MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Sem/2022 - IME

Análise Multivariada

Já vimos ☺

- Análise Descritiva Multivariada, Elipsóides de Concentração de Dados
- Algumas Distribuições Multivariadas, Elipsóides de Confiança, MANOVA
- Técnicas de Análise Multivariada: $Y_{n\times p} = \left(Y_{ij}\right) \in \Re^{n\times p}$ $\Re^{p} \to \Re^{m}$

- Foco na obtenção de Vetores Reducionistas (Escores e cargas)

- Soluções Duais ($\Re^{nxp}$, $\Re^{pxp}$, $\Re^{nxn}$), Representações Biplot

✓ ACP, ACoP, AC, AF

✓ AG

✓ AD

✓ ACC

n > p
Obs Independentes
Var Quantitativas

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

$\Rightarrow$ Componentes Principais (ACP)
$\Rightarrow$ Análise Discriminante (AD)
$\Rightarrow$ Correlação Canônica (ACC)

O Problema n << p

Big data
Big-p

- **Dados dos 3 Experts**:

$$\begin{bmatrix} Y1_{6\times 3} & Y2_{6\times 4} & Y3_{6\times 3} \end{bmatrix} = \boxed{Y_{6\times 10}}$$

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 Transcriptomas**: $Y_{189\times 22.215}$

```
library(devtools)
install_github("genomicsclass/tissuesGeneExpression")
library(tissuesGeneExpression)
data(tissuesGeneExpression)
dim(e)  ## e: contains the expression data
## n=189   p=22215
```

O Problema n << p

Big data
Big-p

- **Dados Breast.TCGA (Bioconductor do R)**

```
X1: breast.TCGA$data.train$mRNA
[1] 150 200
```
Big-p

```
X2: breast.TCGA$data.train$miRNA
[1] 150 184
```
Big-p

```
X3: breast.TCGA$data.train$proteomics
[1] 150 142
```
n>p

```
breast.TCGA$data.train$subtype
Basal   Her2   LumA
   45     30     75   =   150
```

- **Dados** "pencilliumYES" (sparseLDA do R): fungos de 3 espécies são avaliados em variáveis de imagem

$Y_{36\times 3.754}$

```
n=36  p=3.754
G=3: "P. Melanoconidium", "P. Polonicum", "P. Venetum"
```

Normas em Espaços Vetoriais

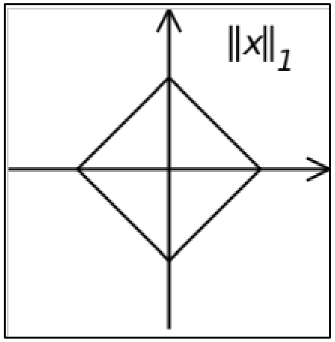
Toda norma || · || induz a uma métrica *d* em espaços vetoriais:

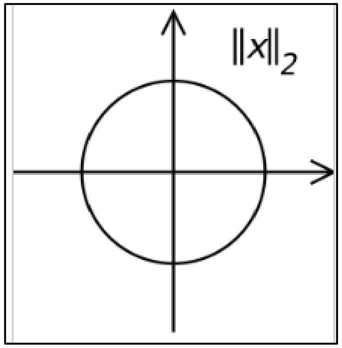
$$\|x-y\| = d(x,y); \qquad \|x\| = d(x,O)$$

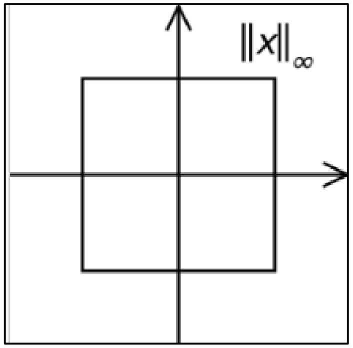
Para $x = (x_1, \dots, x_n) \in \Re^n$ tem-se as normas em Lq:

$$\|x\|_q = \left(\sum_{i=1}^{n} |x_i|^q\right)^{1/q}, \quad 1 \le q < \infty$$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}; \quad \|x\|_2^2 = \sum_{i=1}^{n} |x_i|^2 \qquad \|x\|_\infty = \max_i |x_i|$$







Componentes Principais – n << p

Big data
Big-p

CP CoP

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

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

Já vimos!

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 regularizadas (corrige autovalores negativos) ou soluções penalizadas (eliminar variáveis, isto é, obter autovetores V que atribuem carga “nula” para algumas variáveis)

Normas em Espaços Vetoriais

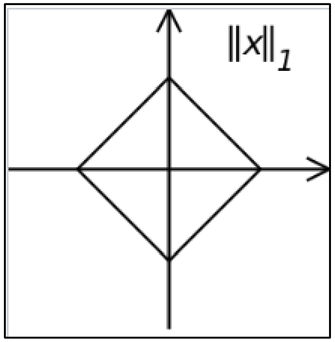
Toda norma || · || induz a uma métrica *d* em espaços vetoriais:

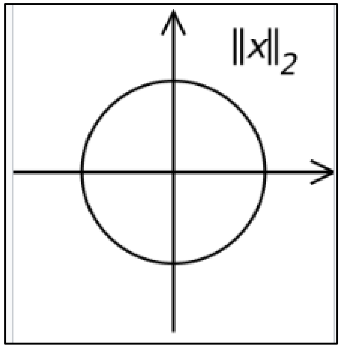
$$\|x-y\| = d(x,y); \qquad \|x\| = d(x,O)$$

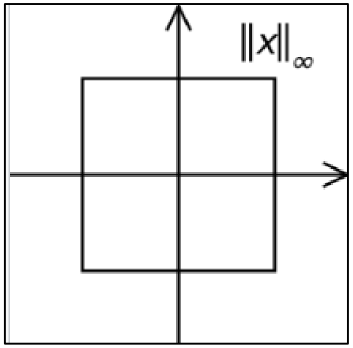
Para $x = (x_1, \ldots, x_n) \in \Re^n$ tem-se as normas em Lq:

$$\|x\|_q = \left(\sum_{i=1}^{n} |x_i|^q \right)^{1/q}, \quad 1 \le q < \infty$$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}; \quad \|x\|_2^2 = \sum_{i=1}^{n} |x_i|^2 \qquad \|x\|_\infty = \max_i |x_i|$$







Componentes Principais – n << p

Big data
Big-p

É facilmente obtido

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

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

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

Como obter os autovetores tais que, $V_j \cong \left(v_{1j}, v_{2j}, \ldots, v_{pj}\right); v_{kj} = 0$ para muitas coordenadas k, mantendo ainda uma alta porcentagem da variância total de Y explicada pelo j-ésimo componente?

Para esse fim, é importante **estabelecer a CORESPONDÊNCIA entre CP e Modelos de Regressão** (Zou, Hastie and Tibshirani, 2006):

CP Regularizado
(parâmetro de regularização λ)

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

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

Distância quadrática
de β à origem

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

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

$$\boxed{\hat{V}_j = \frac{\hat{\beta}}{\left\| \hat{\beta} \right\|_2}; \qquad \hat{Z}_j = Y\hat{V}_j}$$

- Componente Principal Penalizado (LASSO)

Limitação: o número de "cargas" não-nulas é no máximo *n*

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

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

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

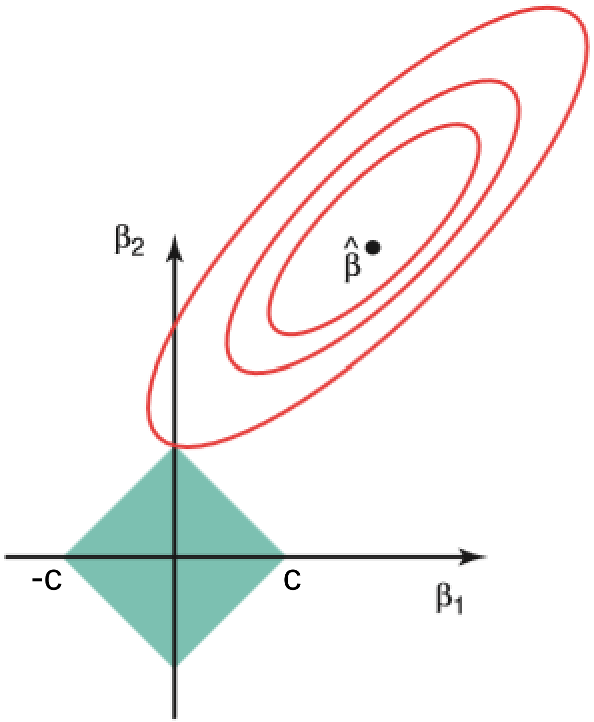
CP Penalizado (LASSO) x CP Regularizado (Ridge Regression)

Penalização na forma de Lagrange:

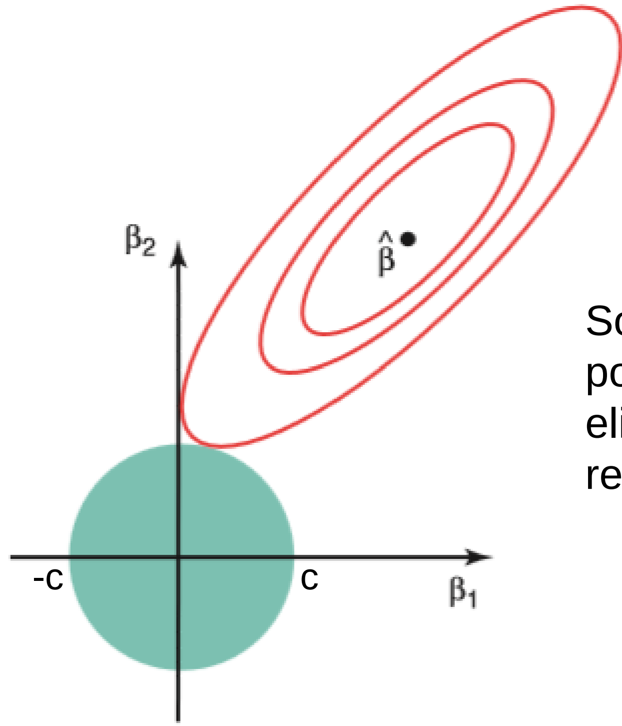
$$\hat{\beta} = \arg\min_{\beta} \left\| Z_j - Y\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1 \qquad \hat{\beta} = \arg\min_{\beta} \left\| Z_j - Y\beta \right\|_2^2 + \lambda \left\| \beta \right\|_2^2$$

Penalização na forma de restrição:

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



Solução mais esparsa: $\beta_1 = 0$



Solução menos esparsa: $\beta_1 \approx 0$

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

Componentes Principais – n << p

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

Componente Principal Regularizado e Penalizado obtidos diretamente das CoP da análise Dual (**Elastic Net**):

CoP: todas as var podem ser selecionadas (não há restrição para cargas nulas)

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

$$\hat{\beta} = \arg\min_{\beta}\left\{\underbrace{\left\|Z_j - Y\beta\right\|_2^2}_{\text{Solução de mínimos quadrados}} + \underbrace{\boxed{\lambda_1\left\|\beta\right\|_2^2}}_{\text{parâmetro de regularização}} + \underbrace{\boxed{\lambda_2\left\|\beta\right\|_1}}_{\text{parâmetro de penalização}}\right\}; \quad \hat{v}_j = \frac{\hat{\beta}}{\left\|\hat{\beta}\right\|_2}; \boxed{\hat{Z}_j = Y\hat{v}_j}$$

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

$\lambda_1; \lambda_2$: em geral, obtidos por validação-cruzada

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

Formalização Geral de Componentes Principais via Modelos de Regressão

Obtenção de Um CPr

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

Obtenção de *m* CPr

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

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

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

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

parâmetro de regularização

m parâmetros de penalização

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

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

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

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

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

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

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

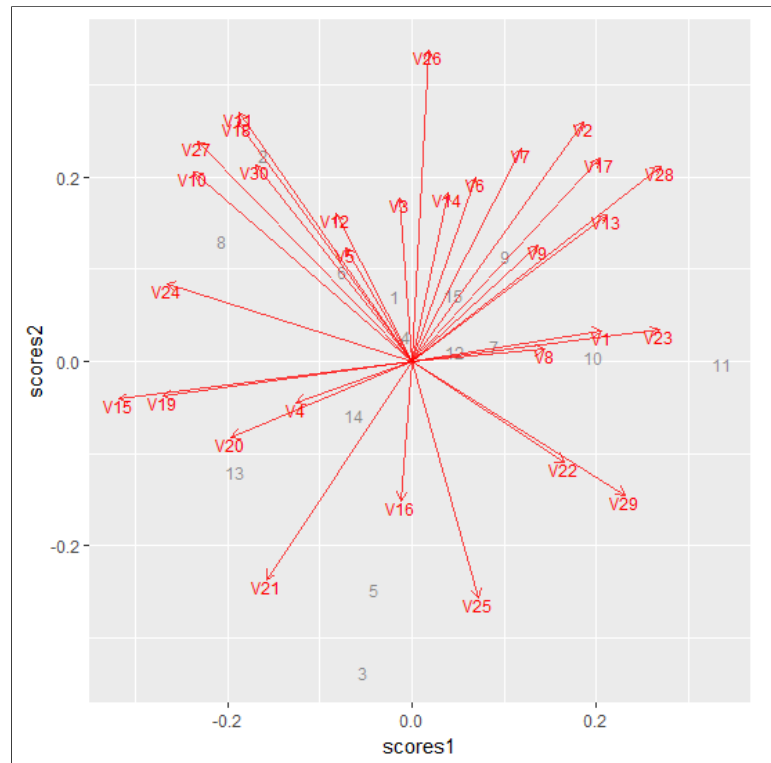
Componentes Principais – n<<p

Sparse Principal Component Analysis

Hui Zou, Trevor Hastie, and Robert Tibshirani

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

Dados Simulados
Biplot (n<p): n=15 p=30
R-prcomp



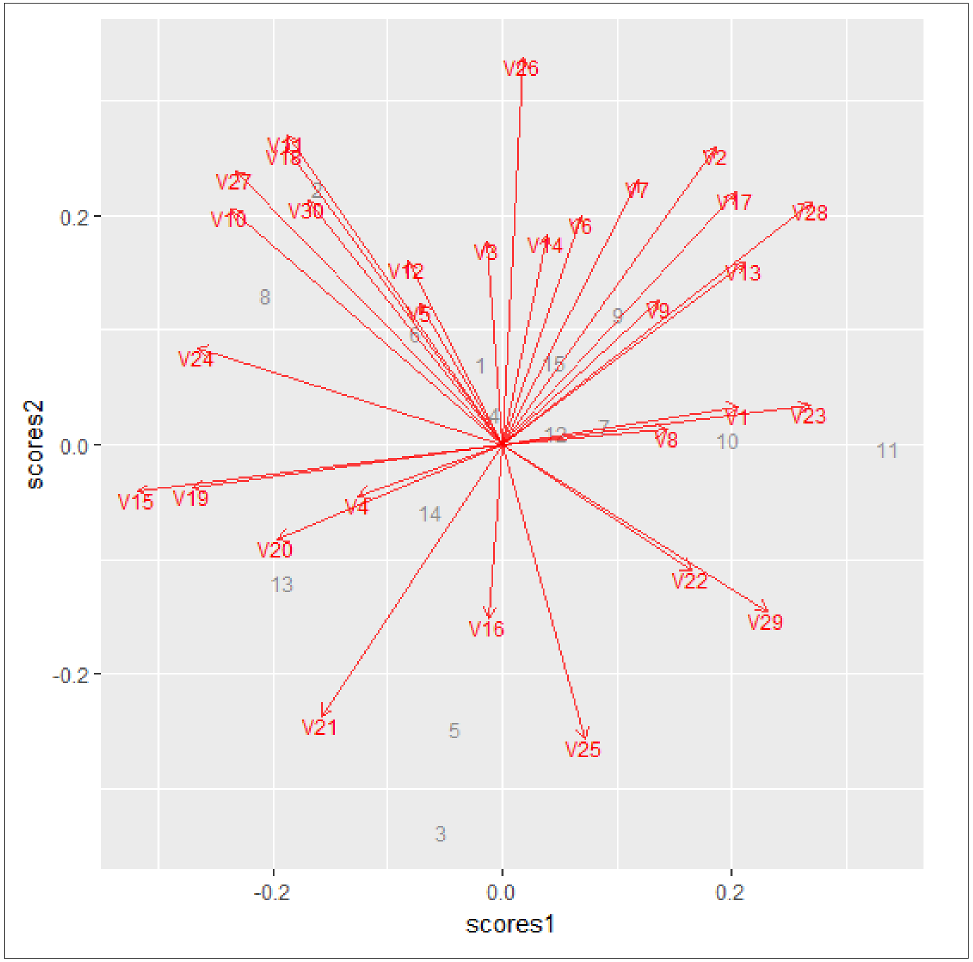
- *R-SPCA do pacote ElasticNet: Componentes Principais Esparsos*

Componentes Principais obtidos da análise dual.

Solução esparsa: obter autovetores com muitas coordenadas nulas!

Componentes Principais – n<<p

Componentes Principais obtidos da análise dual (prcomp-R)



Matriz de cargas

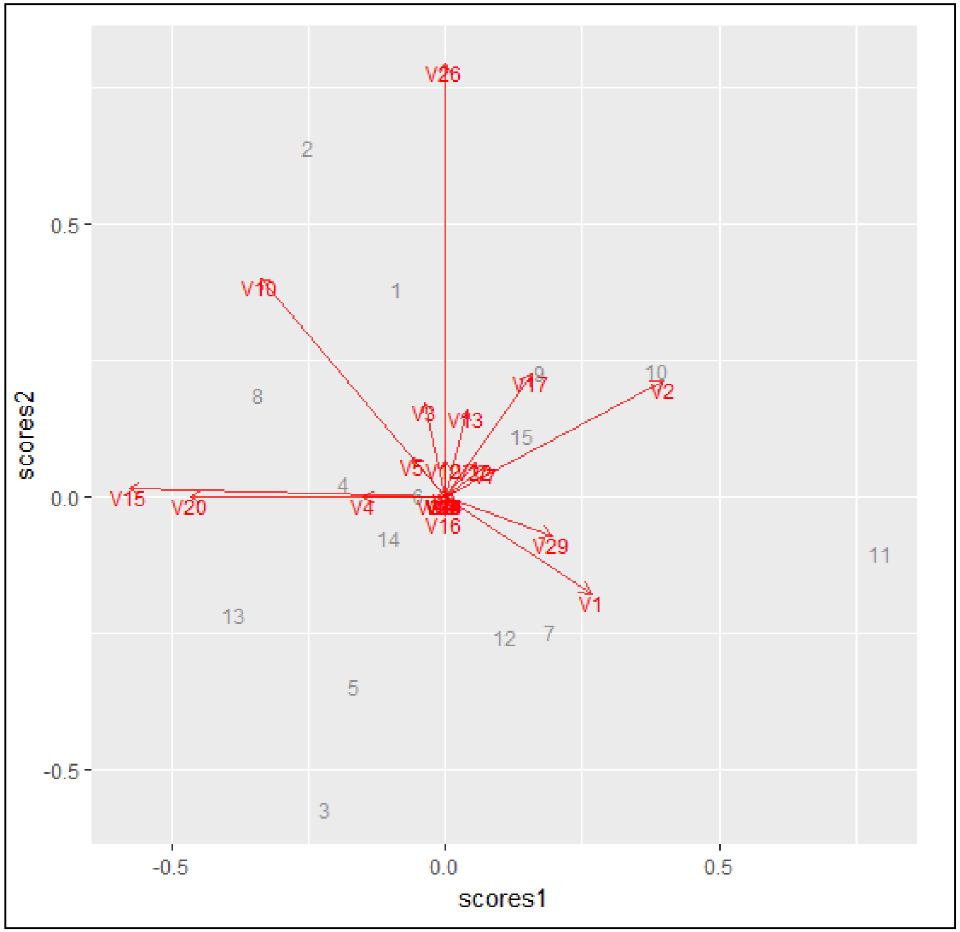
	PC1	PC2
V1	0.20486853	0.03338466
V2	0.18525221	0.26052241
V3	-0.01406721	0.17725332
V4	-0.12560728	-0.04474235
V5	-0.07185638	0.12382278
V6	0.06894920	0.20028957
V7	0.11822653	0.23051465
V8	0.14332703	0.01338871
V9	0.13705858	0.12591629
V10	-0.23708813	0.20630273
V11	-0.18710753	0.26974343
V12	-0.08342832	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

Solução esparsa: obter autovetores com muitas coordenadas nulas!

Componentes Principais – n<<p

Matriz de cargas com esparsidade

Biplot – CP Esparsos: n=15 p=30
(SPCA do pacote ElasticNet-R)



Sparse loadings

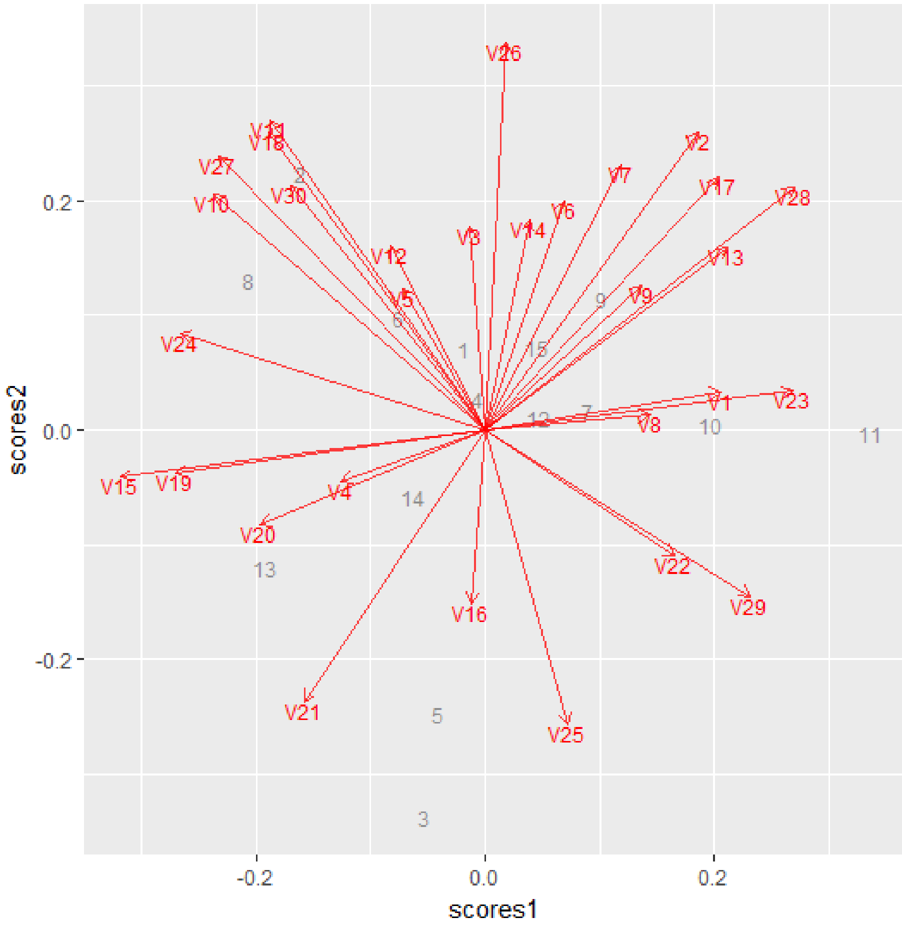
	PC1	PC2
V1	0.266	-0.177
V2	0.398	0.213
V3	-0.040	0.173
V4	-0.151	0.000
V5	-0.062	0.073
V6	0.000	0.000
V7	0.073	0.057
V8	0.000	0.010
V9	0.000	0.000
V10	-0.339	0.401
V11	0.000	0.000
V12	0.000	0.067
V13	0.040	0.160
V14	0.000	0.000
V15	-0.580	0.015
V16	0.000	-0.034
V17	0.157	0.225
V18	0.000	0.000
V19	0.000	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=15 p=30

CoP: Coordenadas Principais

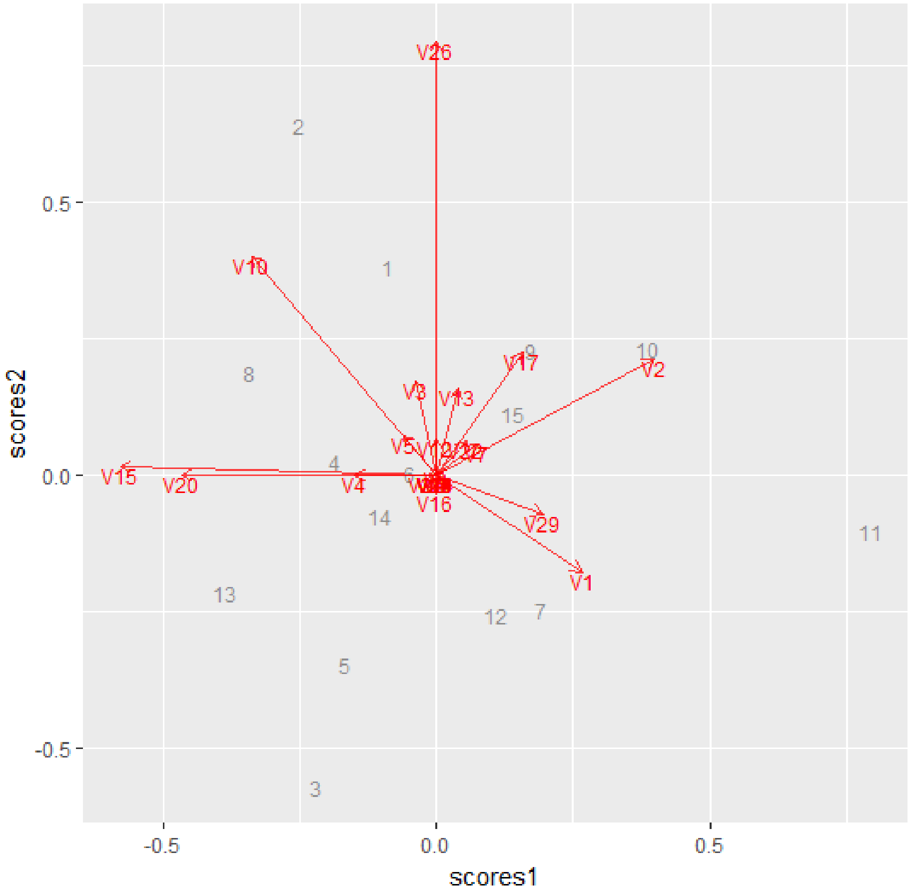
R-prcomp (suporta n<p)



Biplot: n=15, p=30

sCP: CP Esparso

R-SPCA do pacote ElasticNet



Big Data – n << p

```
Dados: Breast.TCGA do Bioconductor do R

X1: breast.TCGA$data.train$mRNA
[1] 150 200
```
Big-p

```
X2: breast.TCGA$data.train$miRNA
[1] 150 184
```
Big-p

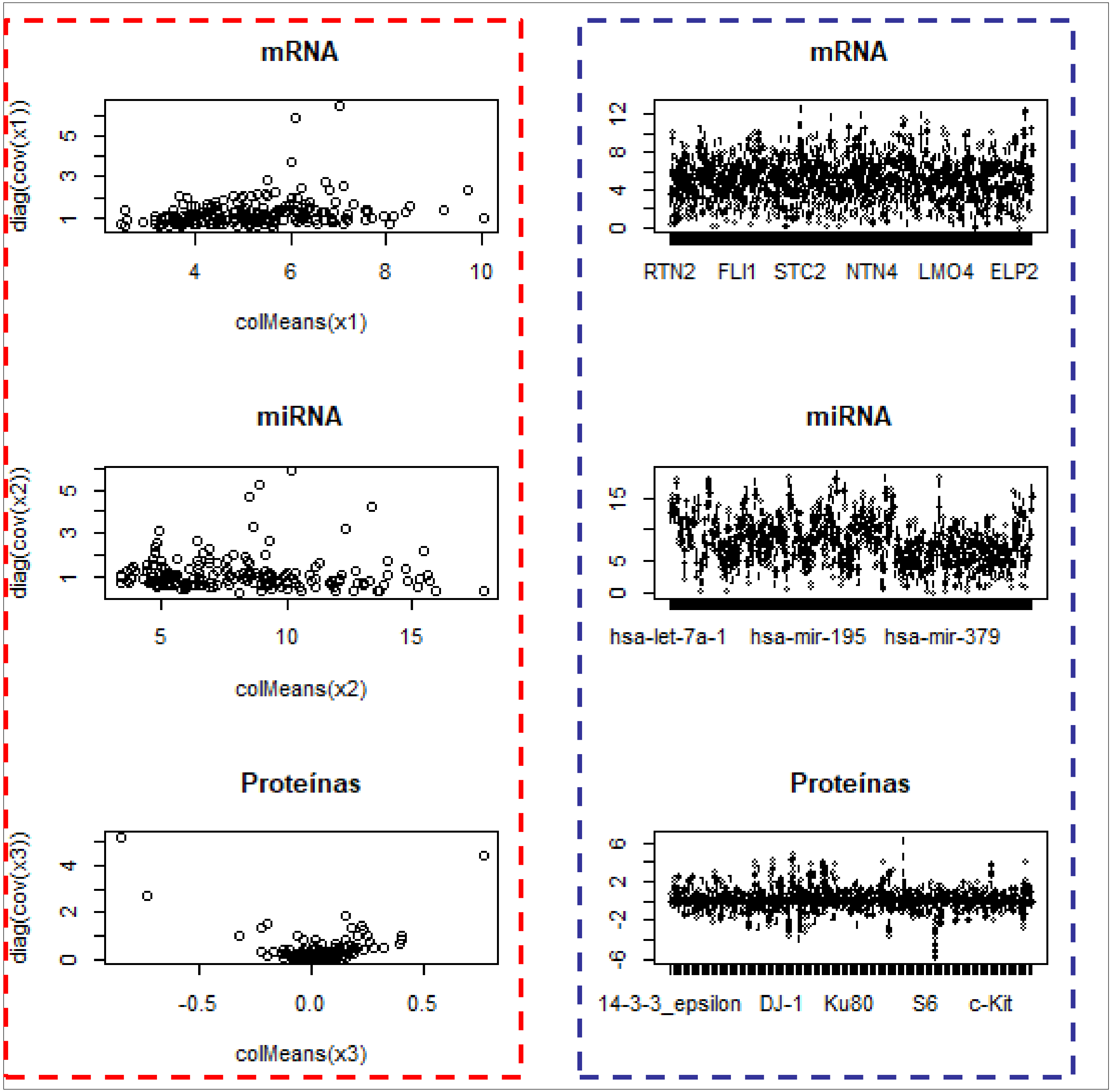
```
X3: breast.TCGA$data.train$proteomics
[1] 150 142
```
n>p

```
breast.TCGA$data.train$subtype
Basal  Her2  LumA
   45    30    75  =  150
```

Dados
breast.TCGA

Gráfico de Dispersão: Variância x Média

Box-plots

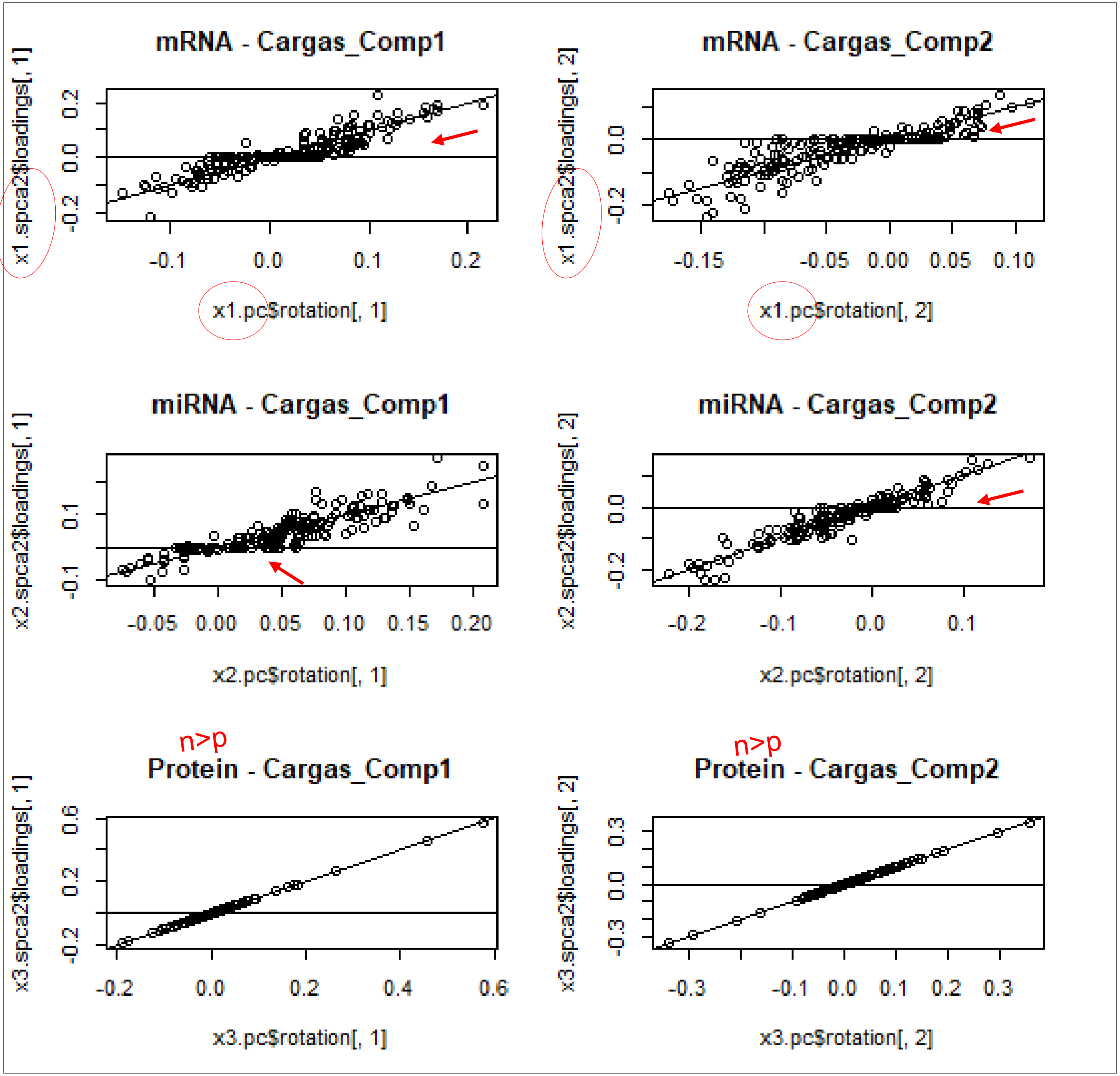


Comparação das Cargas

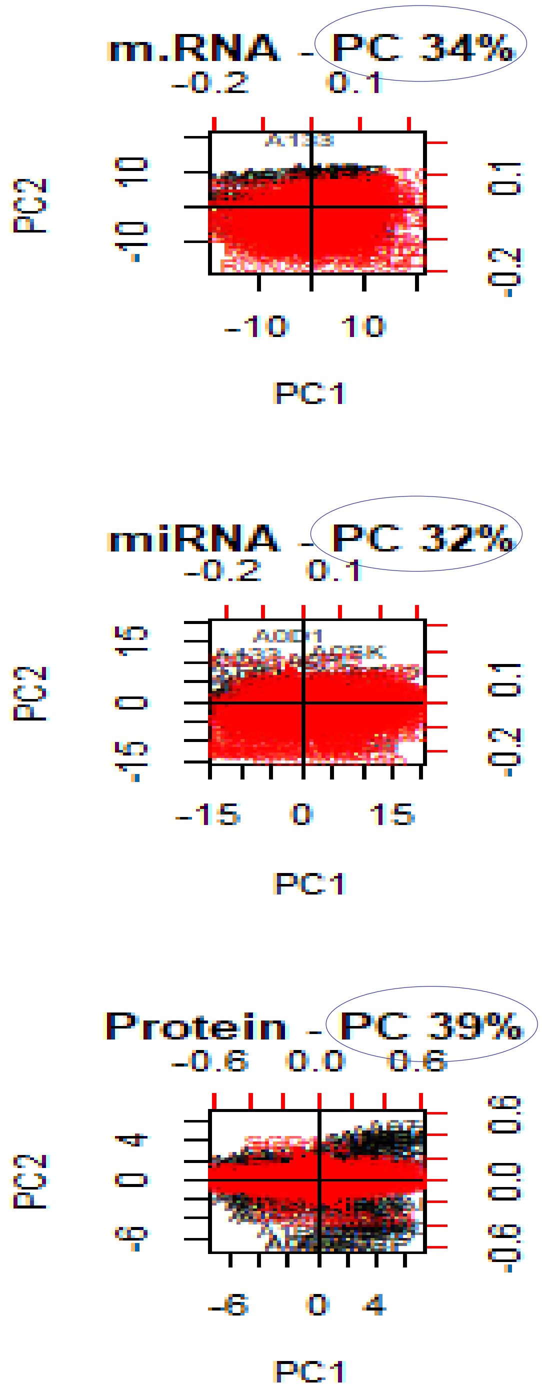
Componentes Principais (`CoP: prcomp`)

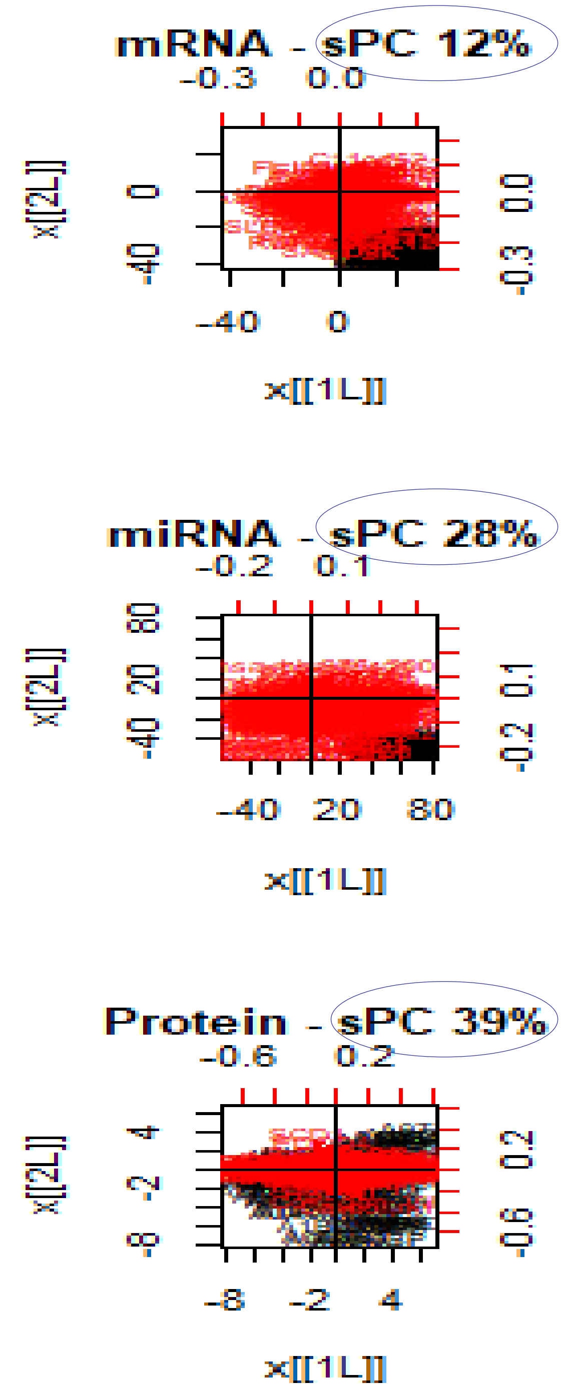
x

Componentes Principais Esparsos (`sCP:elasticNet`)



Biplots

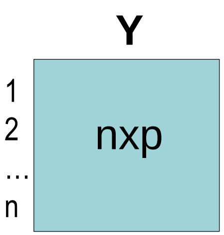




- Indivíduos - Variáveis

Redução de Dimensionalidade
Y → Fatores Latentes

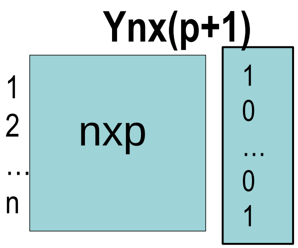
n>p
n<<p
(Big-p)



- **Componente Principal (F_k) e Coordenada Principal (F_k)**

$$\Sigma_{p\times p} = VDV' \;;\quad \boxed{F_k = YV_k}$$
$$Y_{n\times p} = UD^{1/2}V';\quad \boxed{F_k = U_k d_k^{1/2}}$$
$$\left\{ \max_a \frac{a'\Sigma a}{a'a},\quad a'a = 1 \right.$$

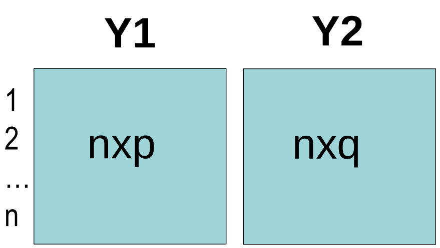
#(autovalores>0)
= min(n,p)
V_{jk}=0?



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

Σ_W inversível?

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



$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

- **Correlação Canônica:**

$$\left\{ \begin{array}{l} \boxed{F_{Y1} = a'Y1;} \quad \max_a \dfrac{a'\boldsymbol{\Sigma}_{12}\Sigma_{22}^{-1}\Sigma_{21}a}{a'\Sigma_{11}a},\quad a'\Sigma_{11}a = 1 \\ \\ \boxed{F_{Y2} = b'Y2;} \quad \max_b \dfrac{b'\boldsymbol{\Sigma}_{21}\Sigma_{11}^{-1}\Sigma_{12}b}{b'\Sigma_{22}b},\quad b'\Sigma_{22}b = 1 \end{array} \right.$$

Σ_{11} inversível?

Σ_{22} inversível?

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

Redução de dimensionalidade: Soluções regularizadas e penalizadas

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

nxm mxp

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

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

$$\arg\min_{A,B} \left\{ \sum_{i=1}^{n} \left\| Y_{i_{p\times 1}} - A_{p\times m} B'_{m\times p} Y_i \right\|_2^2 + \lambda_1 \sum_{k=1}^{m} \left\| \beta_k \right\|_2^2 + \sum_{k=1}^{m} \lambda_{2k} \left\| \beta_k \right\|_1 \right\}; \quad A'A = I_m; \quad B_{p\times m} = (\beta_1, \ldots, \beta_m)$$

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

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

Análise Discriminante Esparsa – n << p

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

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

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

$$\max_a \frac{a'\sum_{g=1}^{G} n_g \left(\bar{Y}_g - \bar{Y}_{.}\right)\left(\bar{Y}_g - \bar{Y}_{.}\right)' a}{a' S_W a} = \boxed{\max_a \frac{a' S_B a}{a' S_W a}} \left\{ \begin{array}{l} S_W^{-1/2} S_B S_W^{-1/2} = P\Lambda P'; \;\; a = S_W^{-1/2} P \\ a_k' S_W a_k = 1, \quad a_j' S_W a_k = 0,\; j \neq k \\ k = 1, \ldots, m \leq \min\left(n, p, G-1\right) \end{array} \right.$$

Tabela de MANOVA: Y=Xβ+e

$X_{n\times G}$: Matriz de incidência de grupos

$P_X = X\left(X'X\right)^{-1} X'$

F.V.	g.l.	Matriz de SQPC
Trat	**G-1**	$H_{p\times p} = \sum_{g=1}^{G} n_g \left(\bar{\mathbf{y}}_g - \bar{\mathbf{y}}\right)\left(\bar{\mathbf{y}}_g - \bar{\mathbf{y}}\right)' = S_B = Y'P_X Y$
Resíduo	**n-G**	$E_{p\times p} = \sum_{g=1}^{G}\sum_{i=1}^{n_g} \left(\mathbf{y}_{gi} - \bar{\mathbf{y}}_g\right)\left(\mathbf{y}_{gi} - \bar{\mathbf{y}}_g\right)' = SS_W = Y'\left(I - P_X\right)Y \Rightarrow S_W = SS_W / \left(n-G\right) = S_c$
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

1. Solução regularizada para corrigir o posto de S_W

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

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

Como obter Ω? Por validação cruzada. E os β's, são esparsos? Não.

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

2. Solução Regularizada e Penalizada: (em `sda` do pacote `sparseLDA_R`)

Problema de predição: G é matriz indicadora de grupos

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

Vetor (Gx1) 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, visualmente indistinguíveis, avaliadas em variáveis de imagem

n=36 (3 sp x 4 linhagens x 3 replicatas), p=3.754

G=3: "*P. Melanoconidium*", "*P. Polonicum*", "*P. Venetum*"

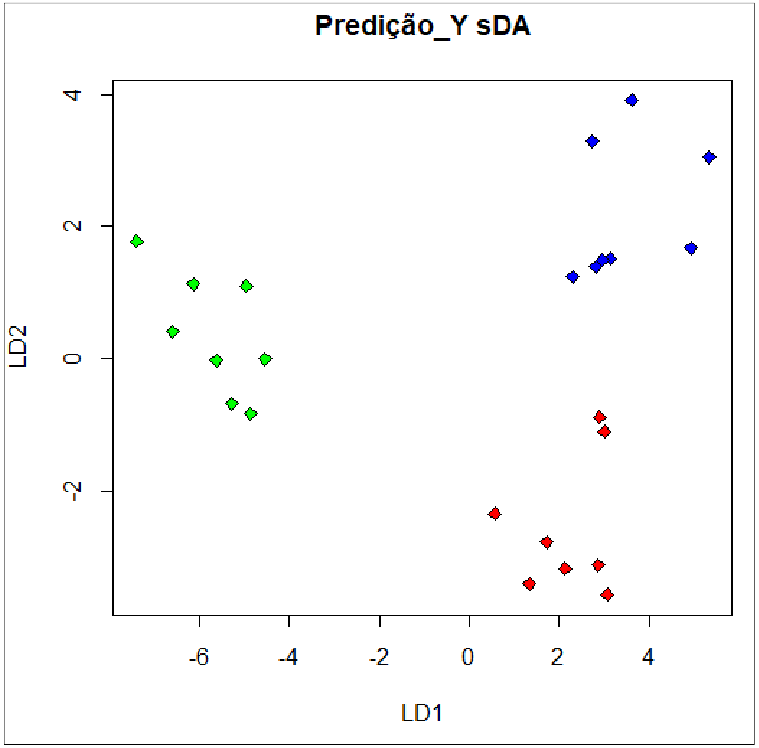
	$LD1 = Y\hat{\beta}_1$	$LD2 = Y\hat{\beta}_2$
[1,]	-3.113028	-3.4122173
[2,]	-3.142295	-3.8733571
[3,]	-2.988152	-1.4446112
[4,]	-2.995266	-1.5567002
[5,]	-2.995715	-1.5637771
[6,]	-3.063970	-2.6392373
[7,]	-2.996005	-1.5683428
[8,]	-2.998670	-1.6103476
[9,]	5.059007	0.9703178
[10,]	5.803007	0.7348022
[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 de 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

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

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

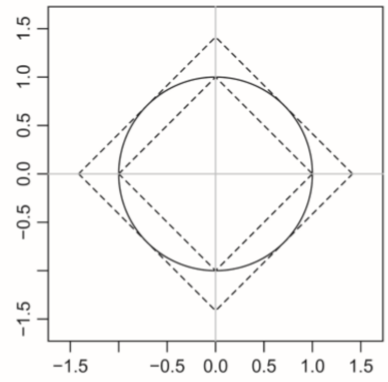
Problemas na otimização quando n<<p n<<q

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)

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

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



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

Idem para U_k

Decomposição de Matrizes - Penalização

PMD: Penalized Matrix Decomposition

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

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

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

PMD(L_1,L_1)

Diferentes algoritmos são propostos para solução deste problema de maximização (Witten et al., 2009, 2015)

Bilinear (em U e V) $\Rightarrow$ U fixo e obter V, V fixo e obter U (2 problemas de otimização linear)

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

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

Decomposição de Matrizes – Penalização Aplicação

- CCA – Esparso: PMD(L_1,L_1)

covariância

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

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

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

A solução panalizada sob a formulação de decomposição em valores singulars de matrizes também pode ser aplicada à Análise de CP:

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

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

Correlação Canônica Esparsa

- Pacote `PMA-R`, função `CCA_R`:

Dados gerados: n=100 p=500 q=1000: $\begin{bmatrix} Y_{1\,100\times 500} & Y_{2\,100\times 1000} \end{bmatrix}$

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

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

$$\rho_c = 0,9735$$

```
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
```