MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Sem/2020 - 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 ($m \leq min(n,p)$)
- ✓ Coordenadas Principais – Escalonamento Multidimensional ($m \leq posto(D_{nxn})$)
- ✓ Análise de Correspondência
- ✓ Análise Fatorial Exploratória (via CP: $m \leq min(n,p)$)
- ✓ Análise Discriminante linear ($m \leq min(n,p, G-1)$) $\leftrightarrow$ MANOVA
- ✓ Análise de Agrupamento

$n > p$ Observações Independentes

Chamo a atenção para:

$\Rightarrow$ Soluções baseadas em quais Critérios?
$\Rightarrow$ Problema Dual de redução de dimensionalidade ($\Re^{nxp}$, $\Re^{pxp}$, $\Re^{nxn}$)
$\Rightarrow$ Possíveis Partições dos dados ($\Re^{pxp}$, $\Re^{nxp}$)
$\Rightarrow$ Representações Bi-Plot

Revisando ☺

Componentes Principais (Pearson, 1901)

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

Suposições: AASn de uma única população com matriz de cov Σ, obs independentes, n>p

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:

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

$$Y_{n\times p} \rightarrow Z_{n\times m};\ Z_{ki} = V_k{}'Y_i$$

escore

carga

$$V_k\ ;\quad \arg\max_{\|a_k\|=1} \frac{a_k{}'\Sigma a_k}{a_k{}'a_k}$$

tr(Σ)

$$= \arg\max_{\|a_k\|=1} \sum_{k=1}^{m} Var\left(Z_{ki}\right)$$

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

$$\Sigma = V\Lambda V';\quad \Lambda = D_{\lambda_j};\ VV' = V'V = I_p;\quad \left|\Sigma - \lambda I_p\right| = 0;\quad \Sigma V_k = \lambda_k V_k$$

$$\Sigma_Y = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22}^2 & \ldots & \sigma_{2p} \\ & & \ldots & \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp}^2 \end{pmatrix} = \lambda_1 V_1 V_1' + \ldots + \lambda_m V_m V_m{}' + \ldots + \lambda_p V_p V_p{}';\quad \lambda_1 \geq \ldots \geq \lambda_m \geq \ldots \geq \lambda_p$$

Análise Fatorial Exploratória (Spearman, 1904)

Redução de dimensionalidade $\Rightarrow$ Modelagem estrutural da 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)

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

Suposição:

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

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

$$\Leftrightarrow \Sigma_{p\times p} \cong \Phi_{p\times m}\Phi'_{m\times p} + \Psi_{p\times p}$$

Comunalidade Especificidade

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

Ψ : Matriz (diagonal) de fatores específicos

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

Solução via Componentes Principais

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

$$\Rightarrow Z_{ip\times 1} = V'\,Y_i$$

$$\Rightarrow Y_i = V\,Z_i = \left(V\Lambda^{1/2}\right)\left(\Lambda^{-1/2}Z_i\right)$$

Cargas fatoriais

Fatores comuns (CP padronizados)

$\rightarrow$ Solução via MVS

Análise Discriminante Linear (Fisher, 1938)

Análise supervisionada

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

Obter combinações lineares para a máxima separação dos grupos:

	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}

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

p variáveis mais grupo

... 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 $\Rightarrow \Sigma_g = \Sigma$

Solução (linear) de Fisher:

$$\frac{l'\sum_{g=1}^{G}\left(\bar{Y}_g - \bar{Y}_.\right)\left(\bar{Y}_g - \bar{Y}_.\right)' l}{l' S_W l} = \frac{l'\, SS_B\, l}{l' S_W l}$$

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

$$SS_T = SS_B + SS_W$$

← MANOVA

MANOVA: DCA com Um Fator em G níveis

Aplicar Técnicas Multivariadas nos componentes da Matriz de "Covariância"

Tabela de MANOVA

Decomposição em $\Re^{p \times p}$

F.V.	g.l.	Matriz de SQPC (SS)	Matriz de QMPC (S)
Fator	**G-1**	$H_{p\times p} = \sum_{g=1}^{G} n_g \left(\bar{Y}_g - \bar{Y}\right)\left(\bar{Y}_g - \bar{Y}\right)'$	$S_B = H/(G-1)$
Resíduo	**G(K-1)=n-G**	$E_{p\times p} = \sum_{g=1}^{G}\sum_{i=1}^{n_g} \left(Y_{ig} - \bar{Y}_g\right)\left(Y_{ig} - \bar{Y}_g\right)'$	$S_W = E/(n-G)$ $=S_c$
TOTAL	**n-1**	$T = H + E = \sum_{g=1}^{G}\sum_{k=1}^{b}\left(Y_{ig} - \bar{Y}\right)\left(Y_{ig} - \bar{Y}\right)'$	$S_T = T/(n-1)$

n_g=K n=GK

$H_0 : \boldsymbol{\mu}_g = \boldsymbol{\mu}$

$$\Lambda^* = \frac{|E|}{|H+E|}$$ Lambda de Wilks

$$Y_{ig} \overset{iid}{\sim} N_p(\mu;\Sigma) \Rightarrow \left(\frac{n-p-2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p,2(n-p-2)}$$

MANOVA: Partição dos Dados

Modelo ANOVA: Delineamento Completamente Aleatorizado com Um fator em G níveis

$$y_{ig\ 1\times1} = \bar{y} + (\bar{y}_g - \bar{y}) + (y_{ig} - \bar{y}_g)$$

Identidade útil para formulação do esqueleto da ANOVA

Resposta

efeito de Tratamento "ENTRE"

Resíduo "DENTRO"

Modelo MANOVA: $Y_{ig} \in \Re^p$

$$Y_{ig\ p\times1} = \bar{Y} + (\bar{Y}_g - \bar{Y}) + (Y_{ig} - \bar{Y}_g)$$

$$Y_{n\times p} = \bar{Y} + (\bar{Y}_g - \bar{Y}) + (Y - \bar{Y}_g)$$

Decomposição da Matrix de Resposta

Decomposição em $\Re^{nxp}$

Aplicar Técnicas Multivariadas nos componentes da Matriz de Dados

Exemplo

Duas variáveis avaliadas em unidades amostrais submetidas a 3 tratamentos

T1		T2		T3	
Y11	**Y12**	**Y21**	**Y22**	**Y31**	**Y32**
9	3	0	4	3	8
6	2	2	0	1	9
9	7			2	7
8	4	1	2	2	8

Média geral = (4 , 5)

$$Y_{8\times2} = \bar{Y}_{8\times2} + (\bar{Y}_g - \bar{Y})_{8\times2} + (Y - \bar{Y}_g)_{8\times2}$$

$$\begin{pmatrix} 9 & 3 \\ 6 & 2 \\ 9 & 7 \\ 0 & 4 \\ 2 & 0 \\ 3 & 8 \\ 1 & 9 \\ 2 & 7 \end{pmatrix} = \begin{pmatrix} 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \end{pmatrix} + \begin{pmatrix} 4 & 3 \\ 4 & 2 \\ 4 & 7 \\ -3 & 4 \\ -3 & 0 \\ -2 & 3 \\ -2 & 3 \\ -2 & 3 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ -2 & -2 \\ 1 & 3 \\ -1 & 2 \\ 1 & -2 \\ 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix}$$

Dependendo do problema, pode haver interesse no componente do Efeito de Tratamento ou no componente Residual (resposta normalizada)

Redução de Dimensionalidade
Obtenção de Vetores Reducionistas

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

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

- Análise Fatorial Exploratória (via CP): $\Rightarrow F_{ki} = \lambda^{-1/2} Z_{ki} \quad \Sigma_{p\times p} = \Phi_{p\times m\times p}\Phi' + \Psi$

Diag

- Análise Discriminante (Linear de Fisher): $f\left(\Sigma_W^{-1}\Sigma_B; a\right) = \frac{a'\Sigma_B a}{a'\Sigma_W a}, \quad a'\Sigma_W a = 1 \quad \Rightarrow X_{ki} = a_k{}'Y_i$

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

$$|\Sigma_W^{-1}\Sigma_B - I_p| = 0$$

$$\Sigma_T = \Sigma_B + \Sigma_W$$

Componentes Principais – Coordenadas Principais
Solução via Espaços Duais

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

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

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

m Componentes Principais (CP)
$m \le r$

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

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

$=Y$

Escalonamento Multidimensional:
m Coordenadas Principais (CoP)
obtidas da Matriz de Distâncias

Análise no espaço $\Re^{nxp}$

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

Equivalência entre os Componentes Principais e as Cordenadas Principais

Flexibilidade de obtenção dos escores nos casos em que n>p (obter CP) ou n<<p (obter CoP)

Componentes Principais – Coordenadas Principais

Equivalência das Soluções em Espaços Duais

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

$$H = I_n - n^{-1}1_n1_n' \qquad (HY)_{n\times p} = U_{n\times n}\begin{pmatrix}\Lambda_r^{1/2} & 0\\ 0 & 0\end{pmatrix}V'_{p\times p}$$

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

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

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

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

Coordenadas Principais

Componentes Principais

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)
  → 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: $\lambda_{eigen} = \left(\lambda_{svd} / \sqrt{n-1} \right)^2$
- 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)

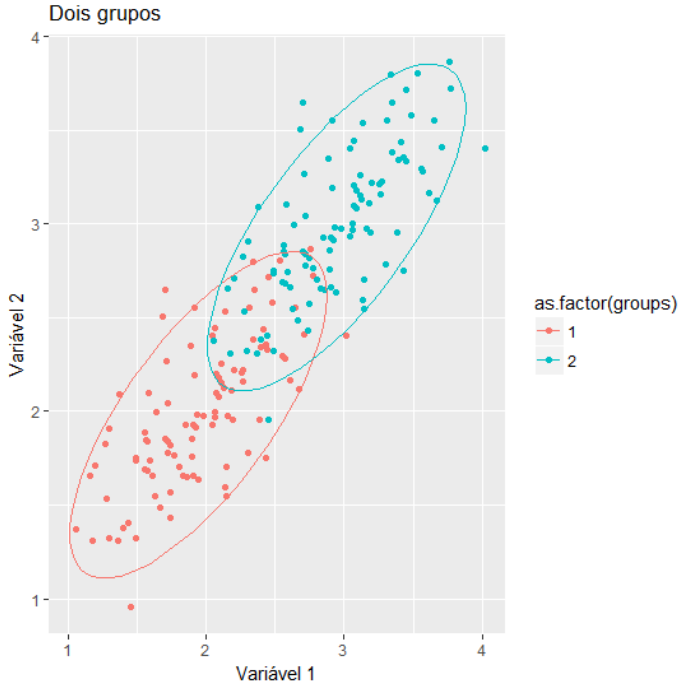
Onde estão os Vetores Reducionistas?

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%).

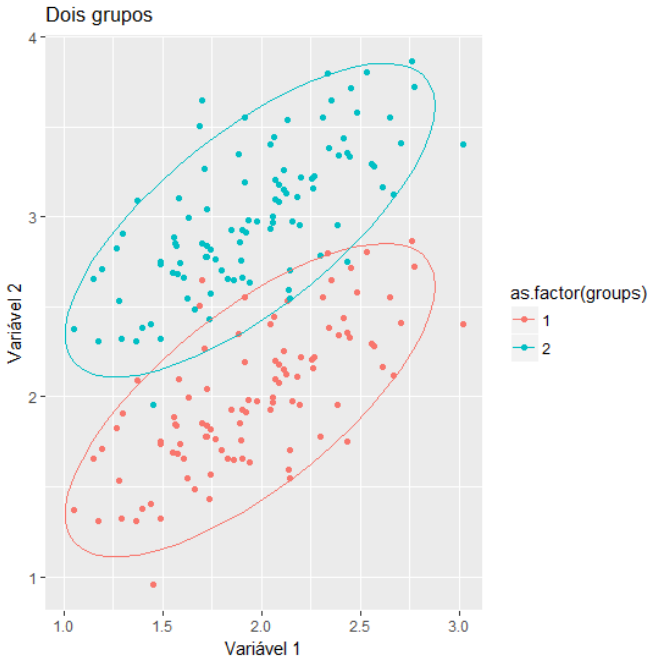
Exemplo 1



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

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

Exemplo 2



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

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

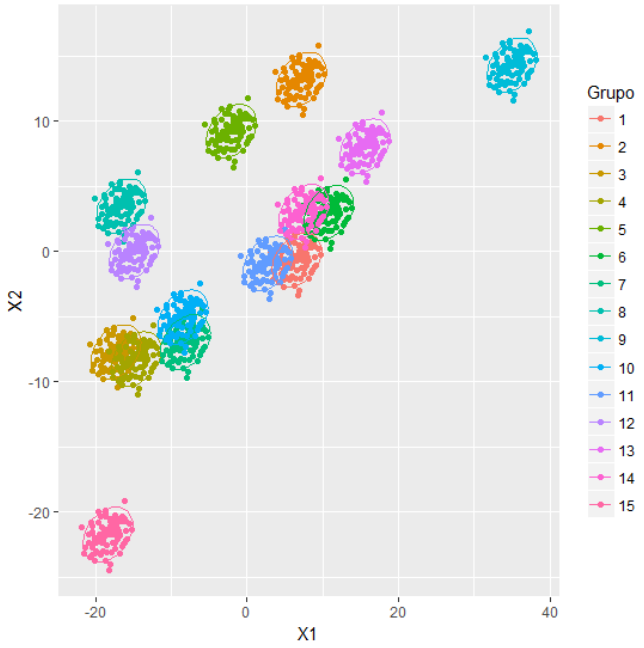
Onde estão os Vetores Reducionistas?

Obter a direção do Eixo Discriminante.

Observações independentes ENTRE e DENTRO de grupos.

Exemplo 3: “Sinais Iguais”

$$T = B + W$$

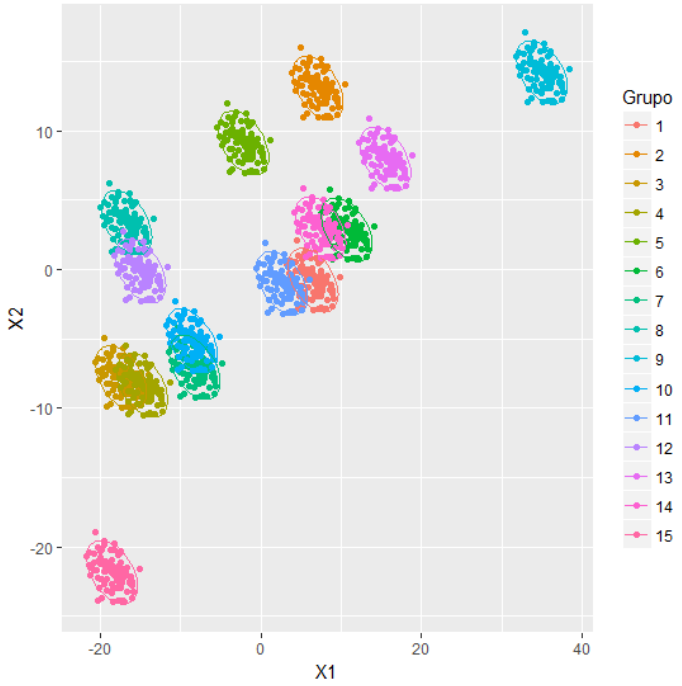


$$G = 15,\ n_g = 100,\ \mu_{\cdot} = (0,0)$$

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

Exemplo 4: “Sinais Opostos”

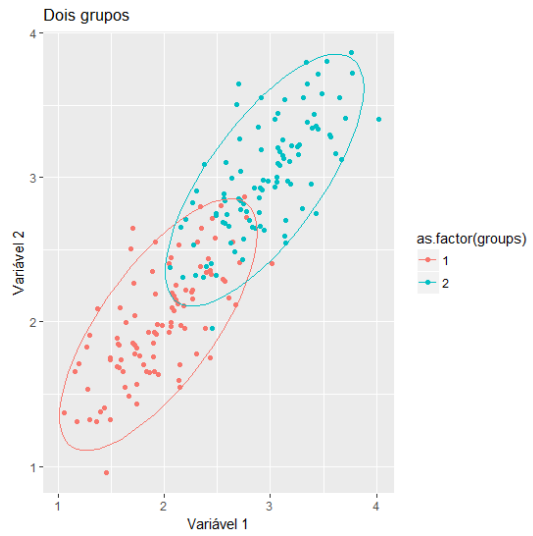
$$T = B + W$$



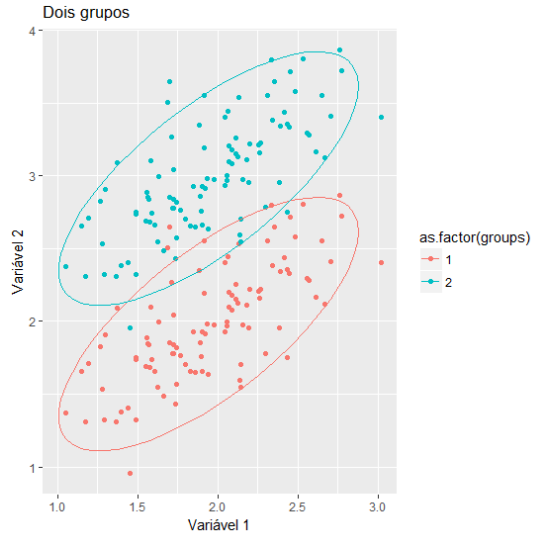
$$G = 15,\ n_g = 100,\ \mu_{\cdot} = (0,0)$$

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

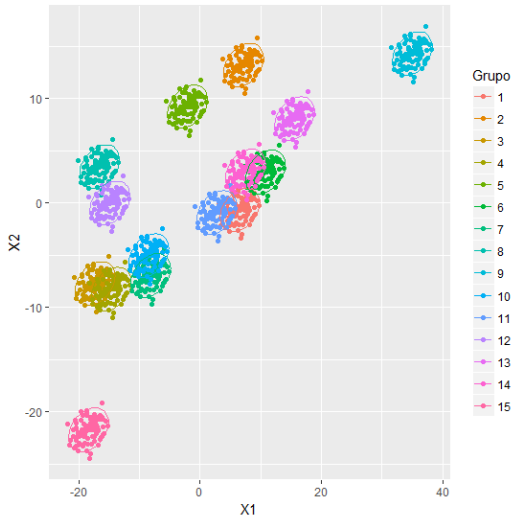
Exemplo 1



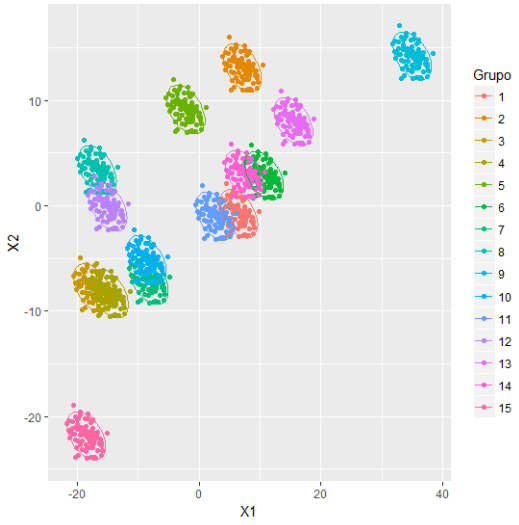
Exemplo 2



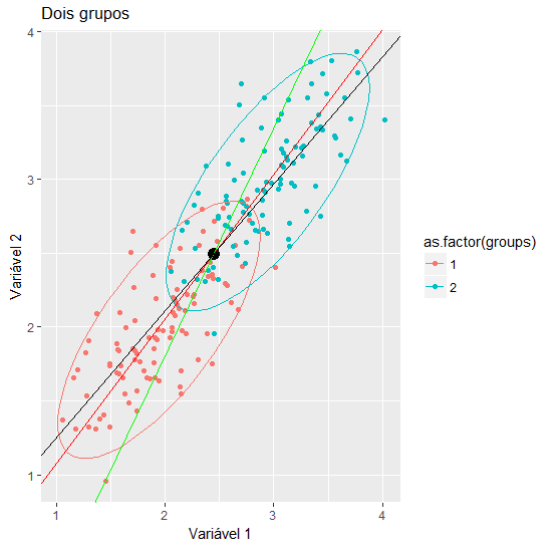
Exemplo 3



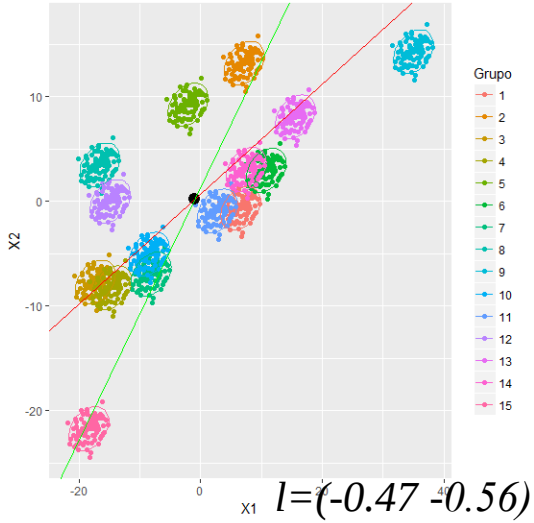
Exemplo 4



Exemplo 1

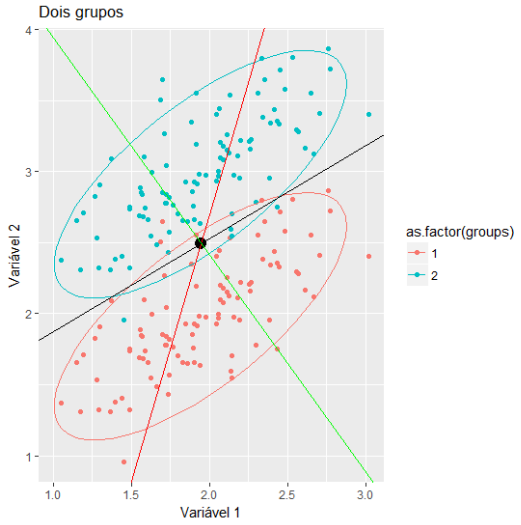


Exemplo 3

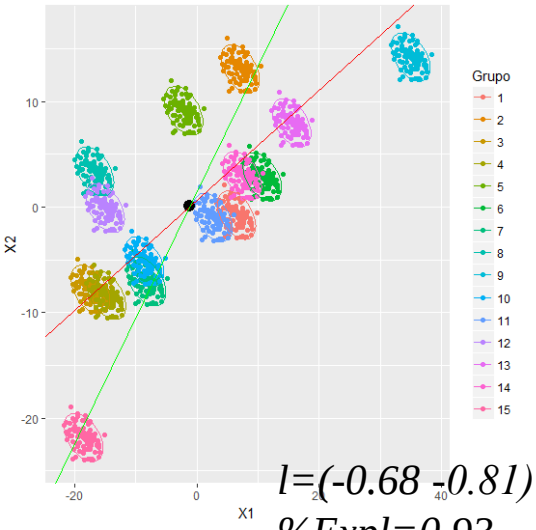


l=(-0.47 -0.56)'
%Expl=0.78

Exemplo 2



Exemplo 4



l=(-0.68 -0.81)'
%Expl=0.93

Preto:reta de MQ Vermelho:vetor de CP Verde:vetor discriminante

Biplots

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

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

Matriz de "escores dos CP"

Matriz de "pesos"

$YY' = U\Lambda U'$ Análise em $\Re^{nxn}$

$YY' = U\Lambda U'$ Análise em $\Re^{pxp}$

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

$m=2$

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

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

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*

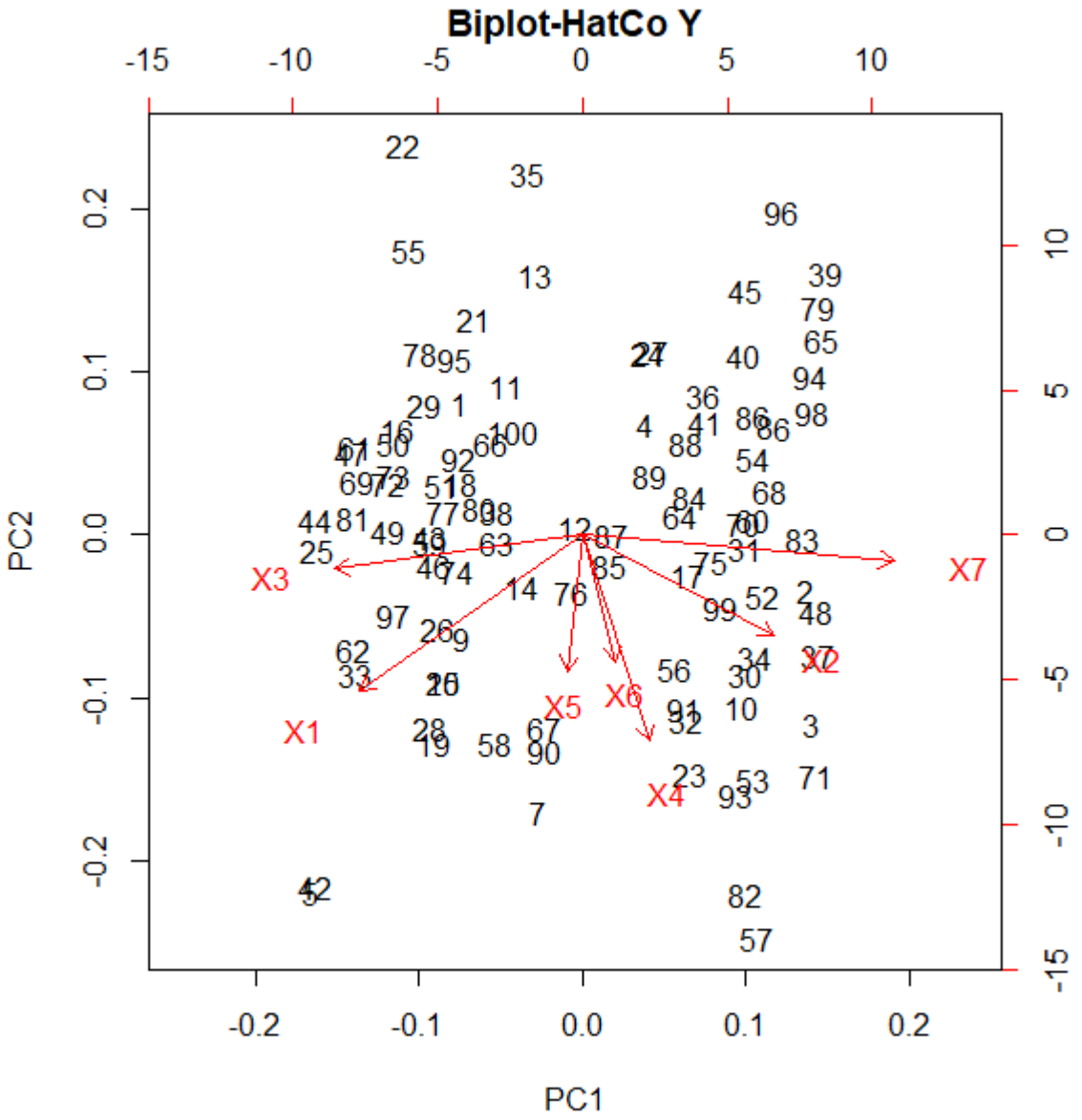
c=0: linhas em coordenadas principais e colunas em coordenadas padronizadas
c=1: linhas em coordenadas padronizadas e colunas em coordenadas principais
c=1/2: representação interativa

Representação Biplot

Dados HatCo (Hair, 2005)
Y: dados originais
n =100
p = 7
Variância Total=10,048

Análise de Componentes principais

	PC1	PC2
DesvioPadrão	2.1627	1.4567
VarExpl	0.4655	0.2112
VarExplAc	0.4655	0.6767

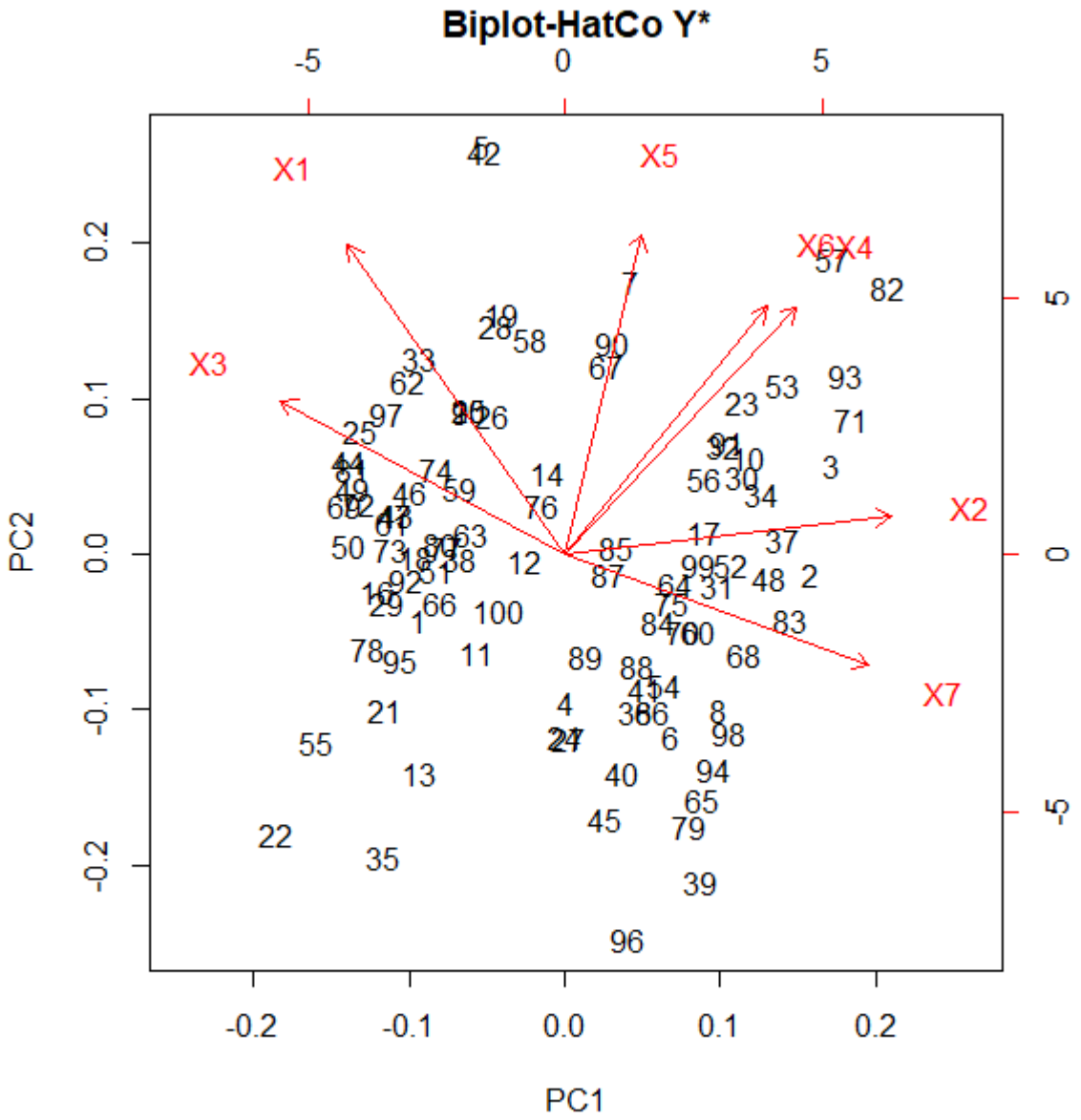


Representação Biplot

Dados HatCo (Hair, 2005)
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



Dúvidas?

Por favor,

Peço que cada aluno me envie, por meio do Chat, pelo menos “1” pergunta sobre o conteúdo abordado até o momento.

Ainda, temos tempo para discutir a Lista 03?