MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Sem/2020 - IME

Análise Multivariada

Já vimos ☺

- Análise Descritiva Multivariada
- Elipsóides de Dispersão e de Confiança, MANOVA
- Metodologias Clássicas de redução de dimensionalidade: $Y_{n\times p}$; $\Re^p \rightarrow \Re^m$
  - ✓ ACP *(m≤min(n,p))*, ACoP *(m≤min(n,p))*, AC *(m≤min(I-1,J-1))*, AF *(m≤min(n,p))*
  - ✓ AG
  - ✓ AD *(m≤min(n,p,G-1))*, ACC *(m≤min(n,p,q))*
  - ✓ Soluções Duais ($\Re^{nxp}$, $\Re^{pxp}$, $\Re^{nxn}$), Representações Biplot
  - ✓ PCR (Regressão via CP), PLS (Regressão via MQ Parciais)
- Integração de Bancos de Dados
- Diagrama de Caminhos ou Grafos

Var. quantitativas
n > p
Obs iid

Caso: n << p (*big p*): Soluções Regularizadas e Penalizadas

⇒ Componentes Principais
⇒ Análise Discriminante
⇒ Correlação Canônica

Métodos via Fatoração de Matrizes
e via Modelos de Regressão!

O Problema n << p

Big data
Big-p

- **Dados "gasoline"**: $[Y_{60\times1} \quad X_{60\times401}]$

- **Dados dos 3 Experts**:

$$[Y1_{6\times3} \quad Y2_{6\times4} \quad Y3_{6\times3}] = Y_{6\times10}$$

TABLE 1.
Wine tasting data from Abdi and Valentin (2007).

		Expert 1			Expert 2				Expert 3		
Wine	Oak-type	Fruity	Woody	Coffee	Red fruit	Roasted	Vanillin	Woody	Fruity	Butter	Woody
1	1	1	6	7	2	5	7	6	3	6	7
2	2	5	3	2	4	4	4	2	4	4	3
3	2	6	1	1	5	2	1	1	7	1	1
4	2	7	1	2	7	2	1	2	2	2	2
5	1	2	5	4	3	5	6	5	2	6	6
6	1	3	4	4	3	5	4	5	1	7	5

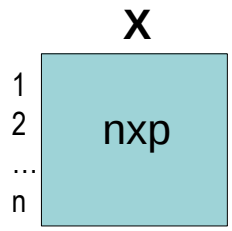
- **Dados dos Ratos Congenitos**: $Y_{50\times35.129}$

- **Dados dos Transcriptomas**: $Y_{189\times22.215}$

```
library(devtools)
install_github("genomicsclass/tissuesGeneExpression")
library(tissuesGeneExpression)
data(tissuesGeneExpression)
dim(e)   ## e contains the expression data
## [1] 22215   189
```

Redução de Dimensionalidade: n>p ⇒ n<<p

Big data
Big-p

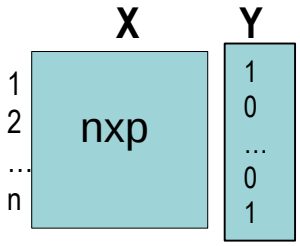


- **Componentes Principais e Coordenadas Principais:**

$$\Sigma_{p\times p} = VDV' \;;\quad F_k = XV_k$$
$$X_{n\times p} = UD^{1/2}V';\quad F_k = U_k d_k^{1/2}$$

$$\left\{ \max_a \frac{a'\Sigma a}{a'a},\ a'a = 1 \right.$$

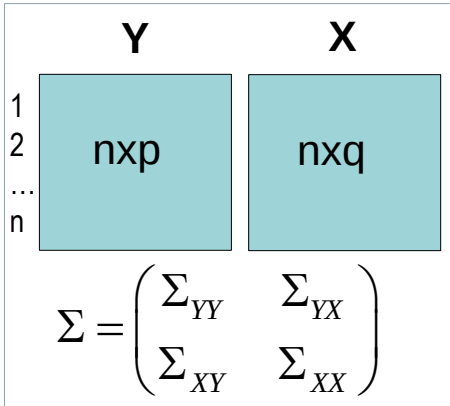
d_k>0? (min(n,p))
V_{jk}=0?



- **Análise Discriminante (Linear de Fisher):**

$$\Sigma_{p\times p} = \Sigma_B + \Sigma_W \;\Rightarrow\; \max_a \frac{a'\Sigma_B a}{a'\Sigma_W a},\ a'\Sigma_w a = 1$$

Σ_w inversível?



$$\Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}$$

- **Correlação Canônica:**

$$\left\{ \begin{array}{l} F_Y = a'Y;\quad \max_a \dfrac{a'\boldsymbol{\Sigma}_{12}\Sigma_{22}^{-1}\Sigma_{21}a}{a'\Sigma_{11}a},\quad a'\Sigma_{11}a = 1 \\ F_X = b'X;\quad \max_b \dfrac{b'\boldsymbol{\Sigma}_{21}\Sigma_{11}^{-1}\Sigma_{12}b}{b'\Sigma_{22}b},\quad b'\Sigma_{22}b = 1 \end{array} \right.$$

Σ_{11} inversível?

Σ_{22} inversível?

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

Componentes Principais Esparsos

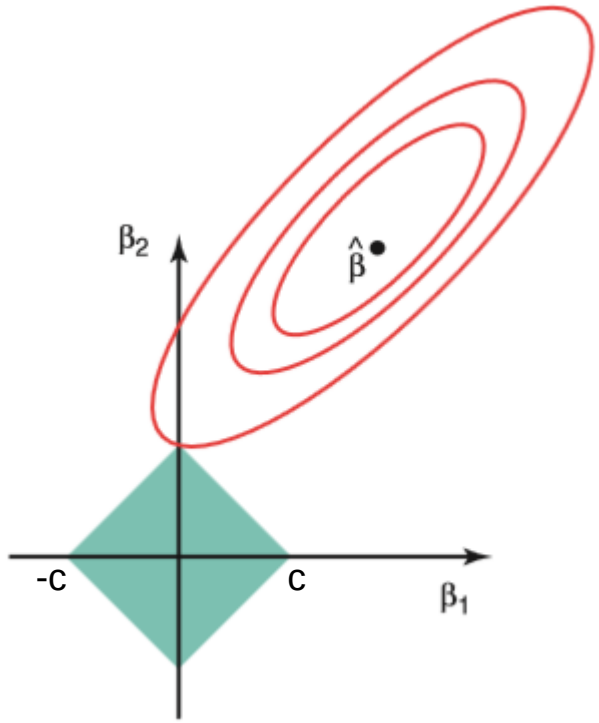
CP Penalizado (LASSO)

Penalização na forma de Lagrange:

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

Penalização na forma de restrição:

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



Solução mais esparsa: $\beta_1 = 0$

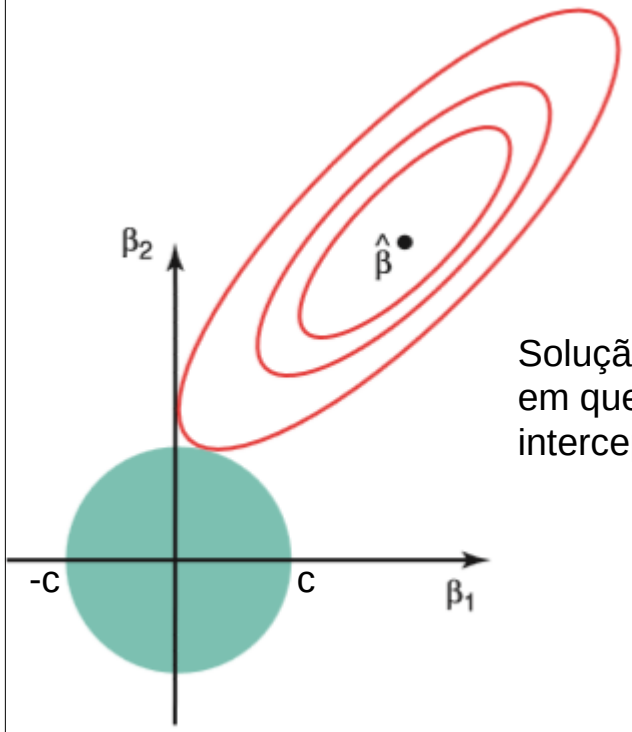
CP Regularizado (Ridge Regression)

Regularização na forma de Lagrange:

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

Regularização na forma de restrição:

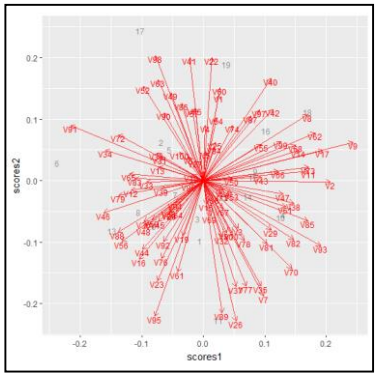
$$\hat{\beta}_{2\times 1}; \min_{\beta} \sum_{i=1}^{n} \left(F_{ik} - Y_i' \beta \right)^2 \text{ sujeito a } \beta_1^2 + \beta_2^2 \le c$$



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

Solução menos esparsa: $\beta_1 \approx 0$

Componentes Principais em Espaços n<<p (*Big-p)*



Redução de dimensionalidade: Soluções regularizadas e penalizadas

$$Y_{n\times p} = UD^{1/2}V' \Rightarrow F_k = U_k d_k^{1/2} = YV_k \overset{n<<p}{\Rightarrow} \hat{F}_k = Y\hat{v}_k$$

(CoP: $UD^{1/2}V'$; CP: $U_k d_k^{1/2}$)

Componente Principal Esparso (*Elastic Net;* Zou et al., 2006))

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|F_k - Y\beta\|_2^2 + \lambda_1 \|\beta\|_2^2 + \lambda_2 \|\beta\|_1 \right\}; \quad \hat{v}_k = \frac{\hat{\beta}}{\|\hat{\beta}\|_2}; \quad \hat{F}_k = Y\hat{v}_k$$

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

$$B_{p\times m} = (\beta_1, \ldots, \beta_m)$$

Componente Principal Esparso via *svd* (Witten et al., 2009)

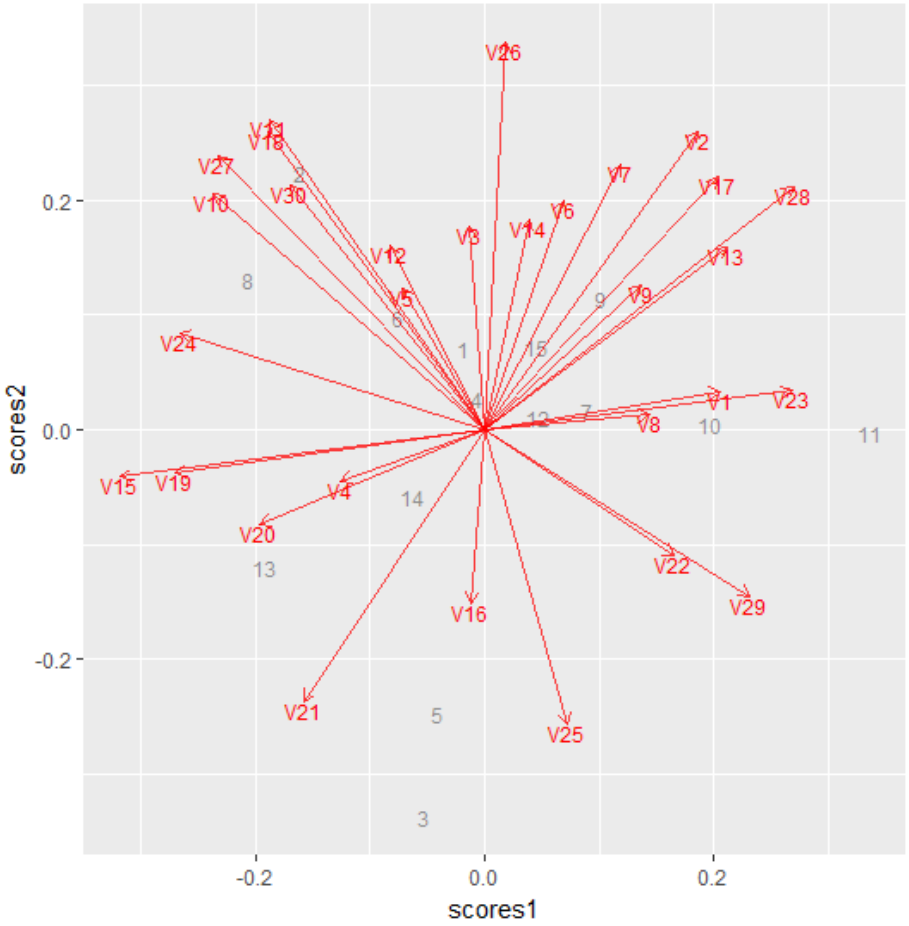
$$\max_{U_k, V_k} U_k' Y V_k; \quad \begin{cases} \|U_k\|_2^2 \le 1 \\ \|V_k\|_2^2 \le 1, \ \|V_k\|_1 \le c_1 \end{cases} \qquad \tilde{F}_k = Y\tilde{V}_k$$

Componentes Principais – n<<p

Biplot: n=15 p=30

CoP: Coordenadas Principais

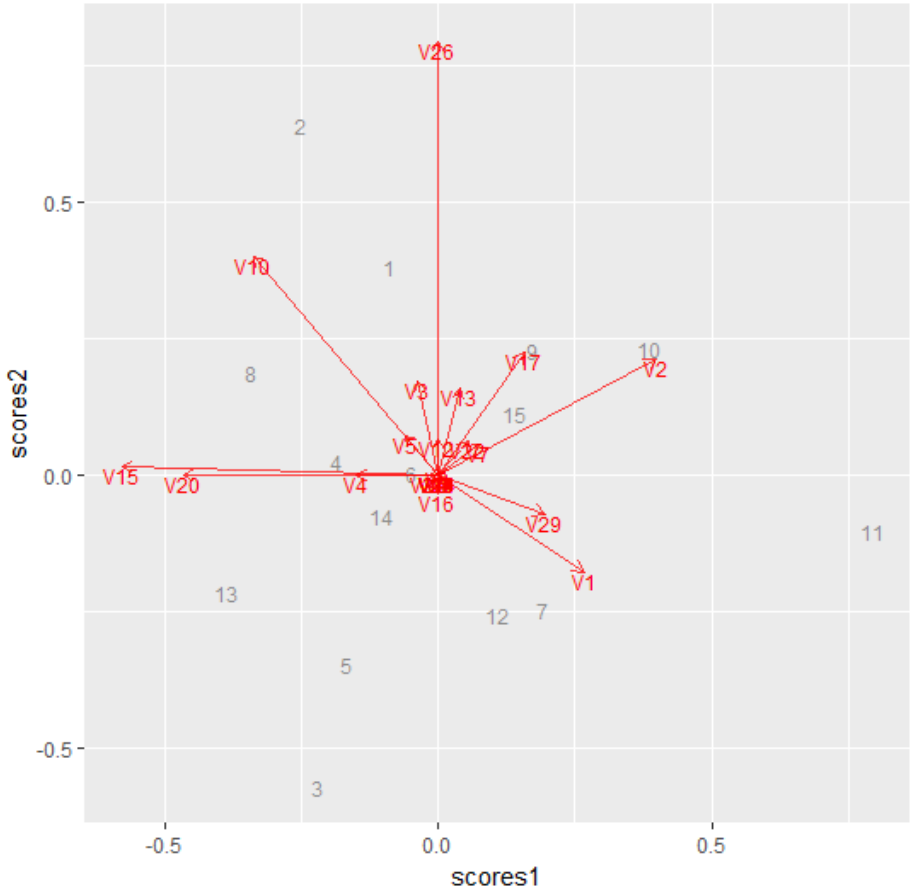
R-prcomp (suporta n<p)



Biplot: n=15, p=30

sCP: CP Esparso

R-SPCA do pacote ElasticNet



Big Data – n << p

Dados: Breast.TCGA do Bioconductor do R

breast.TCGA$data.train$mRNA

```
[1] 150 200
```

Big-p

breast.TCGA$data.train$miRNA

```
[1] 150 184
```

Big-p

breast.TCGA$data.train$proteomics

```
[1] 150 142
```

n>p

breast.TCGA$data.train$subtype

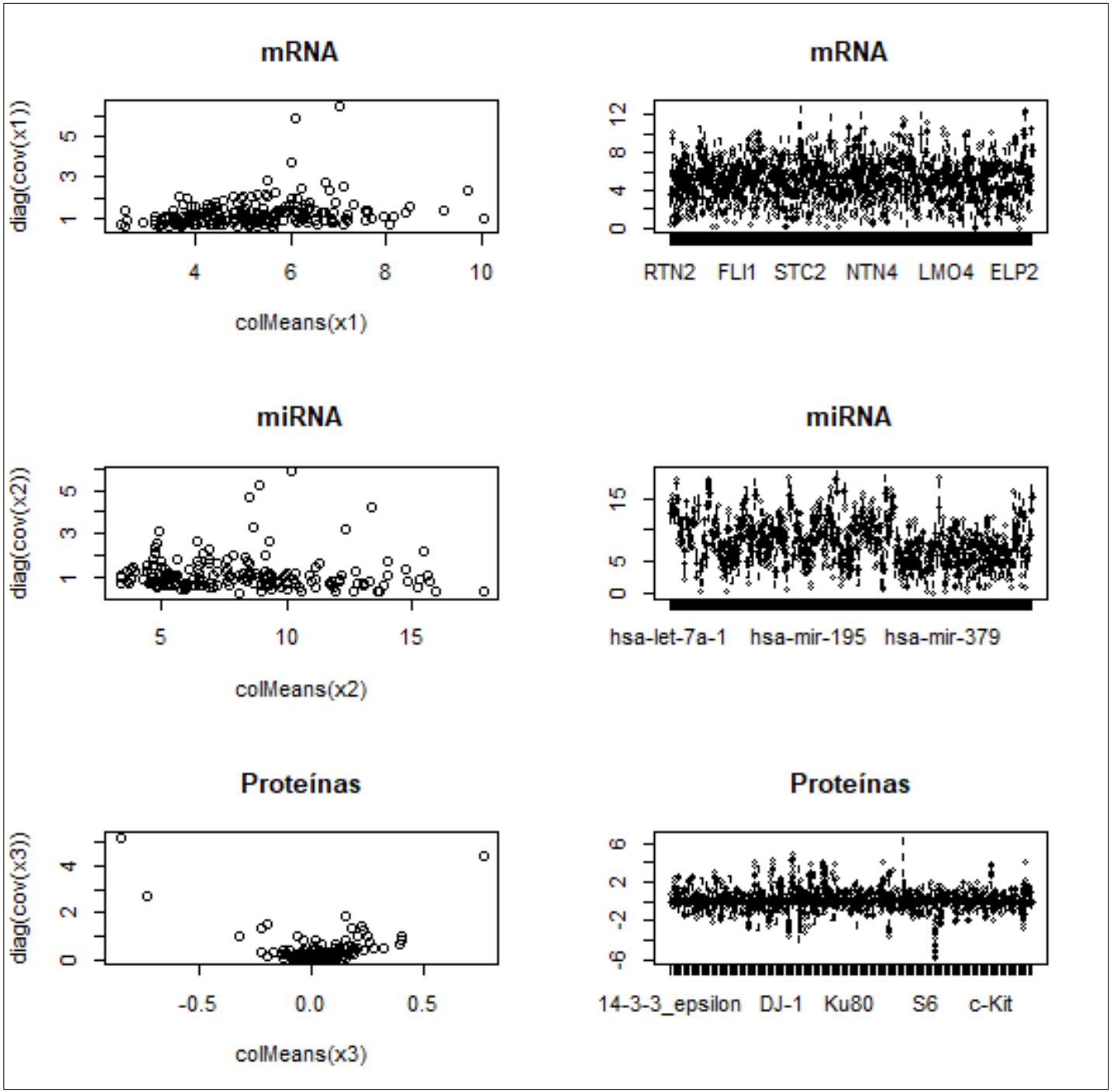
```
Basal  Her2  LumA
   45    30    75
```

Dados breast.TCGA

Dispersão: média x var

Box-plots

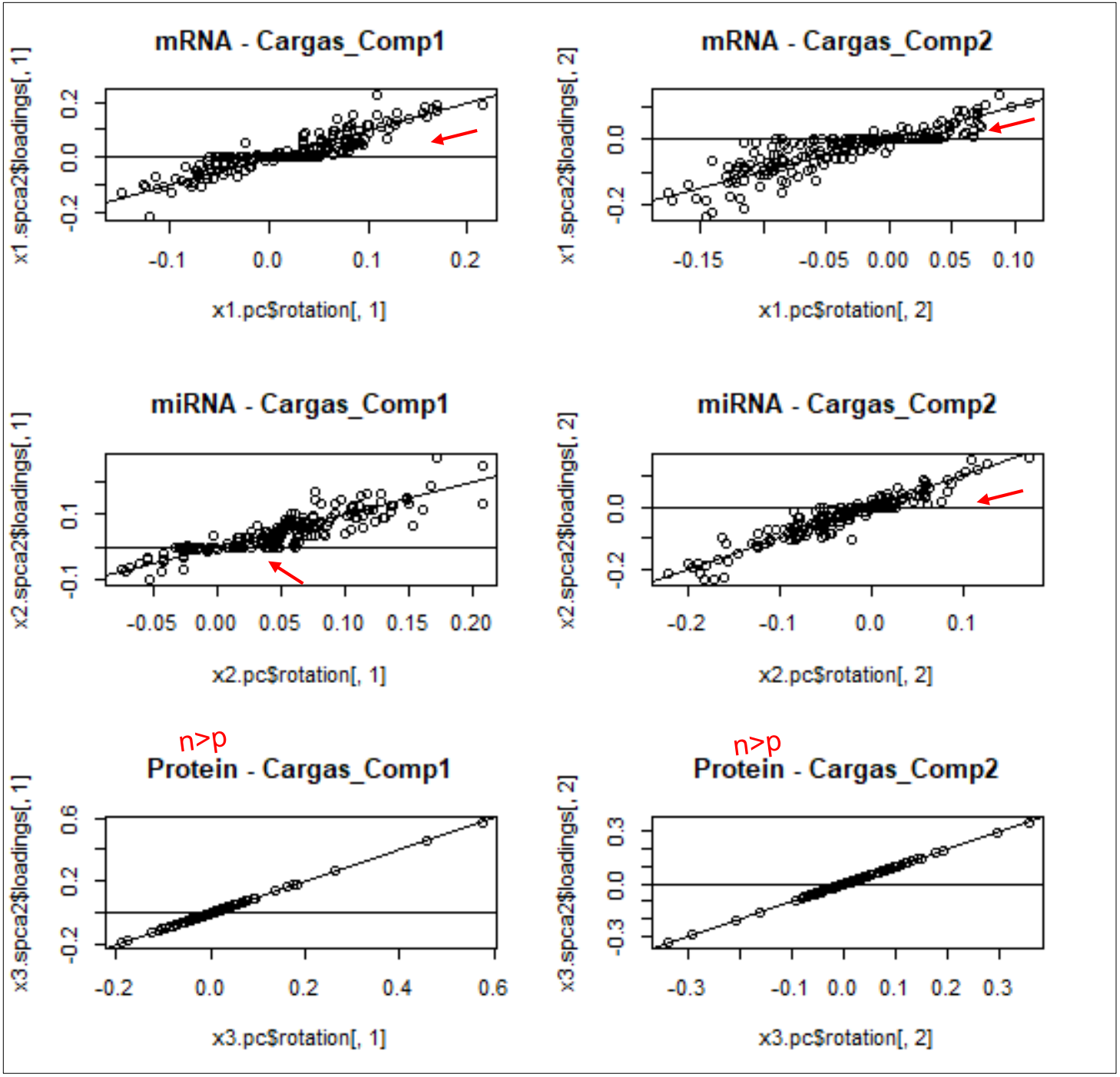
Avaliação da escala: superdispersão, obs atípicas



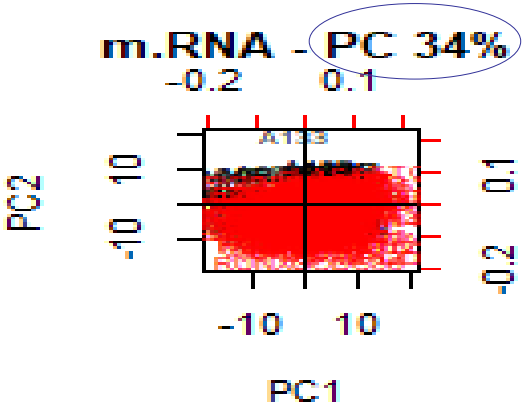
Componentes Principais (`CoP: prcomp`)

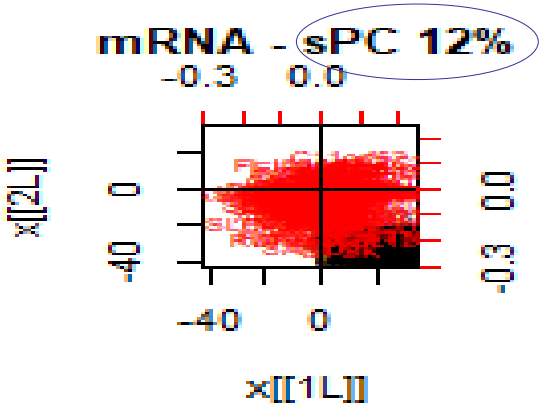
Componentes Principais Esparsos (`sCP:elasticNet`)

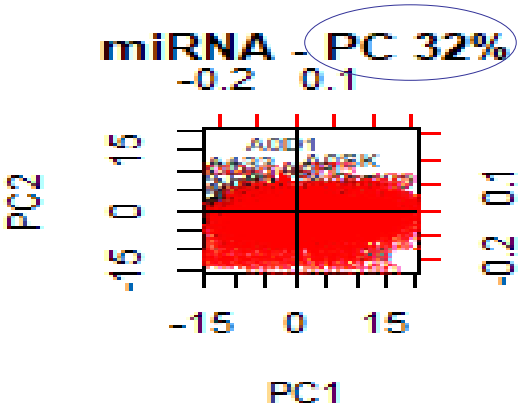
Comparação das Cargas

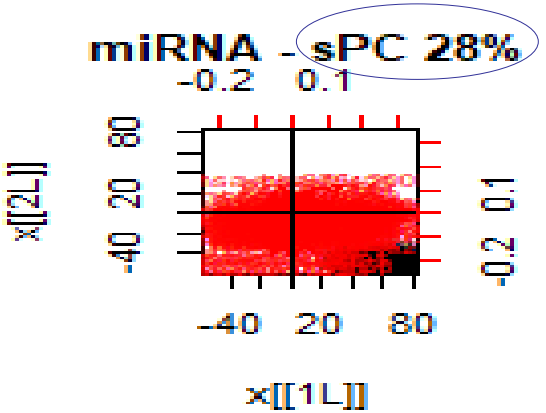


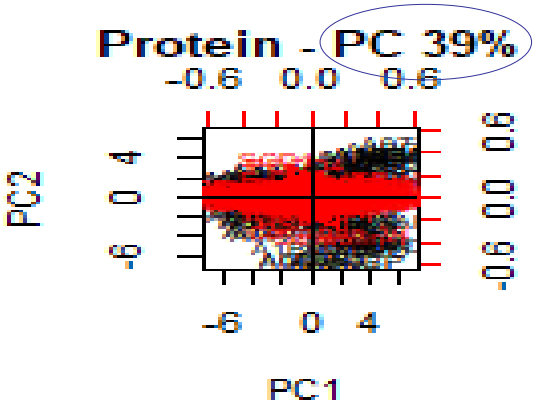
Biplots

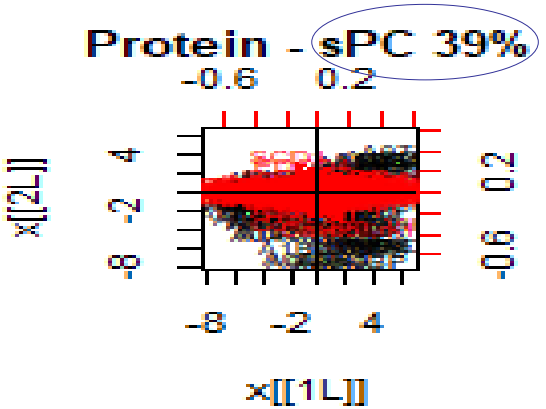












- Indivíduos - Variáveis

Análise Discriminante Esparsa – n << p

Amostra do grupo g $\quad Y_{i\ p\times 1} \mid \tau_g \overset{iid}{\sim} \left(\mu_g ; \Sigma_g\right) \Rightarrow \hat{F}_i = l' Y_i \overset{iid}{\sim} \left(l'\mu_g ; l'\Sigma l\right)$

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

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

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

$$\begin{cases} \max_l l' S_B l \ ; & l_k' S_W l_k = 1, \\ & l_j' S_W l_k = 0,\ j \neq k \\ k = 1, \ldots, m \leq \min\left(n, p, G-1\right) \end{cases}$$

Tabela de MANOVA

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

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

$$\max_{\beta_k} \frac{\beta_k' S_B \beta_k}{\beta_k' S_W \beta_k}; \quad \left| S_W^{-1/2} S_B S_W^{-1/2} - dI_p \right| = 0 \overset{n<<p}{\Rightarrow} S_W \text{ não é p.d (?)}$$

Tornar os autovalores positivos

Solução regularizada para corrigir o posto de S_W

$$\max_{\beta_k} \frac{\beta_k' S_B \beta_k}{\beta_k' (S_W + \Omega) \beta_k}; \quad \left| (S_W + \Omega) - dI_p \right| = 0, \; d > 0$$

$$\Rightarrow \max_{\tilde{\beta}_k} \tilde{\beta}_k' \tilde{S}_B \tilde{\beta}_k \begin{cases} \tilde{S}_B = (S_W + \Omega)^{-1/2} S_B (S_W + \Omega)^{-1/2} \\ \beta_k' (S_W + \Omega) \beta_k' = 1; \; \beta_k' (S_W + \Omega) \beta_j' = 0; \end{cases}$$

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

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

Solução Regularizada e Penalizada: implementada em *sda* do pacote `sparseLDA_R`

Problema de predição de grupos

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

Vetor de pesos dos grupos

Para $\Omega = I_m$: usar algoritmo do *ElasticNet*

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*"

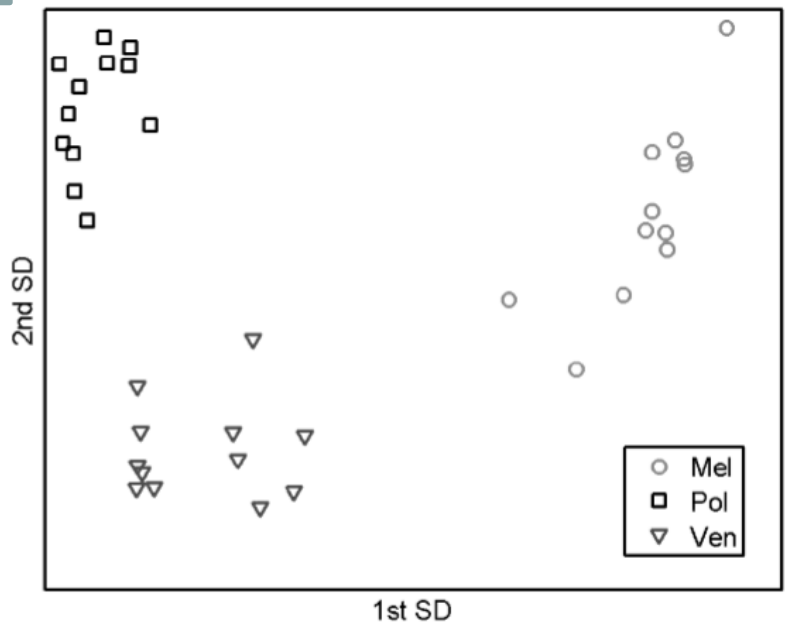
	$Y\hat{\beta}_1$	$Y\hat{\beta}_2$
[1,]	-3.113028	-3.4122173
[2,]	-3.142295	-3.8733571
[3,]	-2.988152	-1.4446112
[4,]	-2.995266	-1.5567002
[5,]	-2.995715	-1.5637771
[6,]	-3.063970	-2.6392373
[7,]	-2.996005	-1.5683428
[8,]	-2.998670	-1.6103476
[9,]	5.059007	0.9703178
[10,]	5.803007	0.7348022
[11,]	2.720020	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

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

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

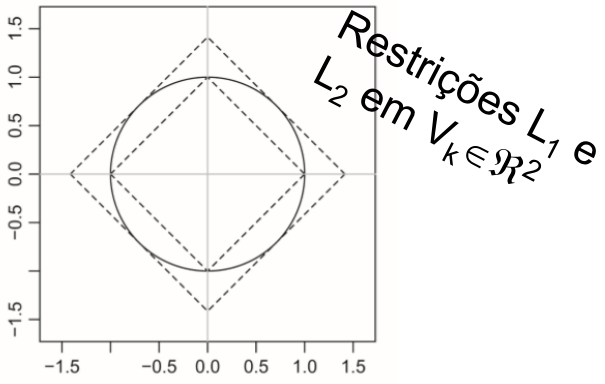
Problemas na otimização quando n<<p n<<q

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

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

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

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



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)

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

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_R`:

Dados gerados: n=100 p=500 q=1000: $\left[Y_{1\,100\times 500} \quad Y_{2\,100\times 1000} \right]$

$$U_{n\times 1} = Y_1 a \quad \Rightarrow \quad n\left(a_j \neq 0\right) = 338;$$

$$V_{n\times 1} = Y_2 b \quad \Rightarrow \quad n\left(b_j \neq 0\right) = 687; \; c_1 = c_2 = 0,5667$$

$$\rho_c = 0,9735$$

```
Set.seed(19)
Set.seed(90)
Num non-zeros u's:  338
Num non-zeros v's:  687
Type of x:  standard
Type of z:  standard
Penalty for x: L1 bound is  0.5666667
Penalty for z: L1 bound is  0.5666667
Cor(Xu,Zv):  0.9735552
```