y_1$$

$$V = b' Y_2$$

$$\Rightarrow \rho_c(U,V) = \frac{Cov(U,V)}{\sqrt{Var(U)}\sqrt{Var(V)}} = \frac{a'\,\boldsymbol{\Sigma}_{12}\,b}{\sqrt{a'\,\boldsymbol{\Sigma}_{11}\,a}\sqrt{b'\,\boldsymbol{\Sigma}_{22}\,b}}$$

a, b; máxima correlação

$$\Rightarrow \frac{a'\,\boldsymbol{\Sigma}_{12}\Sigma_{22}^{-1}\,\Sigma_{21}a}{a'\Sigma_{11}a}$$

$$\Rightarrow \frac{b'\,\boldsymbol{\Sigma}_{21}\Sigma_{11}^{-1}\,\Sigma_{12}b}{b'\Sigma_{22}b}$$

Redução de Dimensionalidade
Obtenção de Vetores Reducionistas

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

- Componentes Principais: $f(\Sigma;a) = \boxed{\dfrac{a'\Sigma a}{a'a}, \quad a'a = 1} \Rightarrow Z_{ki} = a_k{}'Y_i \qquad Cov(Y) = \Sigma$

- Análise Fatorial Exploratória (via CP): $\Rightarrow F_{ki} = \lambda^{-1/2} Z_{ki} \qquad \Sigma = \Phi\Phi' + Diag\left(\Psi_{jj}\right)$

- Análise Discriminante (Linear de Fisher): $f\left(\Sigma_w^{-1}\Sigma_b;a\right) = \boxed{\dfrac{a'\Sigma_b a}{a'\Sigma_w a}, \quad a'\Sigma_w a = 1} \qquad \Sigma = \Sigma_b + \Sigma_w$

$$Y_{n\times p}; n = \sum n_g$$

- Análise de Correlação Canônica:

$$Y_{i(p+q)\times 1} = \begin{pmatrix} Y_{1i\,p\times 1} \\ Y_{2i\,q\times 1} \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

$$\begin{cases} f_1\left(\Sigma_{11}^{-1}\boldsymbol{\Sigma}_{12}\Sigma_{22}^{-1}\,\Sigma_{21};a\right) = \boxed{\dfrac{a'\boldsymbol{\Sigma}_{12}\Sigma_{22}^{-1}\,\Sigma_{21}a}{a'\Sigma_{11}a}, \quad a'\Sigma_{11}a = 1} \\[2ex] f_2\left(\Sigma_{22}^{-1}\boldsymbol{\Sigma}_{21}\Sigma_{11}^{-1}\,\Sigma_{12};b\right) = \boxed{\dfrac{b'\boldsymbol{\Sigma}_{21}\Sigma_{11}^{-1}\,\Sigma_{12}b}{b'\Sigma_{22}b}, \quad b'\Sigma_{22}b = 1} \end{cases}$$

Redução de Dimensionalidade
Apoio do R

- eigen(S) : recebe uma matriz da forma quadrática a ser analisada ($\Re^{pxp}$ ou $\Re^{nxn}$)

- princomp(Y): recebe Y_{nxp} e realiza a decomposição espectral de R ou S (com divisor n)

- prcomp(Y) : recebe Y_{nxp} e realiza a decomposição espectral de R ou S (com divisor n-1)
  $\rightarrow$ suporta n<p

- svd(Y): recebe Y_{nxp} (n<p, n>p) e realiza a decomposição em valores singulares de $\Re^{pxp}$ e $\Re^{nxn}$.

  Para comparar com *eigen* é preciso "padronizar" autovalores: $\lambda_{eigen} = \left(\lambda_{svd} \Big/ \sqrt{n-1} \right)^2$

- cmdscale: recebe a matriz de distâncias D ou Similaridade entre observações e realiza a Análise de Escalonamento Multidimensional (Análise de Coordenadas Principais)
  Ver também os pacotes do R: "sammon" e "isoMDS"

- ca: realiza a Análise de Correspondência.

- factanal: recebe Y_{nxp} ou R e realiza a Análise Fatorial Exploratória, solução por MVS.

- lda: recebe as (p+1)-variáveis e realiza a Análise Discriminante (solução geral)

- cancor(Y_1,Y_2): realiza a Análise de Correlação Canônica

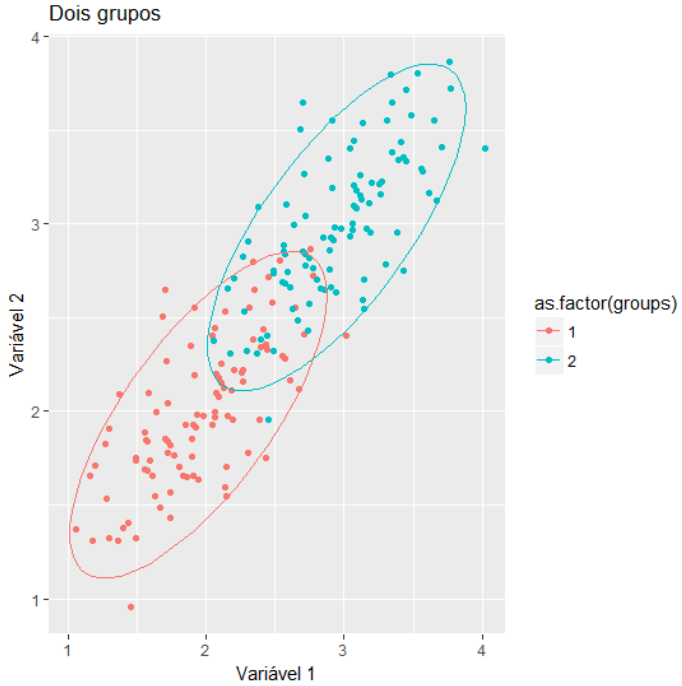
Onde estão os Vetores Reducionistas?

Um gráfico pode valer mais que mil palavras mas pode exigir milhares de palavras para construí-lo. Tukey

Obter a direção do CP e do Eixo Discriminante.

Observações independentes. Indicação da elipse de concentração (95%).

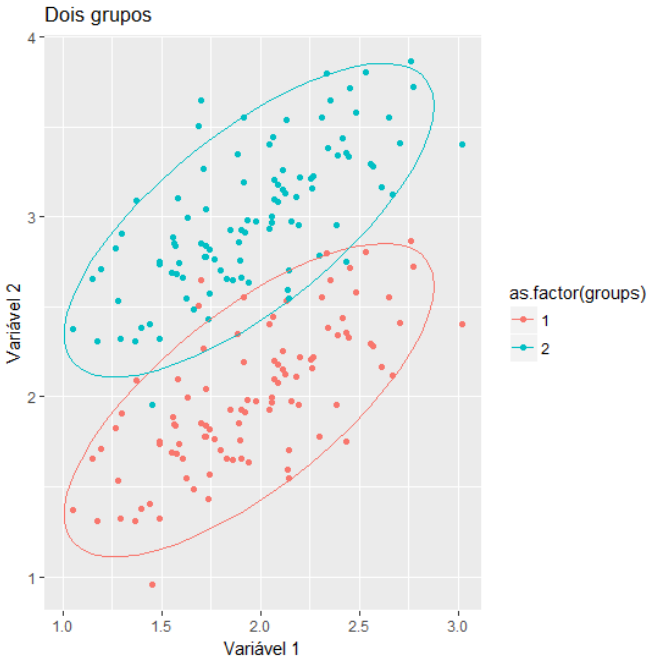
Exemplo 1



$$n = 200,\ p = 2,\ \mu_1 = (2,2),\ \mu_2 = (3,3)$$

$$R_{2\times 2} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}; \sigma = \begin{pmatrix} 0.4 & 0.4 \end{pmatrix}$$

Exemplo 2



$$n = 200,\ p = 2,\ \mu_1 = (2,2),\ \mu_2 = (2,3)$$

$$R_{2\times 2} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}; \sigma = \begin{pmatrix} 0.4 & 0.4 \end{pmatrix}$$

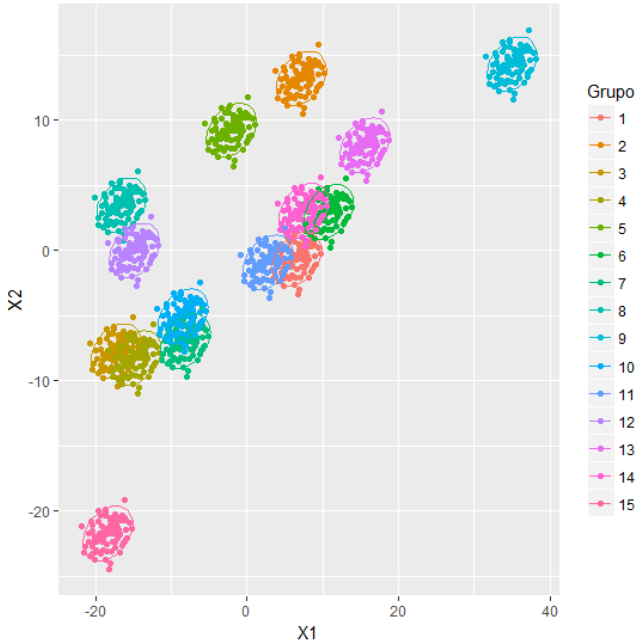
Onde estão os Vetores Reducionistas?

Obter a direção do Eixo Discriminante.

Observações independentes ENTRE e DENTRO de grupos.

Exemplo 3: “Sinais Iguais”

$$T = B + W$$

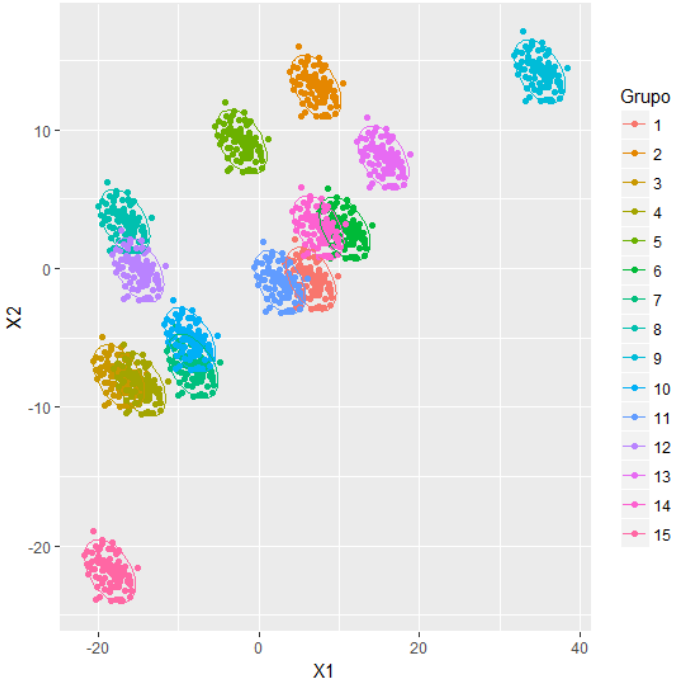


$$G = 15,\ n_g = 100,\ \mu_{.} = (0,0)$$

$$S_b = \begin{pmatrix} 150 & 100 \\ 100 & 150 \end{pmatrix},\ S_w = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

Exemplo 4: “Sinais Opostos”

$$T = B + W$$



$$G = 15,\ n_g = 100,\ \mu_{.} = (0,0)$$

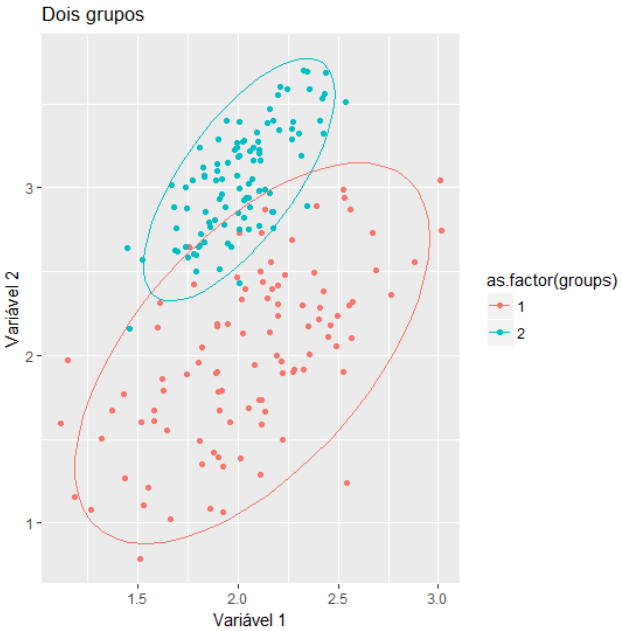
$$S_b = \begin{pmatrix} 150 & 100 \\ 100 & 150 \end{pmatrix},\ S_w = \begin{pmatrix} 2 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

Onde estão os Vetores Reducionistas?

Obter a direção da Variável Canônica.

Observações independentes avaliadas em $\Re^{p+q}$.

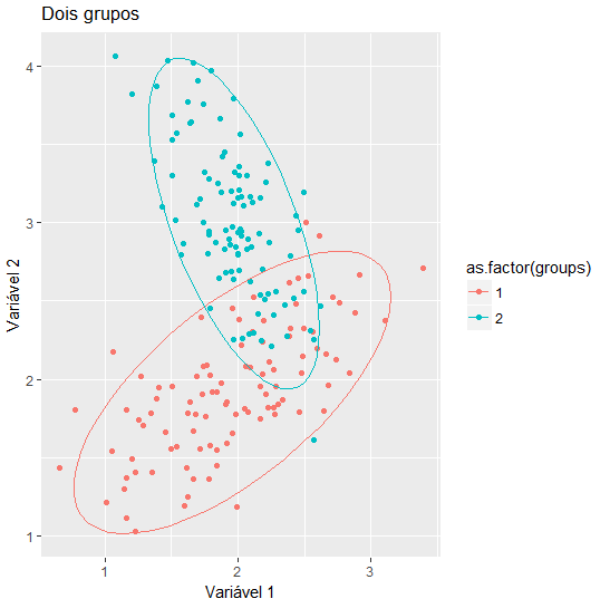
Exemplo 5: Correlações de mesmo sinal



$$R = \begin{pmatrix} 1 & 0.65 & 0.2 & 0.5 \\ 0.65 & 1 & 0.1 & 0.4 \\ 0.2 & 0.1 & 1 & 0.7 \\ 0.5 & 0.4 & 0.7 & 1 \end{pmatrix}$$

$$\sigma_1 = (0.4 \quad 0.5); \sigma_2 = (0.2 \quad 0.3); \quad \mu_1 = (2,2); \mu_2 = (3,3)$$

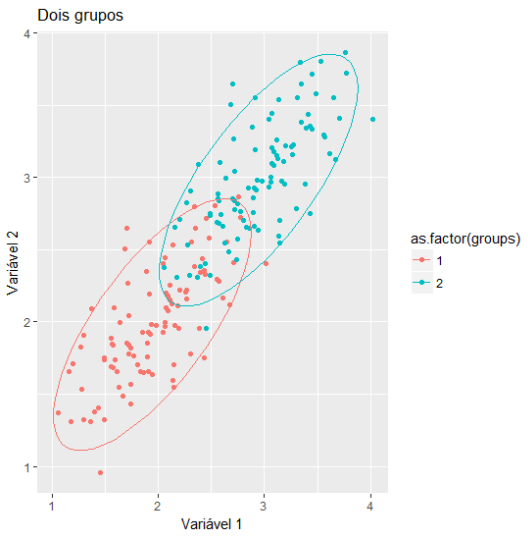
Exemplo 6: Correlações de sinal diferentes



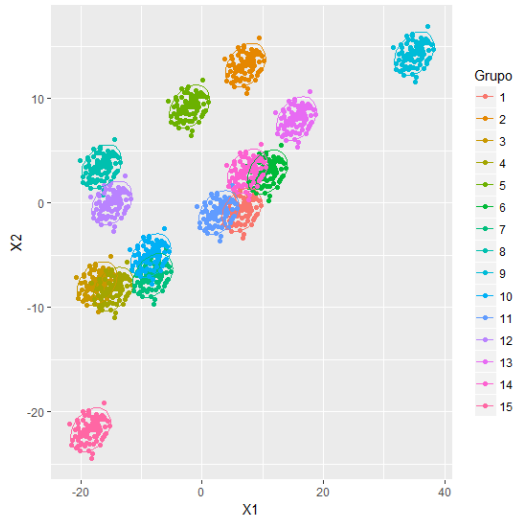
$$R = \begin{pmatrix} 1 & 0.7 & 0.2 & 0.5 \\ 0.7 & 1 & 0.3 & 0.4 \\ 0.2 & 0.3 & 1 & -0.7 \\ 0.5 & 0.4 & -0.7 & 1 \end{pmatrix}$$

$$\sigma_1 = (0.5 \quad 0.4); \sigma_2 = (0.3 \quad 0.5); \quad \mu_1 = (2,2); \mu_2 = (3,3)$$

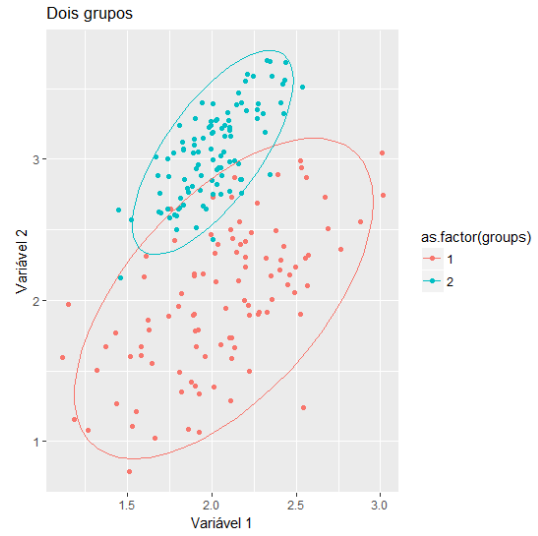
Exemplo 1



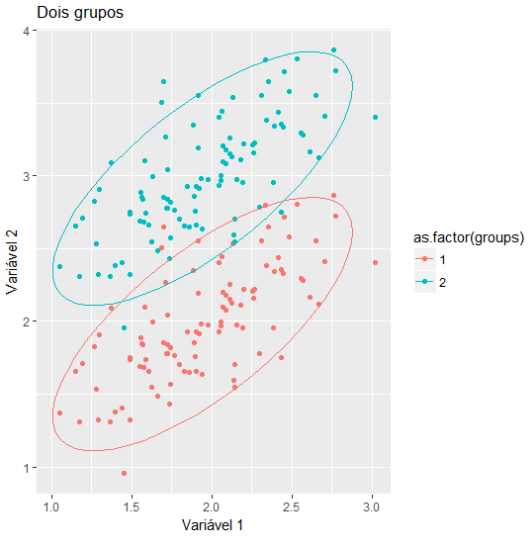
Exemplo 3



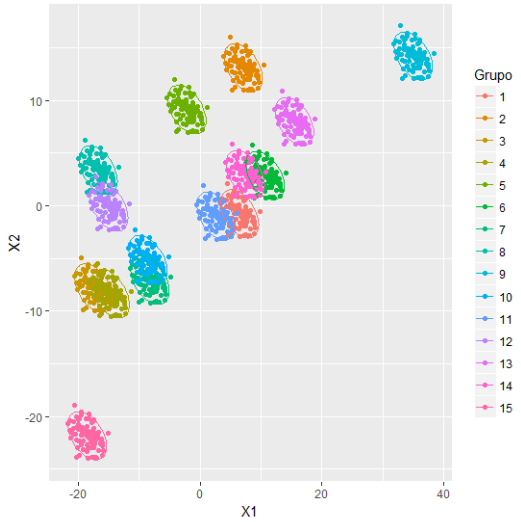
Exemplo 5



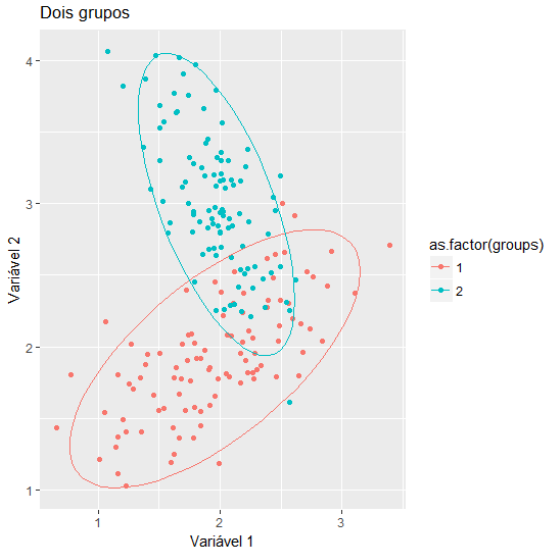
Exemplo 2



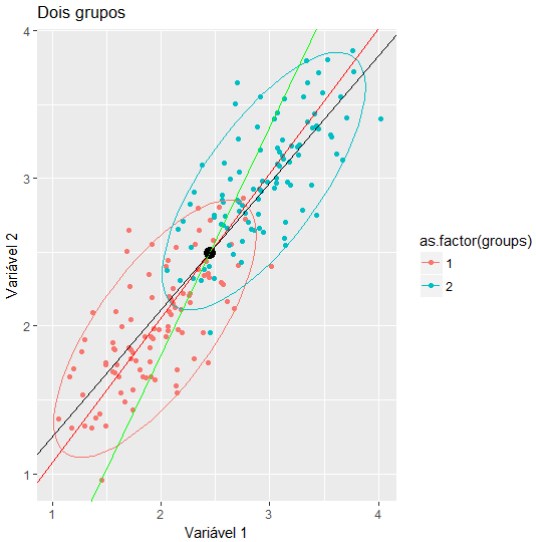
Exemplo 4



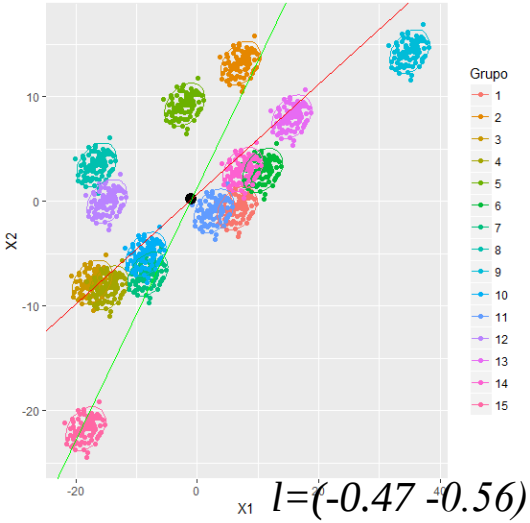
Exemplo 6



Exemplo 1



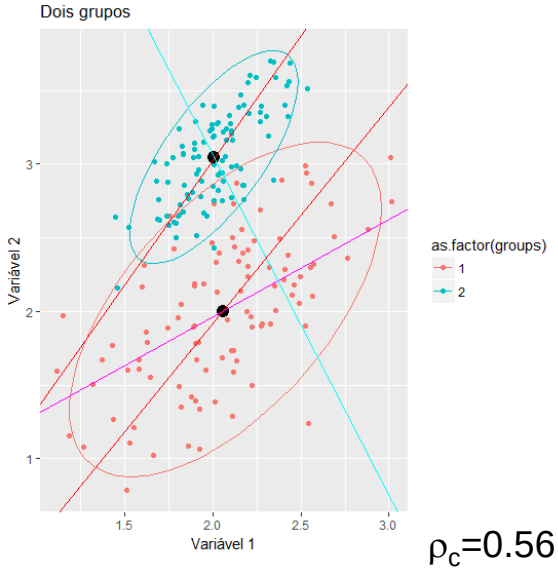
Exemplo 3



l=(-0.47 -0.56)'

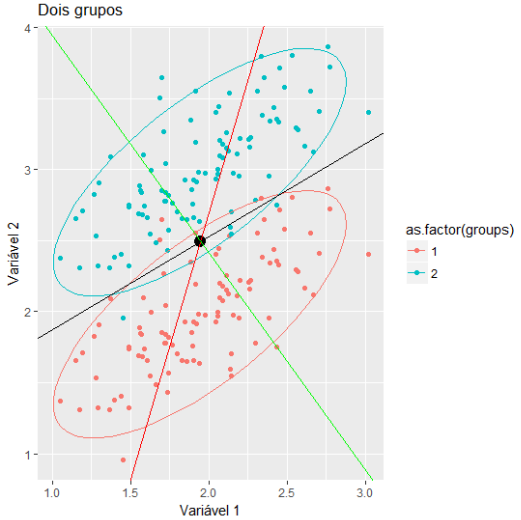
%Expl=0.78

Exemplo 5

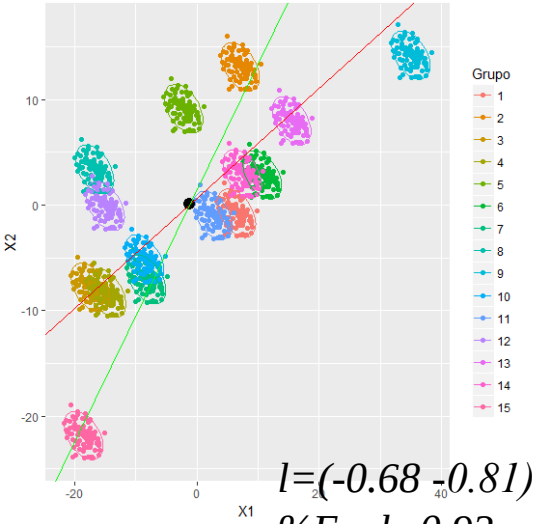


ρ_c=0.56

Exemplo 2



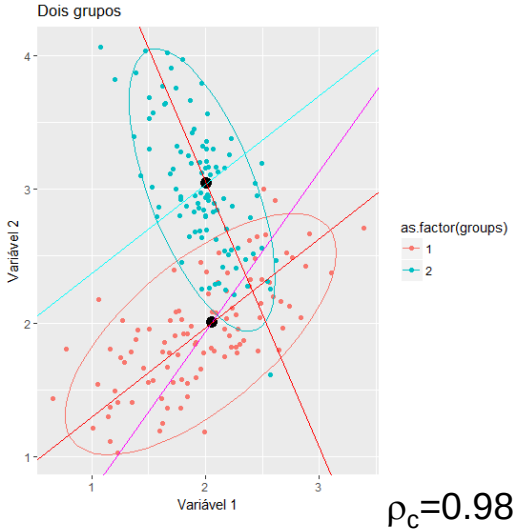
Exemplo 4



l=(-0.68 -0.81)'

%Expl=0.93

Exemplo 6



ρ_c=0.98

Preto:reta de MQ Vermelho:vetor de CP Verde:vetor discriminante Azul e rosa:variáveis canônicas

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: $\Re^{pxp}$

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

r Componentes Principais

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

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

Escalonamento Multidimensional:
r Coordenadas Principais
obtidas da Matriz de Distâncias

Análise no espaço $\Re^{nxp}$

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

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)

Componentes Principais – Coordenadas Principais

Equivalência das Soluções em Espaços Duais

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

HY: linhas de Y centradas

$$(HY)_{n\times p} = U_{n\times n}\begin{pmatrix}\Lambda_r^{1/2} & 0\\ 0 & 0\end{pmatrix}V'_{p\times p}$$

$(n-1)S_u = Y'HY$

Análise em $\Re^{nxn}$ Análise em $\Re^{pxp}$

$$HYY'H = U\Lambda U' \qquad Y'HY = V\Lambda V'$$

$$\boxed{U_{n\times n}\Lambda_m^{1/2}} = \boxed{(HY)_{n\times p}V_{p\times m}} \qquad m\le r$$

Coordenadas Principais

Componentes Principais

Coordenadas Principais

Dados HatCo
7 primeiras observações

	Y1	Y2	Y3	Y4	Y5	Y6	Y7
1	4.1	0.6	6.9	4.7	2.4	2.3	5.2
2	1.8	3.0	6.3	6.6	2.5	4.0	8.4
3	3.4	5.2	5.7	6.0	4.3	2.7	8.2
4	2.7	1.0	7.1	5.9	1.8	2.3	7.8
5	6.0	0.9	9.6	7.8	3.4	4.6	4.5
6	1.9	3.3	7.9	4.8	2.6	1.9	9.7
7	4.6	2.4	9.5	6.6	3.5	4.5	7.6

Matriz de Distância Euclidiana (D)

	1	2	3	4	5	6
2	5.31					
3	6.13	3.62				
4	3.27	3.11	5.16			
5	5.23	7.12	7.86	6.28		
6	5.80	3.47	4.23	3.49	8.33	
7	5.07	4.51	5.32	4.43	3.91	5.09

Escalonamento Multidimensional

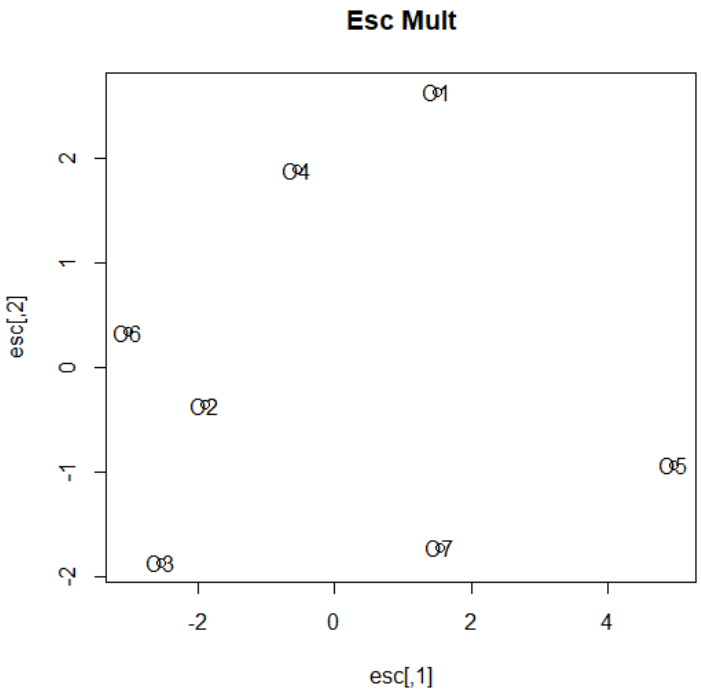
```
> cmdscale(diste)
          [,1]       [,2]
1  1.4981397  2.6360620
2 -1.8968160 -0.3552380
3 -2.5399671 -1.8649995
4 -0.5547189  1.8942258
5  4.9716633 -0.9304384
6 -3.0199393  0.3412407
7  1.5416383 -1.7208525
```

As colunas da matriz X, obtida de D, são denominadas Coordenadas Principais de uma escala construída aos dados

Coordenadas Principais

Representação de Dados a partir de uma Matriz de Distâncias – Dados HatCo

	[,1]	[,2]
1	1.4981397	2.6360620
2	-1.8968160	-0.3552380
3	-2.5399671	-1.8649995
4	-0.5547189	1.8942258
5	4.9716633	-0.9304384
6	-3.0199393	0.3412407
7	1.5416383	-1.7208525



Matriz de Distância Euclidiana Daddos Originais

	1	2	3	4	5	6
2	5.31					
3	6.13	3.62				
4	3.27	3.11	5.16			
5	5.23	7.12	7.86	6.28		
6	5.80	3.47	4.23	3.49	8.33	
7	5.07	4.51	5.32	4.43	3.91	5.09

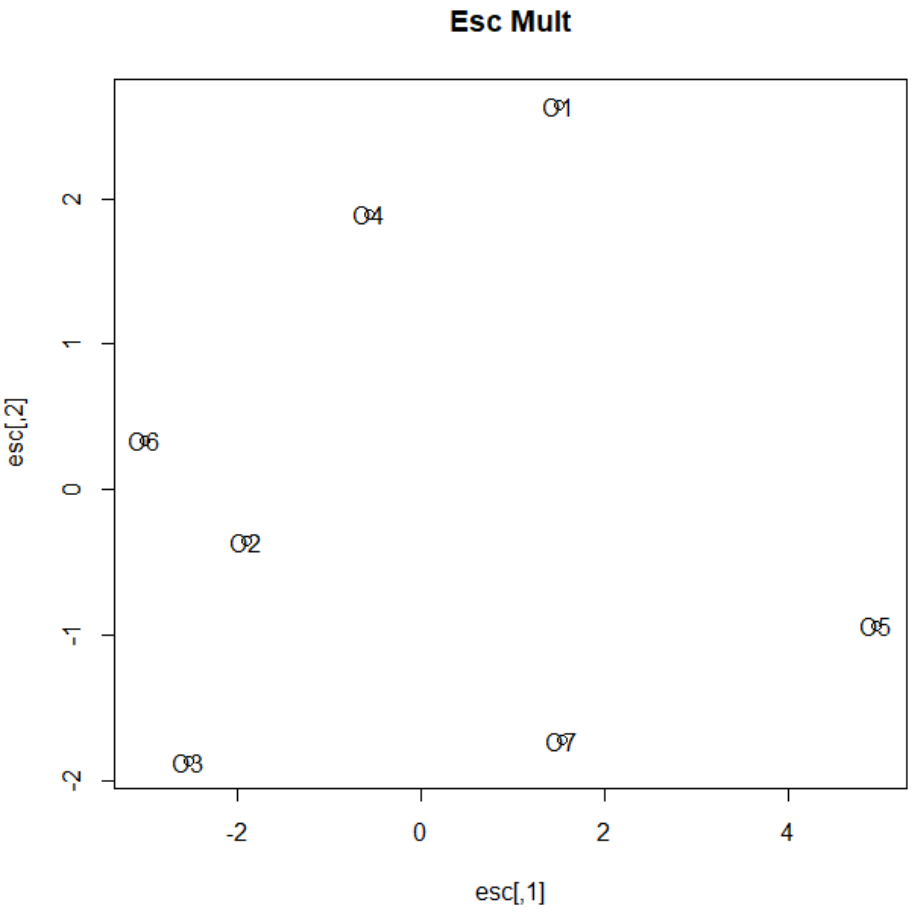
Avaliar a qualidade da representação em $\Re^2$

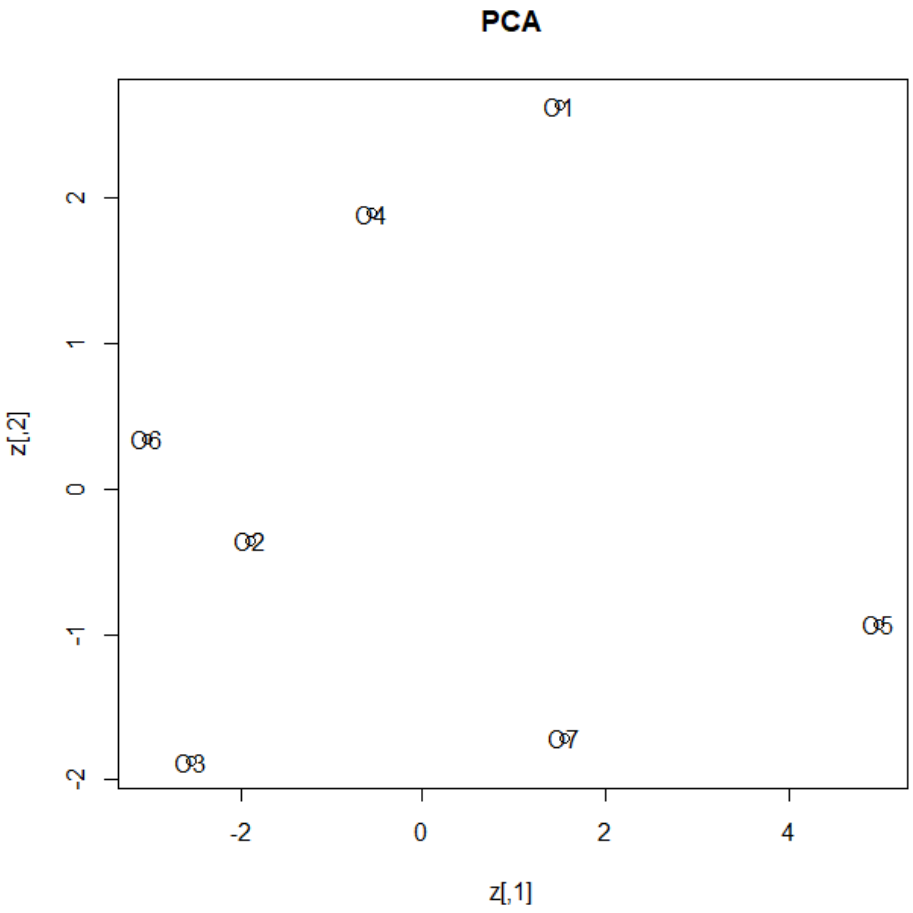
Matriz de Distância Euclidiana das Coordenadas Principais

	1	2	3	4	5	6
2	4.36					
3	5.23	3.73				
4	2.39	3.14	5.01			
5	4.93	5.30	6.34	5.49		
6	4.10	3.47	3.81	2.86	6.81	
7	4.62	3.54	4.56	4.48	2.77	4.81

Componentes Principais e Coordenadas Principais Equivalência

Coordenadas Principais extraídas de uma Matriz de Distância Euclidiana
$\Rightarrow$ Representação equivalente aos Componentes Principais em $\Re^2$





Biplots

Biplot: representação gráfica simultânea de n observações e p variáveis em $\Re^2$

$$Y_{n\times p} = \boxed{U_{n\times n}\ \Lambda^{1/2}}\ \boxed{V'_{p\times p}}$$

Matriz de "escores dos CP"

Matriz de "pesos"

$$\begin{cases} YY' = U\Lambda U' & \text{Análise em } \Re^{nxn} \\ YY' = U\Lambda U' & \text{Análise em } \Re^{pxp} \end{cases}$$

$\Re^p \rightarrow \Re^m$

$m=2$

$$Y_{n\times p};\quad Y \approx \left[U_1\ U_2\right]_{n\times 2} \Lambda_2^{1/2} \left[V_1\ V_2\right]'_{2\times n} = \left[U_1\ U_2\right] \Lambda_2^{1/2-c/2+c/2} \left[V_1\ V_2\right]'$$

$$Y \approx \left(U_1\lambda_1^{1/2-c/2}\quad U_2\lambda_2^{1/2-c/2}\right)\left(\lambda_1^{c/2}V_1\quad \lambda_2^{c/2}V_2\right)'$$

Análises sob os mesmos autovalores

$U_1\lambda_1^{1/2-c/2} \times U_2\lambda_2^{1/2-c/2}$ n pontos

$\lambda_1^{c/2}V_1 \times \lambda_2^{c/2}V_2$ p pontos

c=0: linhas em coordenadas principais e colunas em coordenadas padronizadas

c=1: linhas em coordenadas padronizadas e colunas em coordenadas principais

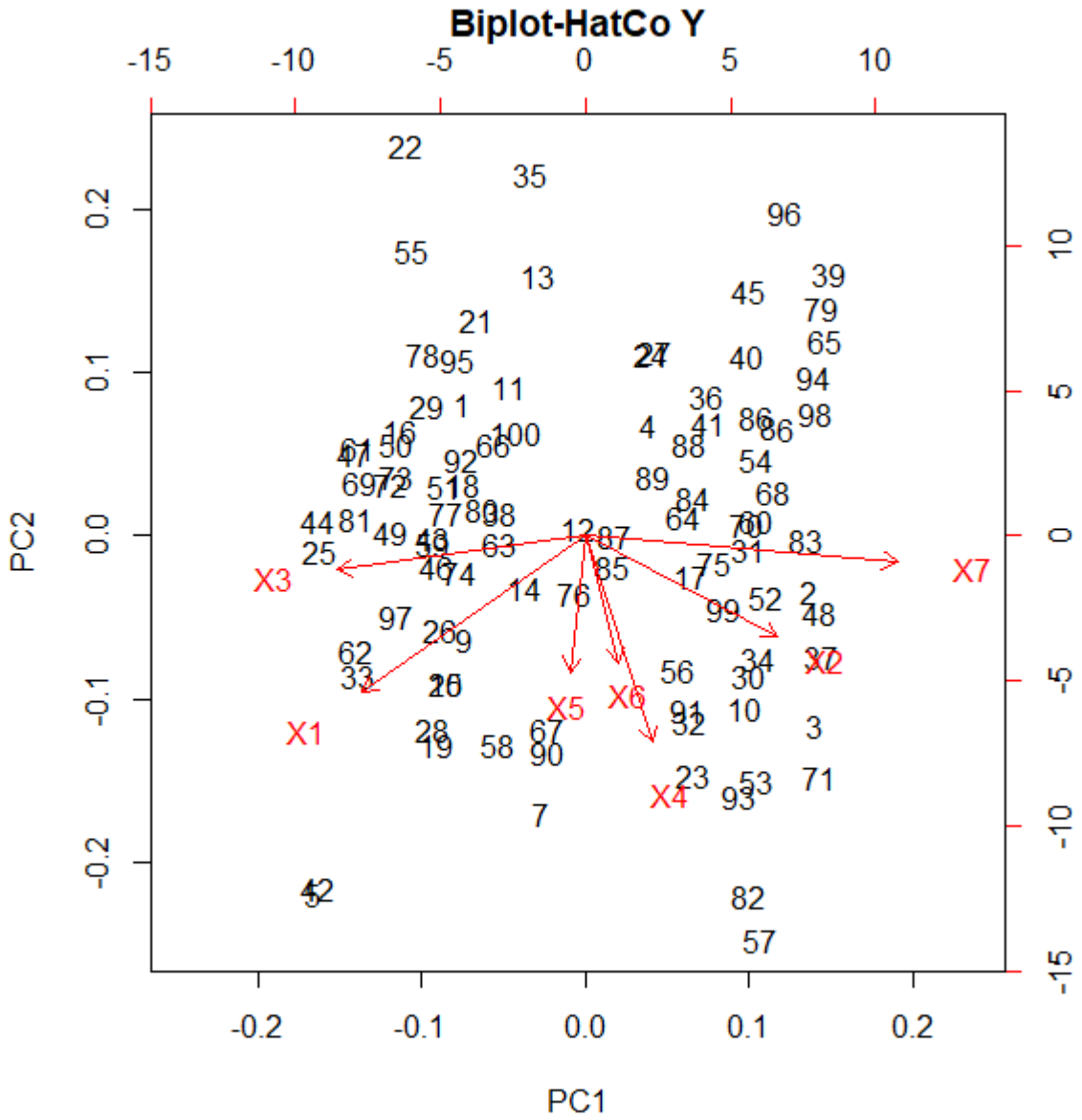
c=1/2: representação interativa

Representação Biplot

Dados HatCo (Y)
n =100
p = 7
Variância Total=10,048

Análise de Componentes principais

	PC1	PC2
Desvio Padrão	2.1627	1.4567
VarExpl	0.4655	0.2112
VarExpl_Acumul	0.4655	0.6767

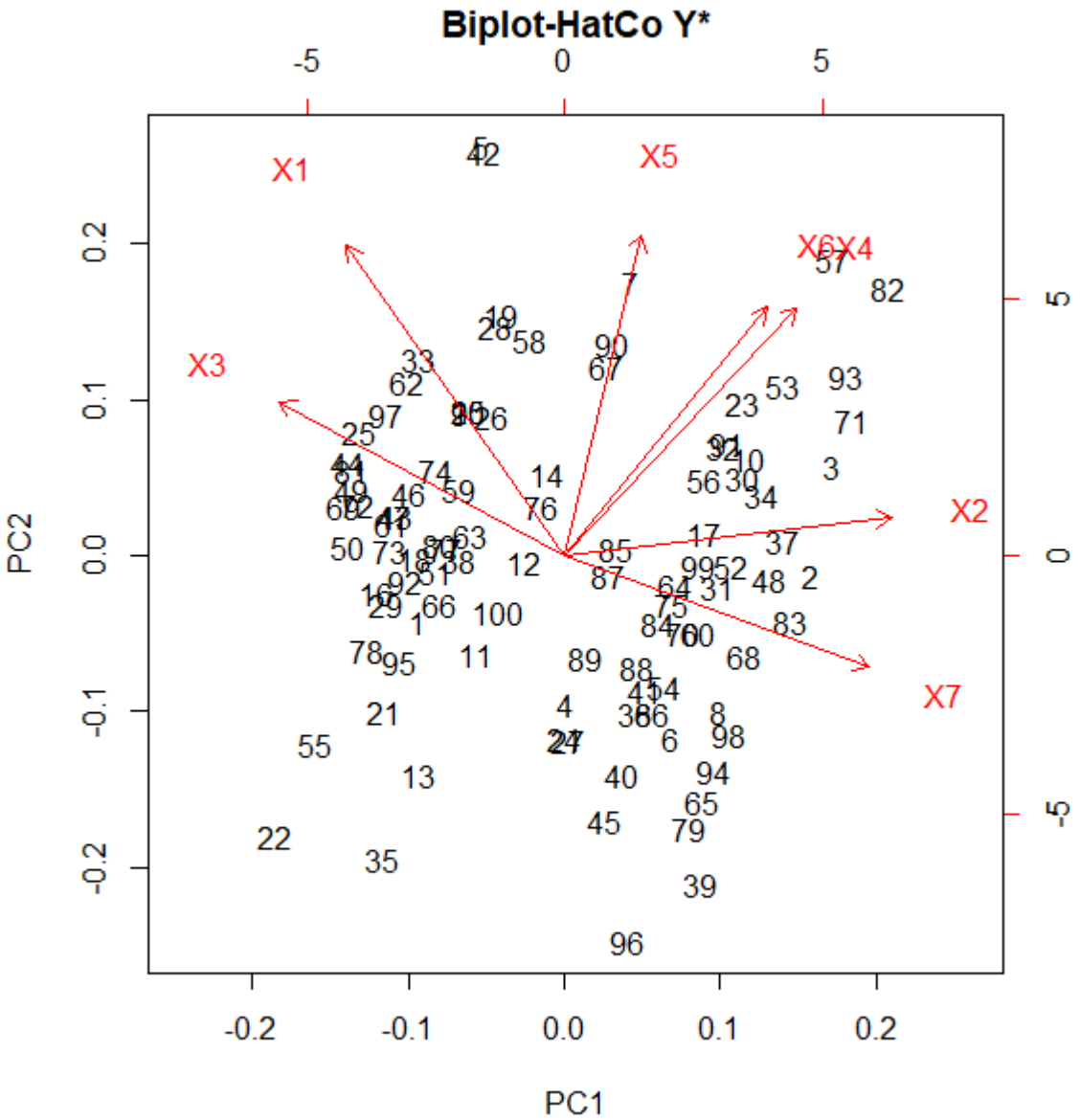


Representação Biplot

Dados HatCo
Dados Padronizados (Y*)
n =100
p = 7
Variância Total=7

Análise de Componentes principais:

	PC1	PC2
Desvio Padrão	1.5893	1.4562
VarExpl	0.3608	0.3029
VarExplAcuml	0.3608	0.6637



Componentes Principais – n << p

Big data
Big-p

$$Y_{n\times p} = U_{n\times n}\begin{pmatrix}\Lambda_m^{1/2} & 0\\ 0 & 0\end{pmatrix}V'_{p\times p} \quad \text{Equivalência} \quad \Leftrightarrow \quad \boxed{Y_{n\times p}V_{p\times m} = U_{n\times n}\Lambda_m^{1/2}}$$

CP CoP

m ≤ min(n,p) $Y_{n\times p} \approx U_{n\times m}\Lambda_m^{1/2}V'_{m\times p}$

Os Componentes Principais podem ser obtidos da análise em $\Re^{pxp}$ ou $\Re^{nxn}$.

Para n > p: realizar a análise em $\Re^{pxp}$ (Decomposição de S: PCA clássico)

Para n < p: realizar a análise em $\Re^{nxn}$ (Decomposição da matriz B obtida de D: Escalonamento Multidimensional ou Coordenadas Principais)

Para n << p: há interessante em soluções penalizadas para eliminar variáveis redundantes da análise, isto é, obter autovetores V que atribuem carga nula a algumas variáveis.

Componentes Principais – n << p

Big data
Big-p

É facilmente obtido

$$Y_{n\times p} = U_{n\times n}\begin{pmatrix}\Lambda_m^{1/2} & 0\\ 0 & 0\end{pmatrix}V'_{p\times p} \quad \text{Equivalência} \quad \Leftrightarrow \quad Y_{n\times p}V_{p\times m} = U_{n\times n}\Lambda_m^{1/2}$$

$$Y = U\Lambda^{1/2}V'; \quad Z_j = Y V_j \overset{n<<p}{\Rightarrow} \quad Z_j = U_j d_j^{1/2}$$

j-ésimo Componente Principal de $\Re^{nxn}$.
Considere: $\Lambda=(dj)$

Como obter os autovetores tais que, $V_j \cong \nu_j = \left(\nu_{1j}, \nu_{2j}, \ldots, \nu_{pj}\right); \nu_{kj} = 0$ para muitas coordenadas k mantendo uma alta porcentagem da variância total explicada ?

Primeiro, é interessante estabelecer a CORESPONDÊNCIA entre CP e Modelos de Regressão (Zou, Hastie and Tibshirani, 2006):

CP Regularizado

$$\hat{\beta} = \arg\min_{\beta} \left\|Z_j - Y\beta\right\|_2^2 + \lambda \left\|\beta\right\|_2^2; \quad \lambda > 0, \ \hat{v}_j = \frac{\hat{\beta}}{\left\|\hat{\beta}\right\|_2}; \quad \hat{v}_j = V_j$$

CP conhecido
(obtido da análise em $\Re^{nxn}$)

Solução regularizada
(não depende de λ)

$$n > p: \ \lambda = 0; \ \hat{v}_j = V_j$$

Componentes Principais – n << p

Correspondência entre CP e Modelos de Regressão (Zou, Hastie and Tibshirani, 2006)

- Componente Principal Regularizado (Ridge Regression)

$$\hat{\beta} = \arg\min_{\beta} \left\| Z_j - Y\beta \right\|_2^2 + \lambda \left\| \beta \right\|_2^2$$

$$\left\| \beta \right\|_2^2 = \sum_{j=1}^{p} \beta_j^2$$

parâmetro de regularização $\begin{cases} \lambda \to 0\text{: solução MQ} \\ \lambda \to \infty\text{: } \beta_j \to 0 \end{cases}$

$$\boxed{\hat{v}_j = \frac{\hat{\beta}}{\left\| \hat{\beta} \right\|_2}; \quad \hat{v}_j = V_j}$$

Solução NÃO depende de λ

- Componente Principal Penalizado (LASSO)

Limitação: o número de "pesos" não-nulos é no máximo *n*

$$\hat{\beta} = \arg\min_{\beta} \left\| Z_j - Y\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1$$

$$\hat{v}_j = \frac{\hat{\beta}}{\left\| \hat{\beta} \right\|_2}; \quad \hat{Z}_j = Y\,\hat{v}_j$$

parâmetro de penalização $\begin{cases} \lambda \to 0\text{: solução MQ} \\ \lambda \to \infty\text{: } \beta_j \to 0 \end{cases}$

$$\left\| \beta \right\|_1 = \sum_{j=1}^{p} \left| \beta_j \right|$$

Distância absoluta (norma L_1) do vetor β à origem

Componentes Principais – n << p

CP Penalizado (LASSO) x CP Regularizado (Ridge Regression)

Penalização na forma de Lagrange:

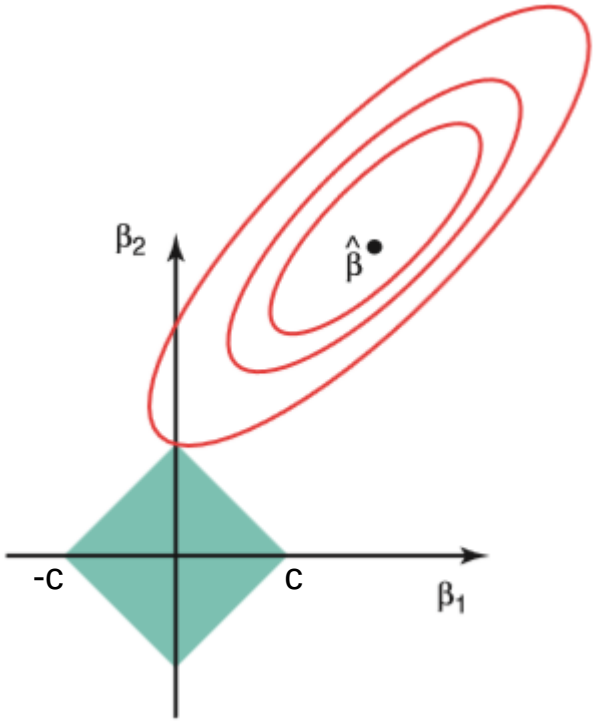
$$\hat{\beta} = \arg\min_{\beta} \left\| Z_j - Y\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1$$

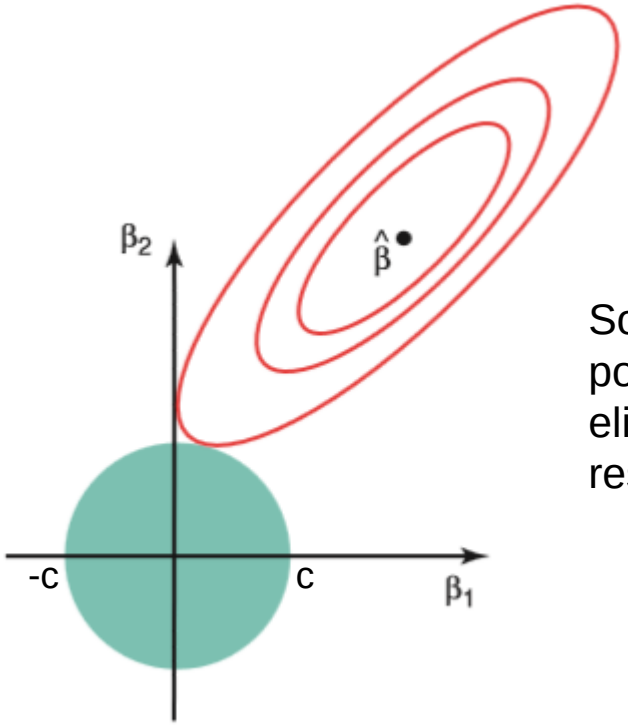
$$\hat{\beta} = \arg\min_{\beta} \left\| Z_j - Y\beta \right\|_2^2 + \lambda \left\| \beta \right\|_2^2$$

Penalização na forma de restrição:

$$\hat{\beta}_{2\times1}; \min_{\beta} \sum_{i=1}^{n} \left(Z_{ij} - Y_i' \beta \right)^2 \text{ sujeito a } \boxed{|\beta_1| + |\beta_2| \le c}$$

$$\hat{\beta}_{2\times1}; \min_{\beta} \sum_{i=1}^{n} \left(Z_{ij} - Y_i' \beta \right)^2 \text{ sujeito a } \boxed{\beta_1^2 + \beta_2^2 \le c}$$





Solução: primeiro ponto em que a elipse intercepta a restrição.

Solução mais esparsa: $\beta_1=0$

Solução menos esparsa: $\beta_1 \approx 0$

Componentes Principais – n << p

Correspondência entre CP e Modelos de Regressão (Zou, Hastie and Tibshirani, 2006)

Vantagem: todas as var podem ser selecionadas (não há limitação no número de "pesos" não-nulos)

Componente Principal Regularizado e Penalizado (Elastic Net):

$$Y_{n\times p} = U\Lambda^{1/2}V' \overset{n<<p}{\Rightarrow} Z_j = U_j d_j^{1/2} \Rightarrow \hat{Z}_j = Y\hat{v}_j$$

$$\hat{\beta} = \arg\min_{\beta}\left\{\left\|Z_j - Y\beta\right\|_2^2 + \boxed{\lambda_1\left\|\beta\right\|_2^2} + \boxed{\lambda_2\left\|\beta\right\|_1}\right\}; \quad \hat{v}_j = \frac{\hat{\beta}}{\left\|\hat{\beta}\right\|_2}; \quad \boxed{\hat{Z}_j = Y\hat{v}_j}$$

Solução de mínimos quadrados

parâmetro de regularização

parâmetro de penalização

Quando $n > p$: $\lambda_1 = \lambda_2 = 0; \boxed{\hat{v}_j = V_j}$

$\lambda_1; \lambda_2$: obtidos por validação-cruzada ou fixados

Componentes Principais – n<<p
Correspondência com Modelos de Regressão

Formalização Geral de Componentes Principais via Modelos de Regressão

$$\arg\min_{\alpha,\beta} \sum_{i=1}^{n} \left\| Y_{i_{p\times 1}} - \alpha_{p\times 1}\beta'_{1\times p} Y_i \right\|_2^2 + \lambda \left\|\beta\right\|_2^2; \quad \left\|\alpha\right\|_2^2 = 1 \quad \Rightarrow \hat{\beta} \propto V_1$$

$$\arg\min_{A,B} \sum_{i=1}^{n} \left\| Y_{i_{p\times 1}} - A_{p\times m} B'_{m\times p} Y_i \right\|_2^2 + \lambda \sum_{j=1}^{m} \left\|\beta_j\right\|_2^2; \quad \lambda > 0, B = \left(\beta_j\right), A'A = I_m$$

$$\Rightarrow \hat{\beta}_j \propto V_j$$

m CP Regularizados e Penalizados podem ser obtidos diretamente de Y:

$$\arg\min_{A,B} \sum_{i=1}^{n} \left\| Y_{i_{p\times 1}} - A_{p\times m} B'_{m\times p} Y_i \right\|_2^2 + \lambda_1 \sum_{j=1}^{m} \left\|\beta_j\right\|_2^2 + \sum_{j=1}^{m} \lambda_{2j} \left\|\beta_j\right\|_1;$$

parâmetro de regularização

m parâmetros de penalização

$$A'A = I_m; B_{p\times m} = (\beta_1, \ldots, \beta_m); \ \hat{v}_j = \frac{\hat{\beta}_j}{\left\|\hat{\beta}_j\right\|_2}; \quad \hat{Z}_j = Y\ \hat{v}_j; j = 1, \ldots, m$$

Componentes Principais – n<<p
Correspondência com Modelos de Regressão

m CP Regularizados e Penalizados podem ser obtidos diretamente de Y:

$$\arg\min_{A,B} \sum_{i=1}^{n} \left\| \boxed{Y_{i_{p\times 1}}} - A_{p\times m} B'_{m\times p}\, Y_i \right\|_2^2 + \lambda_1 \sum_{j=1}^{m} \left\| \beta_j \right\|_2^2 + \sum_{j=1}^{m} \lambda_{2j} \left\| \beta_j \right\|_1;$$

parâmetro de regularização

m parâmetros de penalização

$$A'A = I_m;\ B_{p\times m} = (\beta_1, \ldots, \beta_m);\ \hat{v}_j = \frac{\hat{\beta}_j}{\left\| \hat{\beta}_j \right\|_2};\quad \hat{Z}_j = Y\,\hat{v}_j;\ j = 1, \ldots, m$$

PC Esparso: Variância Explicada (Shen and Huang, 2008)

$$\hat{Z} = \left(\hat{Z}_1, \hat{Z}_2, \ldots, \hat{Z}_m \right);\quad \hat{Z}_j = Y\hat{v}_j$$

$$\hat{V}_{p\times m} = \left(\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_m \right) \Rightarrow tr\left(\hat{Y}'\hat{Y} \right);\quad \hat{Y}_{n\times p} = Y_{n\times p} \hat{V} \left(\hat{V}'\hat{V} \right)^{-1} \hat{V}$$

$$\Rightarrow \frac{tr\left(\hat{Y}'\hat{Y} \right)}{tr\left(Y'Y \right)}$$

Componentes Principais – n<<p

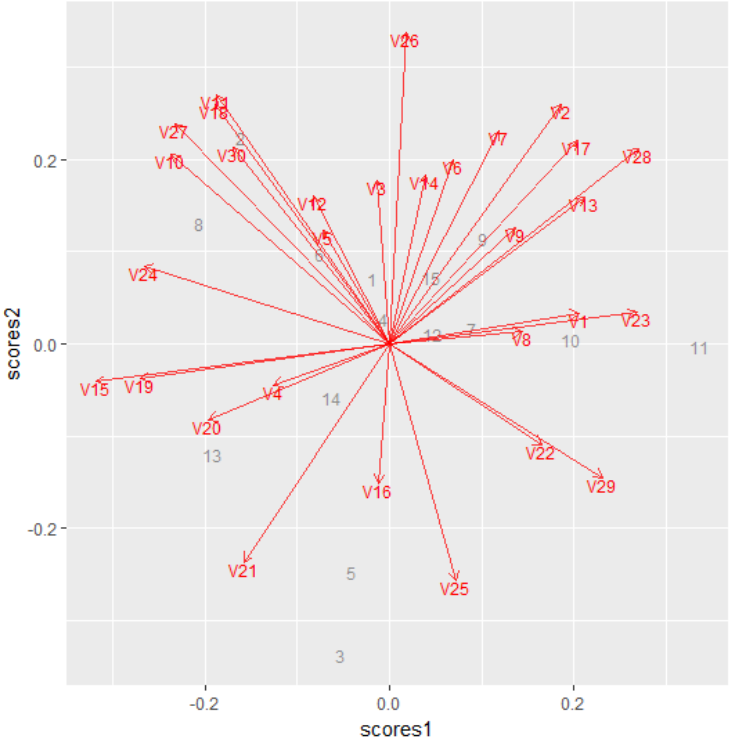
Sparse Principal Component Analysis

Hui ZOU, Trevor HASTIE, and Robert TIBSHIRANI

©2006 American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America

Journal of Computational and Graphical Statistics, Volume 15, Number 2, Pages 265–286
DOI: 10.1198/106186006X113430

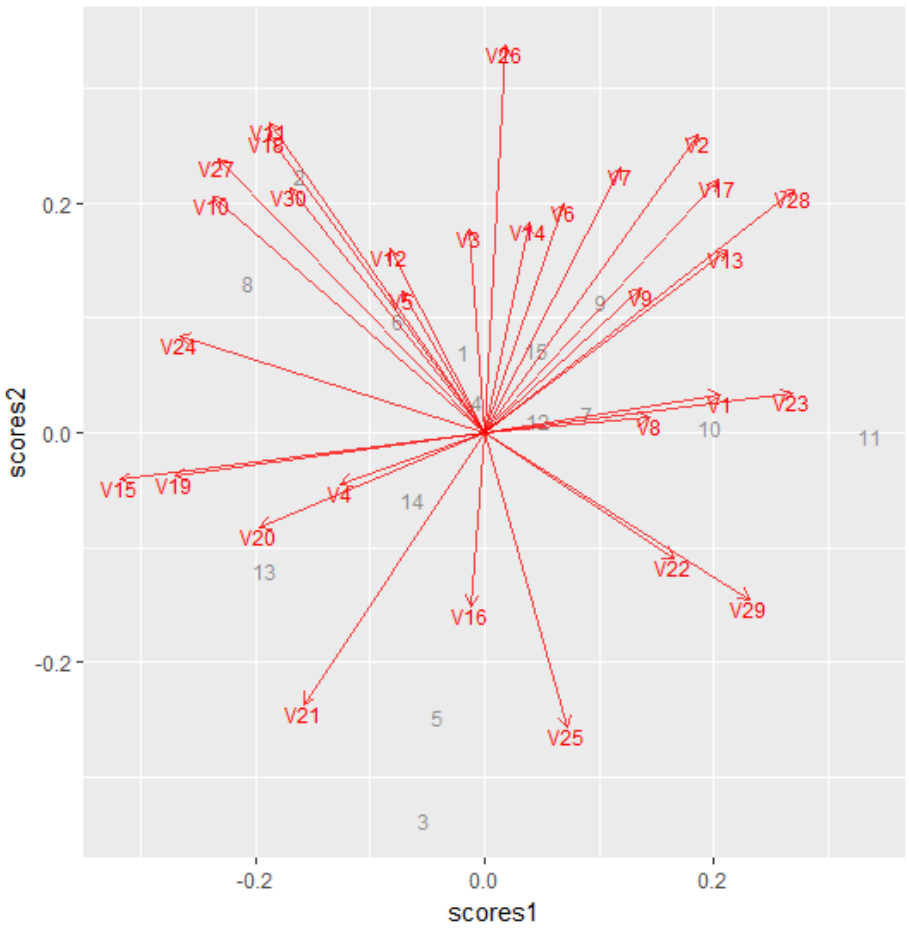
Biplot – (n<p): n=15 p=30
R-prcomp



- *R-SPCA do pacote ElasticNet: Componentes Principais Esparsos*

Solução esparsa: obter autovetores com muitas coordenadas nulas.

Componentes Principais – n<<p

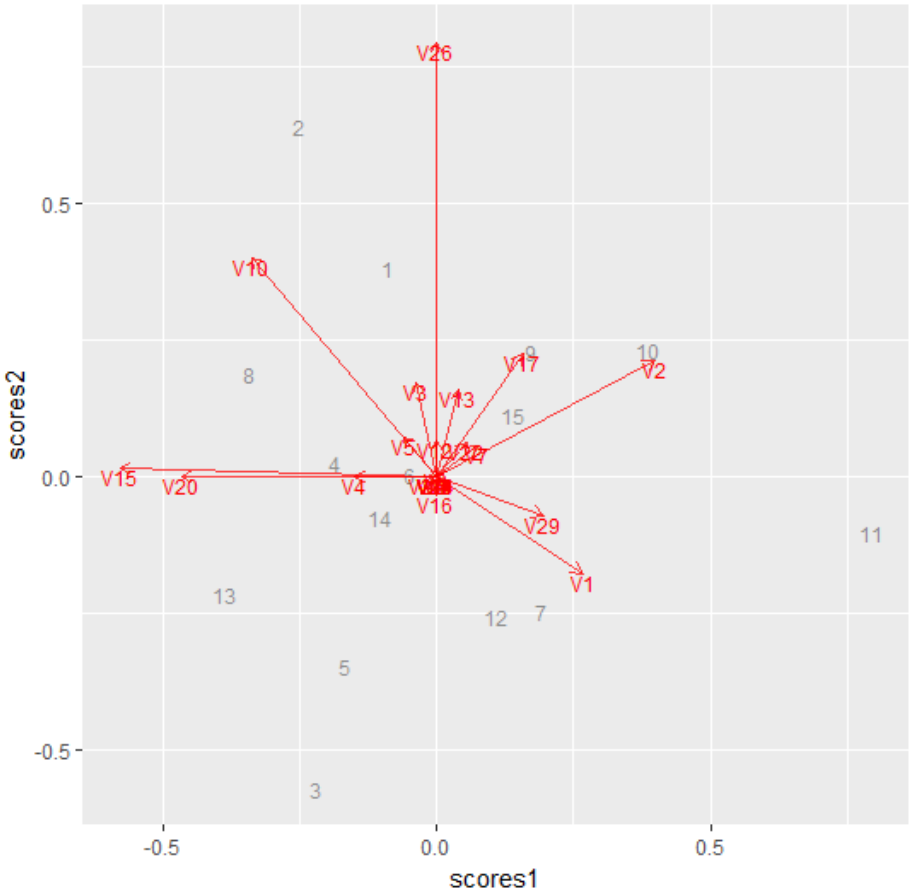


Matriz de pesos

	PC1	PC2
V1	0.20486853	0.03338466
V2	0.18525221	0.26052241
V3	-0.01406721	0.17725332
V4	-0.12560728	-0.04474235
V5	-0.07185638	0.12382278
V6	0.06894920	0.20028957
V7	0.11822653	0.23051465
V8	0.14332703	0.01338871
V9	0.13705858	0.12591629
V10	-0.23708813	0.20630273
V11	-0.18710753	0.26974343
V12	-0.08342832	0.15999298
V13	0.21214752	0.15918079
V14	0.03850244	0.18275759
V15	-0.31794351	-0.04081232
V16	-0.01242316	-0.15170092
V17	0.20361151	0.22032788
V18	-0.18979900	0.25906266
V19	-0.26976250	-0.03717325
V20	-0.19657886	-0.08251878
V21	-0.15850189	-0.23668719
V22	0.16536520	-0.10948335
V23	0.26820473	0.03395738
V24	-0.26656515	0.08480063
V25	0.07236291	-0.25674348
V26	0.01797063	0.33741685
V27	-0.23208636	0.23890523
V28	0.26960763	0.21224456
V29	0.23130349	-0.14483632
V30	-0.16921162	0.21346408

Componentes Principais – n<<p

Biplot – CP Esparsos: n=15 p=30



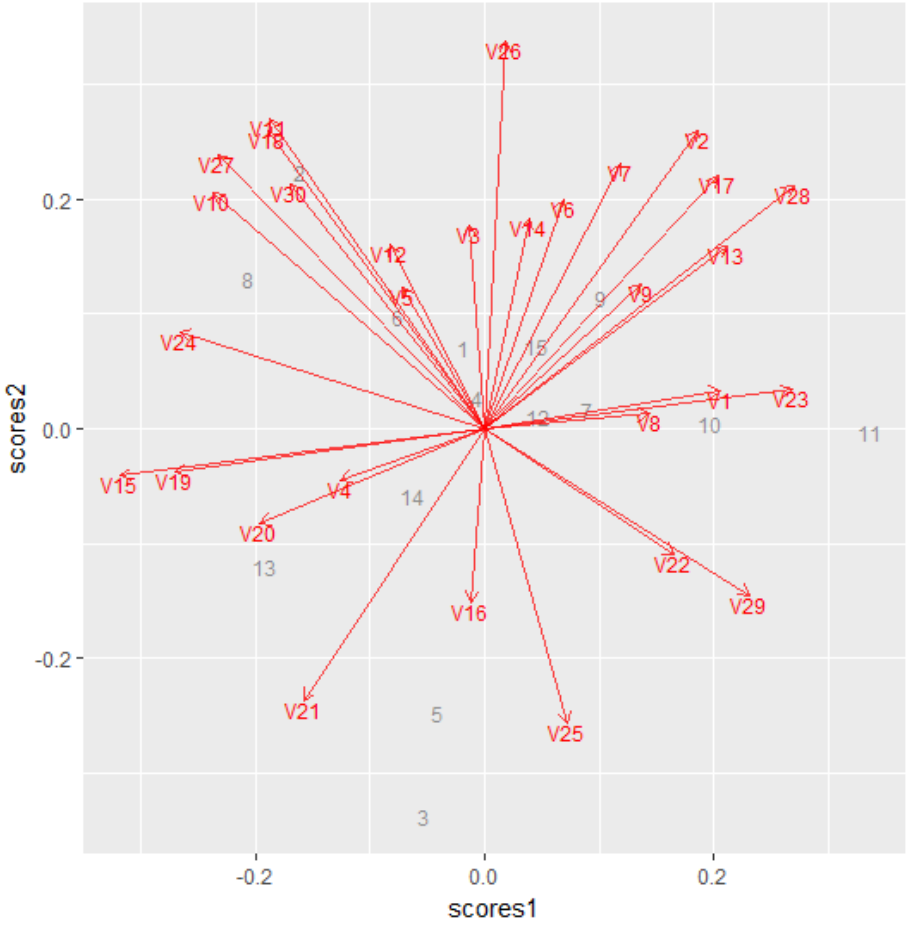
Matriz de pesos com esparsidade

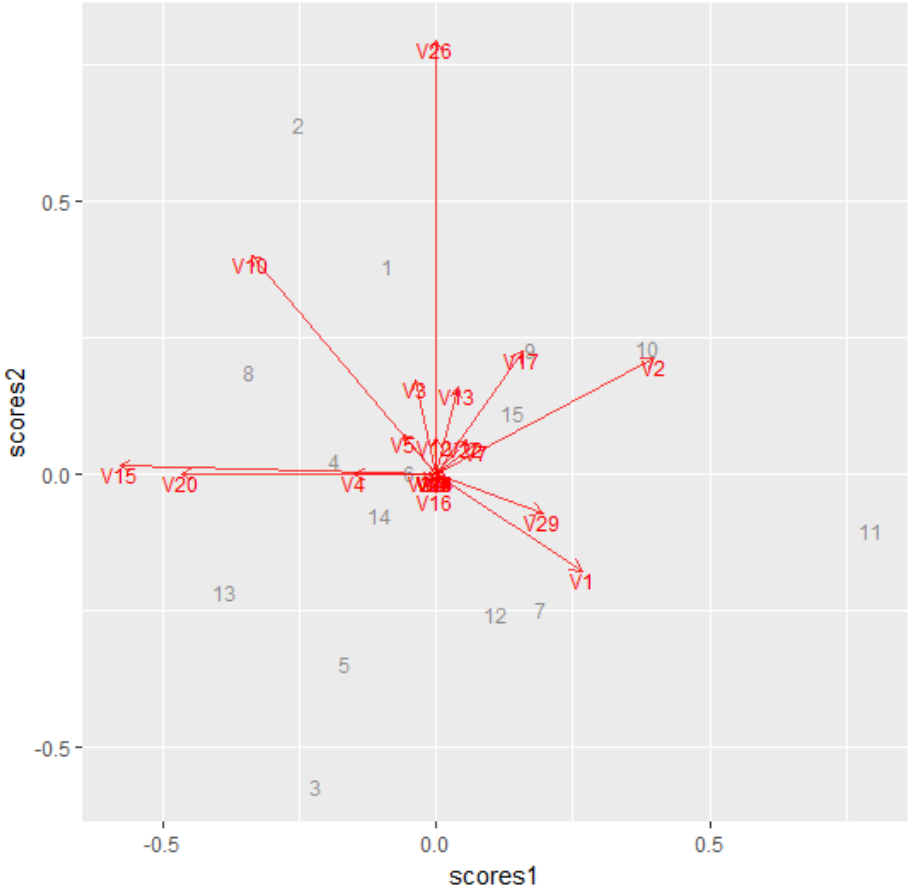
Sparse loadings

	PC1	PC2
V1	0.266	-0.177
V2	0.398	0.213
V3	-0.040	0.173
V4	-0.151	0.000
V5	-0.062	0.073
V6	0.000	0.000
V7	0.073	0.057
V8	0.000	0.010
V9	0.000	0.000
V10	-0.339	0.401
V11	0.000	0.000
V12	0.000	0.067
V13	0.040	0.160
V14	0.000	0.000
V15	-0.580	0.015
V16	0.000	-0.034
V17	0.157	0.225
V18	0.000	0.000
V19	0.000	0.000
V20	-0.467	0.000
V21	0.000	0.000
V22	0.055	0.063
V23	0.000	0.000
V24	0.000	0.000
V25	0.000	0.000
V26	0.000	0.796
V27	-0.018	0.000
V28	0.000	0.000
V29	0.195	-0.071
V30	0.000	0.000

Componentes Principais – n<<p

Biplot – (n<p): n=15 p=30
R-prcomp





Componentes Principais Esparsos – n<<p
Correspondência com Modelos de Regressão

Diferentes Soluções são propostas com o objetivo de uma aproximação de posto m (m≤min(n,p)) para $Y \in \Re^{pxn}$:

Impostas via Lagrange ou desigualdades

$$Y \approx \sum_k d_k U_k V_k'; \quad U_k \in \Re^n; V_k \in \Re^p, d_k \geq 0$$

Obter soluções sujeitas a diferentes restriçoes!

$$\sum_{k=1}^{m} d_k U_k V_k' = \arg\max \left\| Y - \hat{Y} \right\|_2^2; \quad Y = UDV$$

Mas, para n<<p:
Como obter V? Como fazer seleção de variáveis?

- Lee and Seung (2001): fatorização de Matrizes Não-Negativas

$$Y \approx \sum_k U_k V_k'; \quad U_k = \left(u_{ik} \geq 0\right) \in \Re^n; V_k = \left(v_{kj} \geq 0\right) \in \Re^p,$$

- Lazzeroni and Owen (2002): Decomposição em valores singulares

$$\min_{U_k, V_k, d_d} \sum_{i,j} \left(Y_{ij} - \sum_k u_{ik} v_{jk} d_k \right)^2; u_{ik} \in \{0,1\}, v_{jk} \in \{0,1\}, d_k \geq 0$$

Componentes Principais Esparsos – n<<p
Correspondência com Modelos de Regressão

Diferentes Soluções são propostas:

- Jolliffe et al. (2003): critério SCoTLASS (solução baseada em formas quadráticas)

$$\max_V V'\left(Y'Y\right)V; \quad \|V\|_2^2 = 1, \|V\|_1 \le c$$

- Zou et al. (2006)

$$\min_{\theta,V} \|Y - Y\,V\theta'\|_2^2 + \lambda_1 \|V\|_2^2 + \lambda_2 \|V\|_1; \quad \|\theta\|_2^2 = 1, V \in \Re^p, \theta \in \Re^p$$

- Witten et al. (2009): Decomposição em valores singulares

$$\max_{U_k,V_k} U_k'\,Y\,V_k; \quad \begin{cases} \|U_k\|_2^2 \le 1 \\ \|V_k\|_2^2 \le 1, \ \|V_k\|_1 \le c_1 \end{cases}$$

Solução equivalente ao critério SCoTLASS e de fácil de implementação!

Para múltiplos fatores (k=1,…,m)

V_k fixo, obter U_k tal que, $\max_{U_k} U_k'\,Y\,V_k; \ \boxed{\|U_k\|_2^2 \le 1, \ U_k \perp \left(U_1,\ldots U_{k-1}\right)}$

U_k fixo, obter V_k tal que, $\max_{V_k} U_k'\,Y\,V_k; \quad \|V_k\|_2^2 \le 1, \|V_k\|_1 \le c$

Análise Discriminante Esparsa – n << p

$$Y_{i\,p\times 1} \mid \tau_g \overset{iid}{\sim} \left(\mu_g ; \Sigma_g\right) \quad \Rightarrow X_i = l' Y_i \overset{iid}{\sim} \left(l'\mu_g ; l'\Sigma l\right)$$

Solução (linear) de Fisher: Suposição $\Rightarrow \Sigma_g = \Sigma$

Para n>p: S_w^{-1}?

$$\max_l \frac{l' \sum_{g=1}^{G} \left(\bar{Y}_g - \bar{Y}_.\right)\left(\bar{Y}_g - \bar{Y}_.\right)' l}{l' S_w l} = \max_l \frac{l' S_b l}{l' S_w l} \quad \left\{ \begin{array}{l} \max_l l' S_b l \; ; \quad \begin{array}{l} l_j' S_w l_j = 1, \\ l_j' S_w l_k = 0, \; j \neq k \end{array} \\ j = 1, \ldots, m \leq \min\left(n, p, G-1\right) \end{array} \right.$$

Tabela de MANOVA

$G_{n\times G}$: Matriz de incidência de grupos; $P_G = G\left(G'G\right)^{-1} G'$

F.V.	g.l.	Matriz de SQPC
Trat	**G-1**	$H_{p\times p} = \sum_{g=1}^{G} n_g \left(\bar{\mathbf{y}}_g - \bar{\mathbf{y}}\right)\left(\bar{\mathbf{y}}_g - \bar{\mathbf{y}}\right)' = S_b = Y' P_G Y$
Resíduo	**n-G**	$E_{p\times p} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left(\mathbf{y}_{gi} - \bar{\mathbf{y}}_g\right)\left(\mathbf{y}_{gi} - \bar{\mathbf{y}}_g\right)' = S_w = Y'\left(I - P_G\right) Y$
TOTAL	**n-1**	$H + E = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left(\mathbf{y}_{gi} - \bar{\mathbf{y}}\right)\left(\mathbf{y}_{gi} - \bar{\mathbf{y}}\right)' = Y'Y$

Análise Discriminante Esparsa – n << p

- Big-Data: n << p $\max_{\beta_k} \dfrac{\beta_k' S_b \beta_k}{\beta_k' S_w \beta_k}; \quad S_w$ não é p.d

Solução regularizada para corrigir o posto de S_W

Tornar autovalores positivos

$$\max_{\beta_k} \frac{\beta_k' S_b \beta_k}{\beta_k' \left(S_w + \Omega\right) \beta_k}; \quad \left|\left(S_w + \Omega\right) - \lambda I_p\right| = 0, \ \lambda > 0$$

$$\Rightarrow \max_{\tilde{\beta}_k} \tilde{\beta}_k' \tilde{S}_b \tilde{\beta}_k \begin{cases} \tilde{S}_b = \left(S_w + \Omega\right)^{-1/2} S_b \left(S_w + \Omega\right)^{-1/2} \\ \beta_k' \left(S_w + \Omega\right) \beta_k' = 1; \ \beta_k' \left(S_w + \Omega\right) \beta_j' = 0; \end{cases}$$

Mas, como obter Ω? E os β's, são esparsos?

Hastie et al., (1995); Clemmensen et al., (2011, 2016)

Solução Regularizada e Esparsa: implementada na função *sda* do pacote sparseLDA

$$\hat{\beta}_{k\ p\times 1}; \ \min_{\beta_k, \theta_k} \left\{ \left\| G_{n\times G} \theta_k - Y_{n\times p}' \beta_k \right\|_2^2 + \lambda_1 \beta_k' \Omega \beta_k + \lambda_2 \left\| \beta_k \right\|_1 \right\}; \ \theta_k' G' G \, \theta_k = 1, \ \theta_k' G' G \, \theta_j = 0,$$

Vetor de pesos dos grupos

Para $\Omega = I_m$: usar algoritmo do *elastic net*

Análise Discriminante Esparsa – n << p

- Big-Data: n << p

sparseLDA: Dados “penicilliumYES”, espécies de fungos (indistinguíveis) avaliadas em variáveis de imagem
n=36 p=3.754
G=3: "*P. Melanoconidium*", "*P. Polonicum*", "*P. Venetum*"

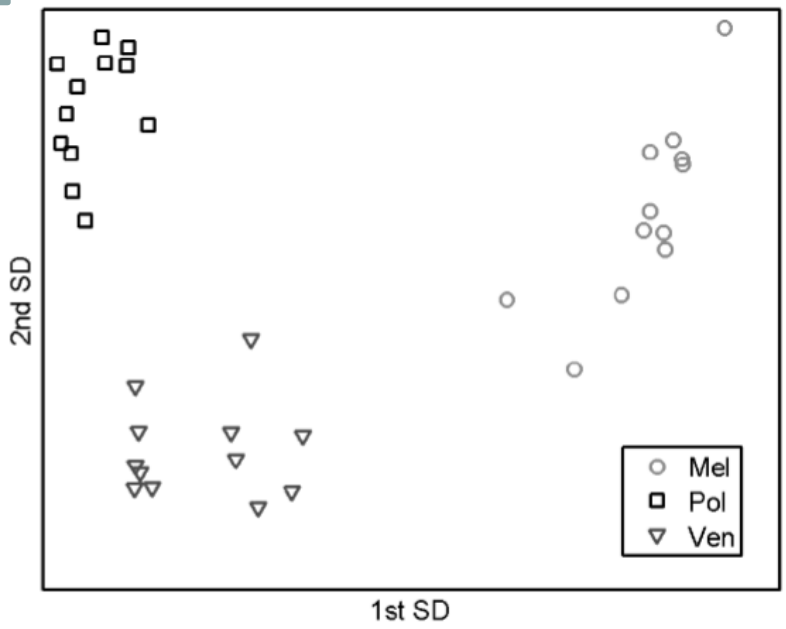
	$Y\hat{\beta}_1$	$Y\hat{\beta}_2$
[1,]	-3.113028	-3.4122173
[2,]	-3.142295	-3.8733571
[3,]	-2.988152	-1.4446112
[4,]	-2.995266	-1.5567002
[5,]	-2.995715	-1.5637771
[6,]	-3.063970	-2.6392373
[7,]	-2.996005	-1.5683428
[8,]	-2.998670	-1.6103476
[9,]	5.059007	0.9703178
[10,]	5.803007	0.7348022
[11,]	2.720020	0.5259276
[12,]	4.546290	-0.6951151
[13,]	6.415934	-0.6321676
[14,]	5.898611	0.1021021
[15,]	5.472054	-1.9047044
[16,]	8.543129	-0.2298407
[17,]	-2.687269	3.2962259
[18,]	-2.666652	3.6210784
[19,]	-2.840601	0.8802641
[20,]	-2.848656	0.7533365
[21,]	-2.733009	2.5755237
[22,]	-2.262692	2.7361895
[23,]	-2.779546	1.8422657
[24,]	-1.346527	3.0923848
...		

Escores da funções discriminantes

	$\hat{\theta}_1$	$\hat{\theta}_2$
[1,]	0.7357055	1.20778203
[2,]	-1.4138227	0.03324864
[3,]	0.6781172	-1.24103066

Matriz de pesos dos grupos

Representação das espécies nas variáveis discriminantes



Correlação Canônica Esparsa – n << p

$$Y_{i\,(p+q)\times 1} = \begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \overset{iid}{\sim} \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right) \Rightarrow \begin{cases} a_k{}'Y_1 \\ b_k{}'Y_2 \end{cases} \quad \max_{a_k, b_k} \rho\left(a_k{}'Y_1, b_k{}'Y_2 \right)$$

$$\Rightarrow \frac{a' \boldsymbol{\Sigma}_{12} \Sigma_{22}^{-1} \Sigma_{21} a}{a' \Sigma_{11} a} \qquad \Rightarrow \frac{b' \boldsymbol{\Sigma}_{21} \Sigma_{11}^{-1} \Sigma_{12} b}{b' \Sigma_{22} b}$$

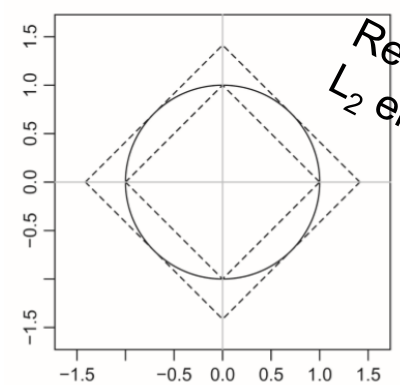
Problemas na otimização quando n<<p n<<q

Inicialmente, considere o problema de obter uma solução penalizada para a decomposição em valores singulares de matrizes:

PMD: Penalized Matrix Decomposition (Witten, Tibshirani, Hastie, 2009, 2015)

$$Y_{n\times p} = UDV', \quad U'U = I_n, V'V = I_p$$

$$\min_{U_k, V_k, d_k} \left\| Y - U_k V_k{}' d_k \right\|_2^2; \quad \begin{cases} \|U_k\|_2^2 \le 1, \|U_k\|_1 \le c_1 \\ \|V_k\|_2^2 \le 1, \|V_k\|_1 \le c_2 \end{cases}$$



Restrições L_1 e L_2 em $V_k \in \Re^2$

Decomposição de Matrizes - Penalização

PMD: Penalized Matrix Decomposition

$$\min_{U_k, V_k, d_k} \left\| Y - U_k V_k' d_k \right\|_2^2 ; \quad \begin{cases} \|U_k\|_2^2 \le 1, \|U_k\|_1 \le c_1 \\ \|V_k\|_2^2 \le 1, \|V_k\|_1 \le c_2 \end{cases}$$

$$\frac{1}{2}\left\| Y - UDV' \right\|_2^2 = \frac{1}{2}\|Y\|_2^2 - \sum_{k=1}^{m} U_k' Y V_k d_k + \frac{1}{2}\sum_{k=1}^{m} d_k^2$$

$$\max_{U_k, V_k} U_k' Y V_k \ ; \quad \begin{cases} \|U_k\|_2^2 \le 1, \|U_k\|_1 \le c_1 \\ \|V_k\|_2^2 \le 1, \|V_k\|_1 \le c_2 \end{cases}$$

PMD(L_1,L_1)

Diferentes algoritmos são propostos para solução deste problema de maximização (Witten et al., 2009, 20015)

Bilinear em U e V $\Rightarrow$ U fixo e obter V, V fixo e obter U

$$V_k \text{ fixo}: \max_{U_k} U_k' Y V_k \ ; \quad \|U_k\|_2^2 \le 1, \|U_k\|_1 \le c_1, 1 \le c_1 \le \sqrt{n}$$

$$U_k \text{ fixo}: \max_{V_k} U_k' Y V_k \ ; \quad \|V_k\|_2^2 \le 1, \|V_k\|_1 \le c_2, 1 \le c_2 \le \sqrt{p}$$

Decomposição de Matrizes – Penalização Aplicação

- CCA – Esparso: PMD(L_1,L_1)

$$\max_{a_k, b_k} a_k' Y_1' Y_2 b_k \; ; \quad \begin{cases} a_k' Y_1' Y_1 a_k \le 1, \; \|a_k\|_1 \le c_1 \\ b_k' Y_2' Y_2 b_k \le 1, \; \|b_k\|_1 \le c_2 \end{cases}$$

Algoritmo proposto: **CCA-P Diagonal**. Assume que para dados em alta dimensão a matriz de covariância diagonal pode ser adotada (Dudoit et al. 2001; Tibshirani et al., 2003)

$$a_k' Y_1' Y_1 a_k \overset{Y_1' Y_1 = I_p}{=} a_k' a_k \le 1, \quad b_k' Y_2' Y_2 b_k \overset{Y_2' Y_2 = I_q}{=} b_k' b_k \le 1$$

- PCA – Esparso: PMD(·,L_1) (Witten et al., 2009, 2015)

$$\max_{V_k} V_k' Y' Y V_k ; \quad \|V_k\|_2^2 \le 1, \; \|V_k\|_1 \le c_2$$

Correlação Canônica Esparsa

- Pacote PMA-R, função CCA:

Dados gerados: n=100 p=500 q=1000

$$U_{n\times 1} = Y_1 a \quad \Rightarrow \quad n\left(a_j \neq 0\right) = 126;$$

$$V_{n\times 1} = Y_2 b \quad \Rightarrow \quad n\left(b_j \neq 0\right) = 311; \; c_1 = c_2 = 0,367$$

$$\rho_c = 0,9612$$

Componentes Principais via Modelos de Regressão

Revisando!

Como obter as variáveis originais a partir das componentes principais?

$$\mathrm{Y} \in \Re^{\mathrm{p}} \quad \Rightarrow \quad \mathrm{Z} = P'\,\mathrm{Y} \in \Re^{p}; \qquad \boxed{Z_j = a_{1j}Y_1 + a_{2j}Y_2 + \ldots + a_{pj}Y_p} \qquad P = (a_{ij}); \quad P' = (a_{ji})$$

$$\mathrm{Z} = P'\,\mathrm{Y} \quad \Leftrightarrow \quad \mathrm{P\,Z} = \mathrm{P\,P'\,Y} = \mathrm{Y} \quad \Rightarrow \quad \mathrm{Y} = \mathrm{P\,Z}$$

$$\boxed{Y_j = a_{j1}Z_1 + a_{j2}Z_2 + \ldots + a_{jp}Z_p}$$

$$Var(Y_j) = \sum_{k=1}^{p} a_{jk}^2\, Var(Z_k) = \sum_{k=1}^{p} a_{jk}^2\, \lambda_k \quad \Rightarrow \quad R_{jk}^2 = \frac{a_{jk}^2\, \lambda_k}{\sigma_{jj}}$$

é o quadrado da correlação entre Y_j e Z_k

Proporção da Var(Y_j) explicada por Z_k . É o coeficiente de determinação da regressão de Y_j em Z_k .

Aplicação de CP em Modelos de Regressão

Considere a seguinte situação da Regressão de Y em p Variáveis Preditoras:

Matriz de dados: $\left(\boxed{Y_{n\times 1}}, \boxed{X_{1\,n\times 1}, \ldots, X_{p\,n\times 1}}\right)_{n\times(p+1)}$

$Y_i \in \Re$

Objetivo: Modelo de Regressão de Y nos Componentes Principais das p Variáveis Preditoras (problemas de Multicolinearidade)

Componentes Principais e Modelos de Regressão
Seleção de Variáveis Preditoras

Redução de Dimensionalidade em Modelos de Regressão

$$\boxed{Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + e_i;} \quad Y_{n\times 1} = 1_n \beta_0 + X_{n\times p}\beta_{p\times 1} + e_{n\times 1}$$

Seleção de variáveis preditoras por PCR (Componentes Principais de Var. Preditoras)

$$X = U\Lambda V'; \quad Z = XV, \quad X = ZV'$$

$$\Rightarrow \quad Y_{n\times 1} = 1_n \beta_0 + Z_{n\times p} V'_{p\times p}\, \beta_{p\times 1} + e_{n\times 1} = \alpha_0 + Z_{n\times p}\alpha_{p\times 1} + e_{n\times 1}; \quad \alpha = V'\beta$$

$$\underset{\text{Estimação em } \Re^{(p+1)}}{X_1,\ldots,X_p} \quad \rightarrow \quad \underset{\text{Estimação em } \Re^{(m+1)}}{Z_1,\ldots,Z_m} \quad ; \qquad Z_{ki} = \sum_{j=1}^{p} a_{jk} X_{ij}, \quad k = 1,\ldots,m; \quad m < p$$

$$\boxed{Y_i = \alpha_0 + \sum_{k=1}^{m} \alpha_k Z_{ki} + e_i}$$

Variáveis Z são não correlacionadas $\Rightarrow$ Útil na seleção de variáveis preditoras.

Restrições impostas pela redução de dimensionalidade

$$\sum_{k=1}^{m} \alpha_k Z_{ki} = \sum_{k=1}^{m} \alpha_k \sum_{j=1}^{p} a_{jk} X_{ij} = \sum_{j=1}^{p} \boxed{\sum_{k=1}^{m} \alpha_k a_{jk}} X_{ij} = \sum_{j=1}^{p} \beta_j X_{ij}$$

Exemplo: Gerar n=50 observações da N_1 com p=45 variáveis preditoras.

Componentes Principais e Modelos de Regressão PLS

PLS: Mínimos Quadrados Parciais

(Tibshirani et al., 2015; Abdi, 2010)

$Y_i \in \Re$

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + e_i; \qquad X_1, \ldots, X_p \rightarrow Z_1, \ldots, Z_m$$

Obter direções (variáveis Z) em X que melhor explicam Y.

Estimação em $\Re^{(p+1)}$

PCR: variáveis Z só explicavam X

$$Y_i = \varphi_{1j} X_{ij} + e_i, \quad j = 1, 2, \ldots, p \xrightarrow{\hat{\varphi}_{1j} \text{ : Solução de MQ}} Z_{1i} = \sum_{j=1}^{p} \hat{\varphi}_{1j} X_{ij}$$

$$X_{ij} = a_{0j} + b_j Z_{1i} + e_{(X_j|Z_1)i} \quad j = 1, .., p \quad \Rightarrow \quad \hat{e}_{(X_j|Z_1)i}$$

$$Y_i = \varphi_{2j} \hat{e}_{(X_j|Z_1)i} + e_i, \quad j = 1, 2, \ldots, p \longrightarrow Z_{2i} = \sum_{j=1}^{p} \hat{\varphi}_{2j} X_{ij}$$

CP de X que melhor predizem Y

$$\Rightarrow \quad Z_1, \ldots, Z_m; \quad Y_i = \alpha_0 + \sum_{k=1}^{m} \alpha_k Z_{ki} + e_i$$

Estimação em $\Re^{(m+1)}$

Regressão PLS: Formulação Geral

(Tibshirani et al., 2015; Abdi, 2010)

$Y_i \in \Re^p$

$$\left(Y_{n\times p} \quad X_{n\times q}\right)_{n\times(p+q)}$$

Obter "Componentes Principais" de X que melhor predizem Y

$$X_{n\times q} = \underbrace{F_{X\,n\times m}}_{\text{Matriz de Escores (Coord. Principais)}} \underbrace{W'_{X\,m\times q}}_{\text{Matriz de Cargas}}$$

$$Y_{n\times p} = \underbrace{F_{Y\,n\times m}}_{\text{Matriz de Escores (Coord. Principais)}} W'_{Y\,m\times p} = F_X \; \underbrace{B\,W_Y{}'}_{\text{Matriz de Cargas}}; \quad B = diag(d_j)$$

Obter $F_{Xj}, F_{Yj};\ \max_{F_{Xj},F_{Yj}} F_{Xj}{}' F_{Yj} = d_j$

Maximizar o produto interno dos Escores

Considerando X e Y normalizadas, Passo1: obter as primeiras soluções

$$\begin{cases} \left(X'Y\right)_{q\times p} = W_x B W_Y{}' \\ \left(X(X'Y)Y'\right)_{n\times n} = F_X F_Y{}' \\ \left(Y(Y'X)X'\right)_{n\times n} = F_Y F_X{}' \end{cases}$$

Algoritmo: obtenção dos escores

Atualizar X e Y, e usar o resíduo do Passo1, até obter *m* componentes.

$$\begin{cases} X^{(2)} = X - X^{(1)} = X - F_{X1} W_{X1}{}' \\ Y^{(2)} = Y - Y^{(2)} = Y - F_{Y1} W_{Y1}{}' d_1 \end{cases}$$

Essas direções não são iguais às da Correlação Canônica (que é uma solução simétrica, diferentemente de PLS que é assimétrica/direcionada)

Regressão PLS: Formulação Geral

Obter "Componentes Principais" de X que melhor predizem Y

$$\left(Y_{n\times p} \quad X_{n\times q}\right)_{n\times(p+q)}$$

$$X_{n\times q} = F_{X\,n\times m} \; W'_{X\,m\times q} \qquad Y_{n\times p} = \boxed{F_{Y\,n\times m}} \; W'_{Y\,m\times p} = \boxed{F_X \; B} W_Y{}'; \quad B = diag$$

$$\left(X'Y\right)_{q\times p} = W_x B W_Y{}'$$

$$\left(X(X'Y)Y'\right)_{n\times n} = F_X F_Y{}'$$

$$\left(Y(Y'X)X'\right)_{n\times n} = F_Y F_X{}'$$

$$\longrightarrow$$

$$X^{(2)} = X - F_{X1} W_{X1}{}'$$

$$Y^{(2)} = Y - F_{Y1} W_{Y1}{}' d_1$$

$$\longrightarrow \left(F_{Xj}, F_{Yj}\right), \; j = 1, \ldots, m$$

$$\left(F_X, F_Y\right)_{n\times m}$$

$$\left(F_X, F_Y\right)_{n\times m} \begin{cases} F_X = XW_X \quad F_Y = YW_Y \\ \hat{Y} = F_X B W_Y{}' = X\boxed{W_X B W_Y{}'} = X\boxed{B_{PLS}}; \quad B_{PLS} = W_X{}' B W_Y \end{cases}$$

Regressão PLS e Correlação Canônica
Intergração de Bancos de Dados

Solução PLS (Partial Least Square):

$$a,b;\quad \max_{a\in\mathbb{R}^p,a\in\mathbb{R}^q} = \frac{\boxed{\left[Cov\left(a'X;b'Y\right)\right]^2}}{\left(a'a\right)\left(b'b\right)} \Rightarrow \quad \Sigma_{XY}\Sigma_{YX}a = \lambda a, \quad b = \Sigma_{YX}a$$

Solução CCA (Canonial Correlation Analysis):

$$a,b;\quad \max_{a\in\mathbb{R}^p,a\in\mathbb{R}^q} \boxed{\left[Corr\left(a'X;b'Y\right)\right]^2}$$

$$\boxed{\left[Cov\left(a'X;b'Y\right)\right]^2 = Var\left(a'X\right)\left[Corr\left(a'X;b'Y\right)\right]^2 Var\left(b'Y\right)}$$

PLS é a CCA com regularizações definidas pelos PC de X e PC de Y

Regressão PLS , CCA e DA
Integração de Bancos de Dados

PLS é a CCA com penalizações definidas pelos PC de X e PC de Y

$$\left[Cov\left(a'X;b'Y\right)\right]^2 = Var\left(a'X\right)\left[Corr\left(a'X;b'Y\right)\right]^2 Var\left(b'Y\right)$$

Bartlet (1938): estabelece conexões entre CCA e e Análise Discriminante (DA) $\Rightarrow$ CCA pode ser realizada com um conjunto X de treinamento e uma matriz Y de variáveis indicadoras da Doença (Matriz G de var. dummy)

Barker and Raynes (2003) propõem PLS-DA

Assim, um procedimento formulado para problemas de regressão com muitas variáveis preditoras, passa a ser usado para predizer estrutura de grupos.

Lê Cao implementa PLS-DA para Integração Supervisionada de BD no mixOmics-R

Soluções mais gerais:

$$\max_{a_j,\, j=1,..,J} \sum_{j\neq k} c_{jk}\, g\left(Cov\left(X_j a_j; X_k a_k\right)\right); \quad \left\|a_j\right\|_2 = 1, \quad |a_j| \leq \lambda_j, \quad j = 1,\ldots,J$$

$$g(x) = x, \quad g(x) = |x|, \quad g(x) = x^2$$

J bancos de dados

$C = \left(c_{jk}\right)$: Pesos às conexões entre BD