MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Semestre/2019

Análise Multivariada

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

Programa da disciplina ☺

- Estatísticas Descritivas Multivariadas, Elipsóides de Concentração
- Distribuição Normal Multivariada: propriedades e Distribuições Amostrais
- Regiões de Confiança, Testes Multivariados, MANOVA, IC Simultâneos, Correções para Múltiplos Testes
- Análises Multivariadas Clássicas (*n>p, iid*): CP, CoP, AC, AF, AAgr, AD, CC e PLS
- Análises Multivariadas: Soluções Esparsas (*n<<p, iid*) - CP, AD e CC
- Componentes Principais em Observações Correlacionadas
- Aprendizado de Estruturas (SEM, Propriedades de Markov, Teoria de Grafos)

Definição de Big-data?

Análise de Dados Multivariados:

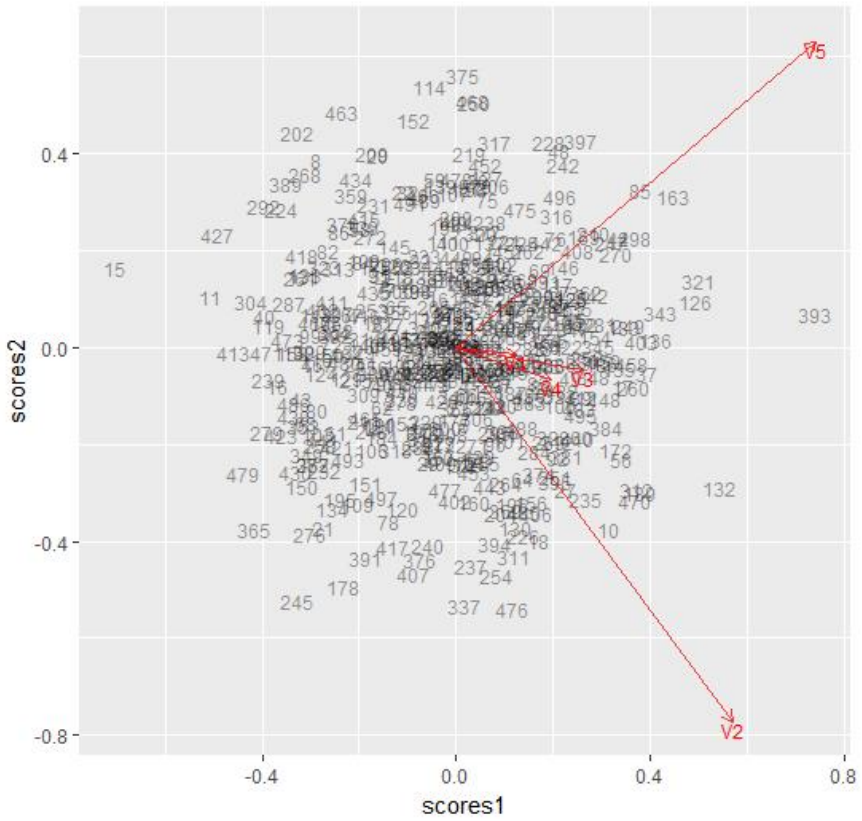
Dimensão do Espaço das Unidades Amostrais: Big-n (n >> p)?
Dimensão do espaço das Variáveis: Big-p (n << p)?

Integração de Bancos de Dados

Dados Multivariados: $Y_{n \times p}$
Problemas em Alta Dimensionalidade

Big-n (n >> p)

Biplot: n=500 p=5

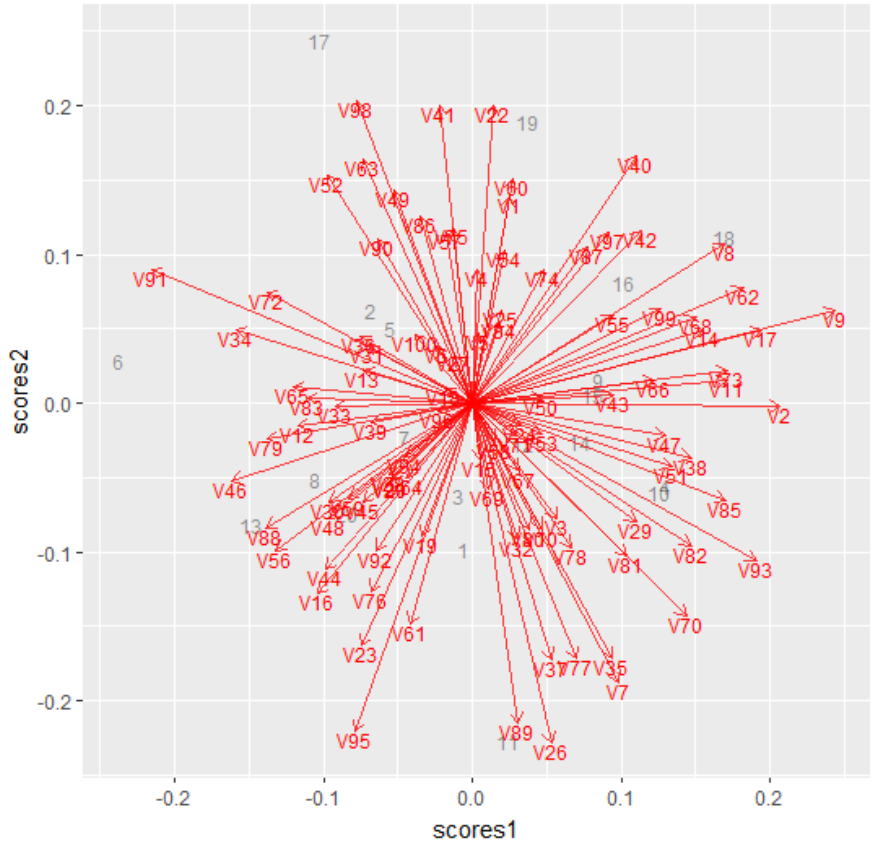


Sumarização e visualização ?

(Black Screen Problem - BSP: *R_alpha blending*)

Big-p (n << p)

Biplot: n=20 p=100



Soluções regularizadas e penalizadas!

Dados Multivariados - O que é Big-Data ?

- **Fokoué, E. (2015):** apresenta uma taxonomia para dados em alta dimensão

  → Grandes Bancos de Dados: n>1000 ou p>50

Razão n/p !

	$\frac{n}{p} < 1$ Information Poverty ($n \ll p$)	$1 \leq \frac{n}{p} < 10$ Information Scarcity	$\frac{n}{p} \geq 10$ Information Abundance ($n \gg p$)
$n > 1000$	Large p, Large n A	Smaller p, Large n B	Much smaller p, Large n C
$n \leq 1000$	Large p, Smaller n D	Smaller p, Smaller n E	Much smaller p, Small n F

Table 1: *In this taxonomy,* A *and* D *pose a lot of challenges.*

- **Matloff, N. (2016 -** Handbook of Big Data): *Big-n, Big-p*

Portnoy (1988): em Big-n, explorar propriedades assintóticas dos EMVS na família exponencial

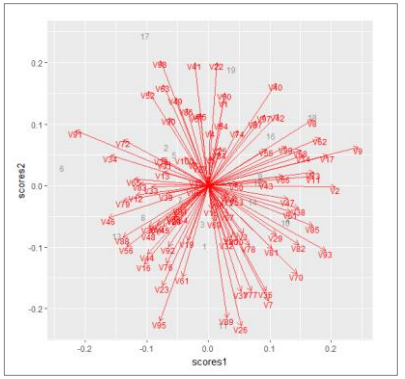
$$\hat{\theta}_{EMVS} \overset{\frac{p^2}{n} \to 0;\, p,n \to \infty}{\sim} N\left(\theta; v(\theta)\right)$$

Limite inf. de Kramer-Rao

Big-p:

$$p > \sqrt{n}$$

Dados Multivariados – *Big-p*



- *Big-p*:

- Dados esparsos: *Tamanho efetivo* de *p* é limitado

Redução de dimensionalidade: Soluções regularizadas e penalizadas

Componente Principal Regularizado e Penalizado: (Elastic Net)

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

CP da solução Dual

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

Decomposição em valores singulares

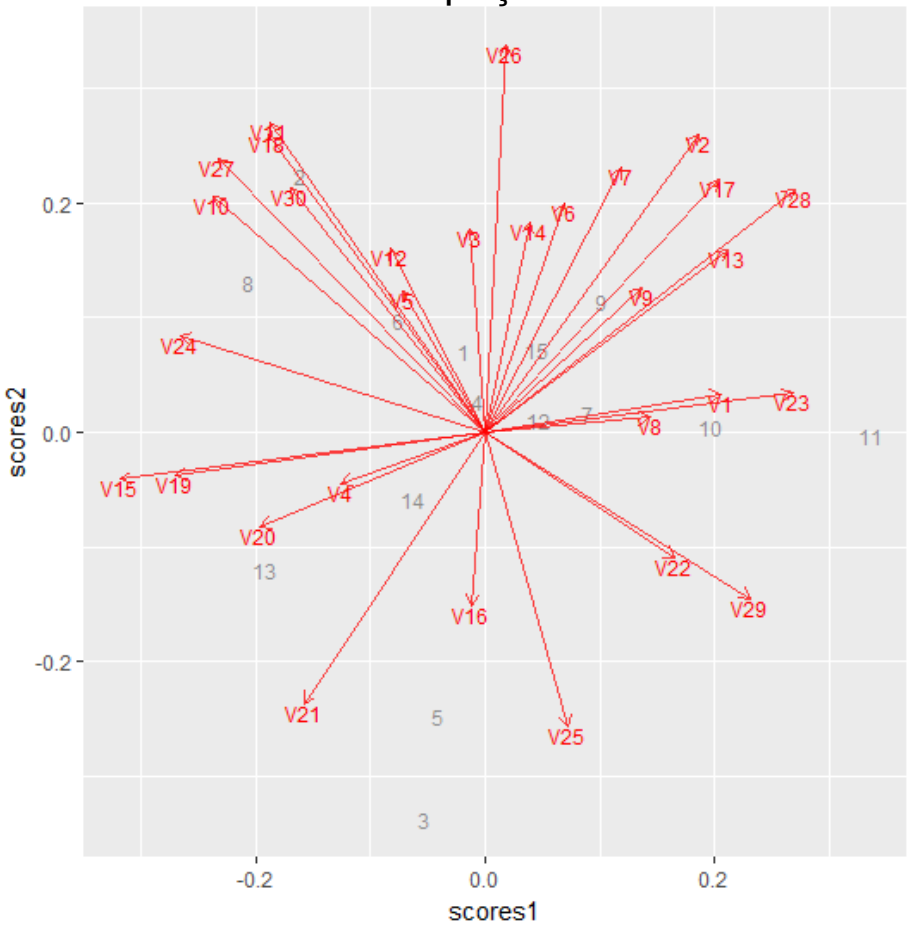
$$\max_{U_k, V_k} U_k'\, Y\, V_k; \quad \begin{cases} \left\| U_k \right\|_2^2 \le 1 \\ \left\| V_k \right\|_2^2 \le 1, \ \left\| V_k \right\|_1 \le c_1 \end{cases}$$

Algoritmo de maximização (Witten et al., 2009)

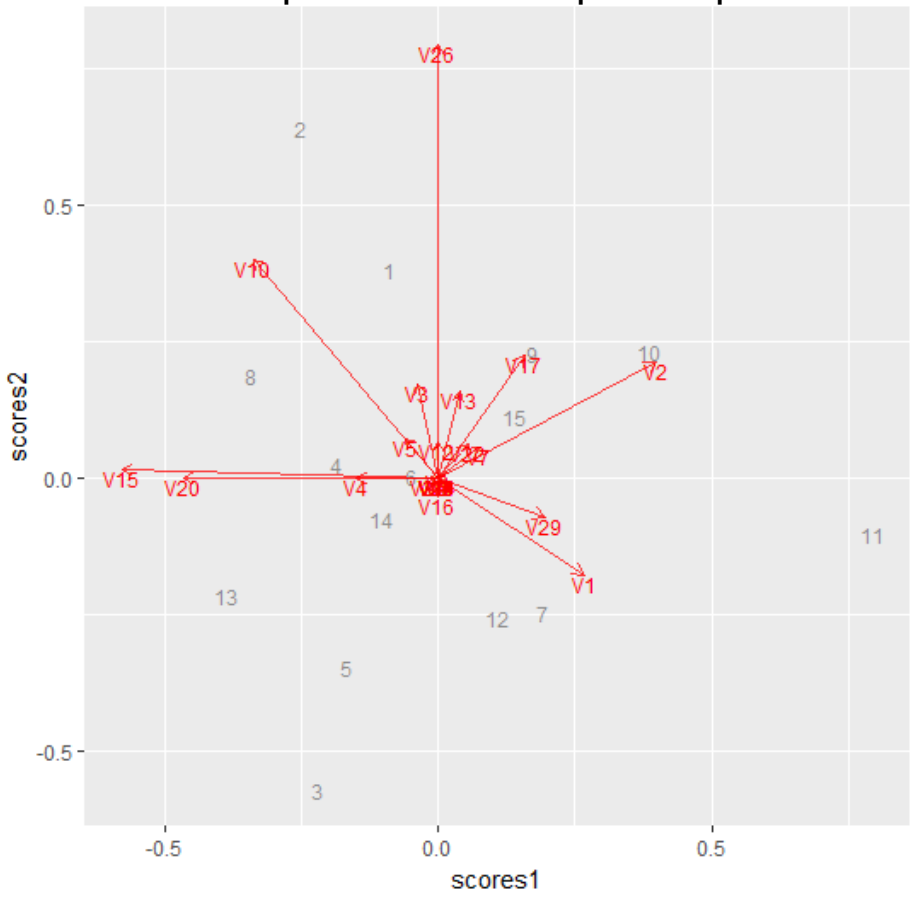
Componentes Principais – n<<p

Representação Biplot: n=15 p=30

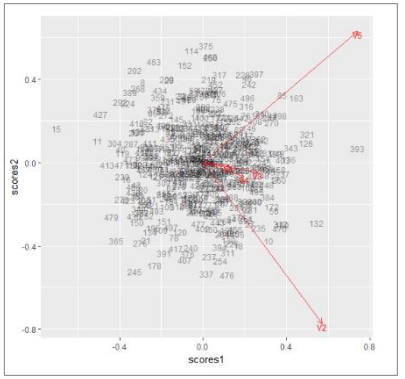
R-prcomp: Componentes Principais em Espaços Duais



R-SPCA do pacote ElasticNet: Componentes Principais Esparsos



Dados Multivariados – *Big-n*



Big-n $\Rightarrow$ erro padrão das estimativas tendem a zero

$$\frac{s}{\sqrt{n}} \to 0 \quad \not\in$$ problema inferencial, somente análise descritiva de dados ??

Não é um consenso!

Muitas análises são realizadas no espaço $\Re^{nxn}$: tempo computacional é o problema

Soluções Paralelizadas (N. Matloff, 2016):

- Particionar os dados via procedimentos de Aleatorização
- Em cada sub-amostra calcular a estimativa de interesse ($\hat{\theta}_g$)
- Obter a média das estimativas

Pacote R: "partools"
Paralelização de cálculos estatísticos em observações *iid*
(SA: Software Alchemy)

$$n = \sum_{g=1}^{G} n_g; \quad n_g = \frac{n}{G} \qquad \Rightarrow \bar{\theta}_n = \frac{1}{G}\sum_{g=1}^{G} \hat{\theta}_g$$

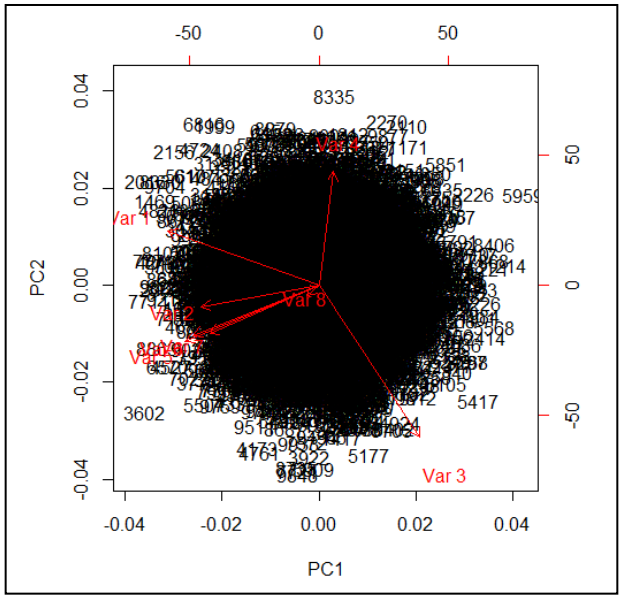
$$\Rightarrow Cov\left(\bar{\theta}_n\right) = \frac{1}{G} Cov\left(\hat{\theta}_g\right)$$

$$\bar{\theta}_n \xrightarrow[n\to\infty, n_g\to\infty]{G \text{ fixo}} \hat{\theta}_{EMVS}$$

Dados Multivariados – *Big-n*

```
> require(partools)
> cls <- makeCluster(2)
> setclsinfo(cls)
> #Gerar dados para ajustar Modelo Linear
> n <- 10000
> p <- 2
> set.seed = 1099
> tmp <- matrix(rnorm((p+1)*n),nrow=n)
> u <- tmp[,1:p] # gerar valores "X"
> # adicionar coluna de valores "Y"
> u <- cbind(u,u %*% rep(1,p) + tmp[,p+1])
> colnames(u) = c("X1","X2","Y")
> #head(u)
> # ajustar lm via solução paralelizada (N. Matloff, 2015)
> # SA: Software Alchemy
> distribsplit(cls,"u")
> #calm(cls,"u[,3] ~ u[,1]+u[,2]")
> calm(cls,"u[,3] ~ u[,1]+u[,2]")$tht
 (Intercept)        u[, 1]        u[, 2]
-0.003110128  1.005448362  1.002561369
>
> # check: resultados devem ser aproximadamente os mesmos
> lm(u[,3] ~ u[,1]+u[,2])
(Intercept)        u[, 1]        u[, 2]
  -0.002909     1.005829     1.002436
```

Dados Multivariados – *Big-n*

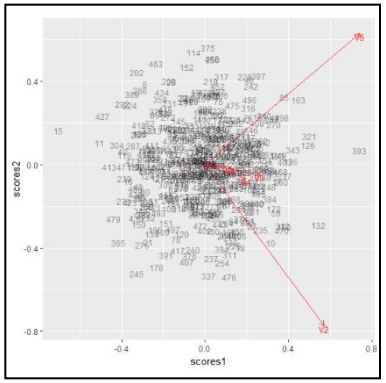



```
> require(partools)
> cls <- makeCluster(2)
> setclsinfo(cls)

#gerar dados multivariados
> n = 10000
> p = 8
> n/p
[1] 1250
> media_pop = c(rdunif(p, 10))
> cov_pop = round(genPositiveDefMat(dim=p)$Sigma,2)
> dados = mvrnorm(n,media_pop,cov_pop)
> biplot(prcomp(dados,scale=TRUE))

#rodar análise de CP via solução paralelizada (N. Matloff, 2015)
#SA: Sottware Alchemy
> distribsplit(cls,"dados")
> caprcomp(cls,'dados,scale=TRUE',8)$sdev
[1] 1.4707854 1.0525281 1.0046278 0.9884619 0.9051253 0.8954517 0.8450872
[8] 0.6382045
> prcomp(dados,scale=TRUE)$sdev
[1] 1.4707662 1.0516048 1.0024852 0.9883842 0.9063294 0.8971400 0.8458021
[8] 0.6382948
```

Dados Multivariados – *Big-n*



Visualização Gráfica: (Norman Matloff)

- Construção do Gráfico de Coordenadas Paralelas
- Controle do BSP
- *Big-n*: representar os pontos (perfis) com os maiores valores de densidade

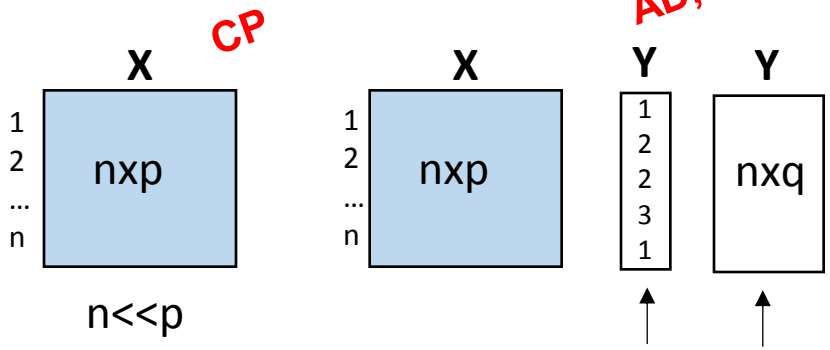
  Estimação da densidade (método dos *k*-vizinhos mais próximos)

N-Integração de Bancos de Dados

Um Único BD

Análise Não-Supervisionada CP

Análise Supervisionada AD, CCg



Integração entre múltiplos níveis de informação avaliados nas mesmas unidades amostrais!

⇒ **Aplicação em Dados MultiOmics**

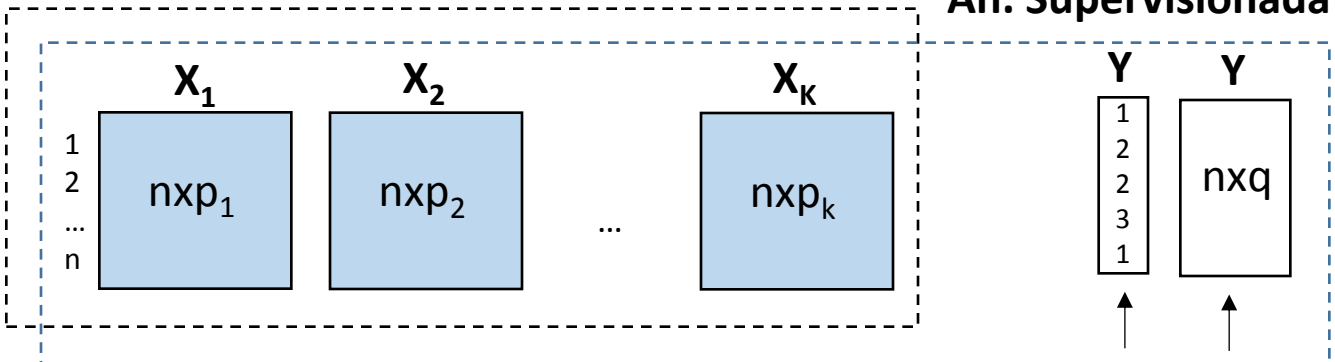
X: Matriz(es) de Dados

Y: Respostas de interesse (Classes de doenças ou Variáveis quantitativas)

Múltiplos BD (Multimodais, Multivisão)

Análise Não-Supervisionada CCg

An. Supervisionada



P-Integração: mesmas variáveis em diferentes unidades amostrais (estudos multicentros, Meta-análise)

Redução de Dimensionalidade e Integração de BD

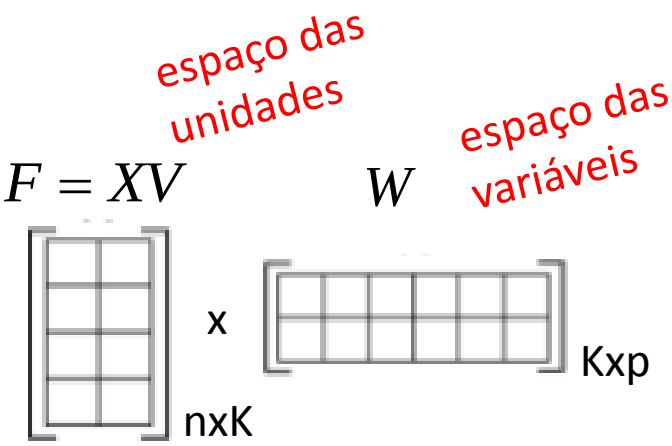
$X_{n\times p}$: Matriz de dados

Soluções via Fatoração do espaço $\Re^{n \times p}$

$$X_{n\times p} = \underbrace{U_{n\times n} D_r^{1/2}}_{=F} \underbrace{V'_{p\times p}}_{=W}$$

Matriz de Escores
(valores latentes)

Matriz de Cargas
(direções de projeção)



Análise de Um Único BD - Fatoração de Matrizes

$X_{n\times p}$: Matriz de dados (padronizados)

Análises de Um Único BD (UniOmics: Lê Cao - mixOmics-R)

Solução via a matriz de cargas

- Componentes Principais Independentes

IPCA (PCA + ICA): PCA com filtros ICA na Matriz de Cargas

$$X_{n\times p} = \underbrace{U_{n\times n}\, D_r^{1/2}}_{F = XV}\; V'_{p\times p} \quad \Rightarrow \quad X\tilde{V}$$

$V \rightarrow \tilde{V}$

Capturar sinais (que não são erros Gaussianos) dos Componentes Principais

Aplicar ICA em V ("denoising" V)
Algoritmo FastICA

ICA: Análise de Componentes Independentes
(Comon, 1994)

Integração de Múltiplos Bancos de Dados

Solução PLS (Partial Least Square):

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

Solução CCA (Canonial Correlation Analysis):

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

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

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

Integração de Bancos de Dados

Regressão PLS , CCA e DA

Solução PLS

Solução CCA

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

Bartlet (1938): propõe **CCA-DA** $\Rightarrow$ CCA pode integrar X (matriz quantitativa) e Y (matriz de variáveis indicadoras da Doença (var. dummy)

Barker and Raynes (2003): propõem **PLS-DA**

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

Soluções mais gerais:

Lê Cao (2018): propõem Correlação Canônica Generalizada (**GCCA**)

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

J bancos de dados

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

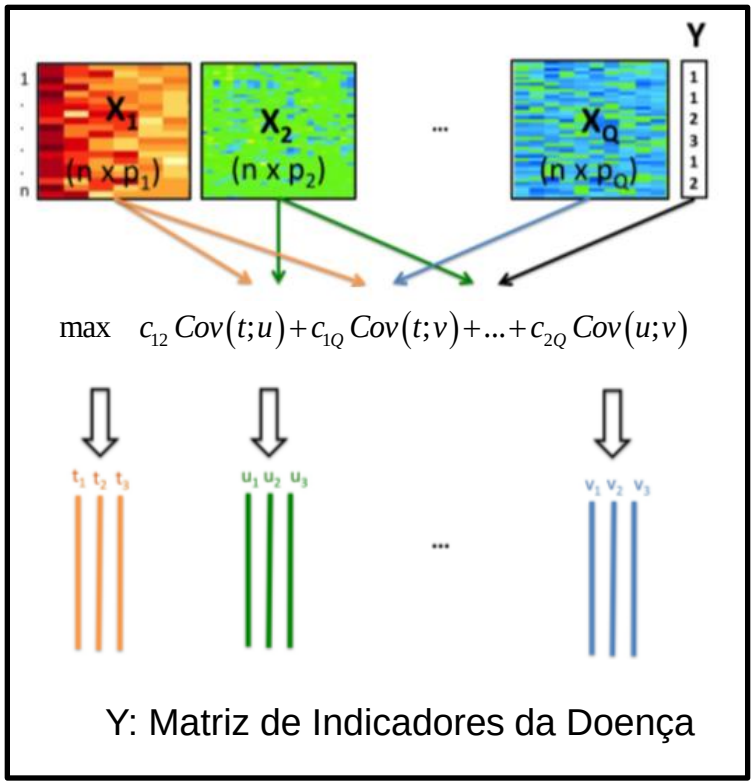
$C = \left(c_{jk}\right)$: atribui pesos às conexões entre BD (Teoria de Grafos)

Integração de Bancos de Dados
Correlação Canônica Generalizada

Lê Cao (2018): GCCA

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

J bancos de dados



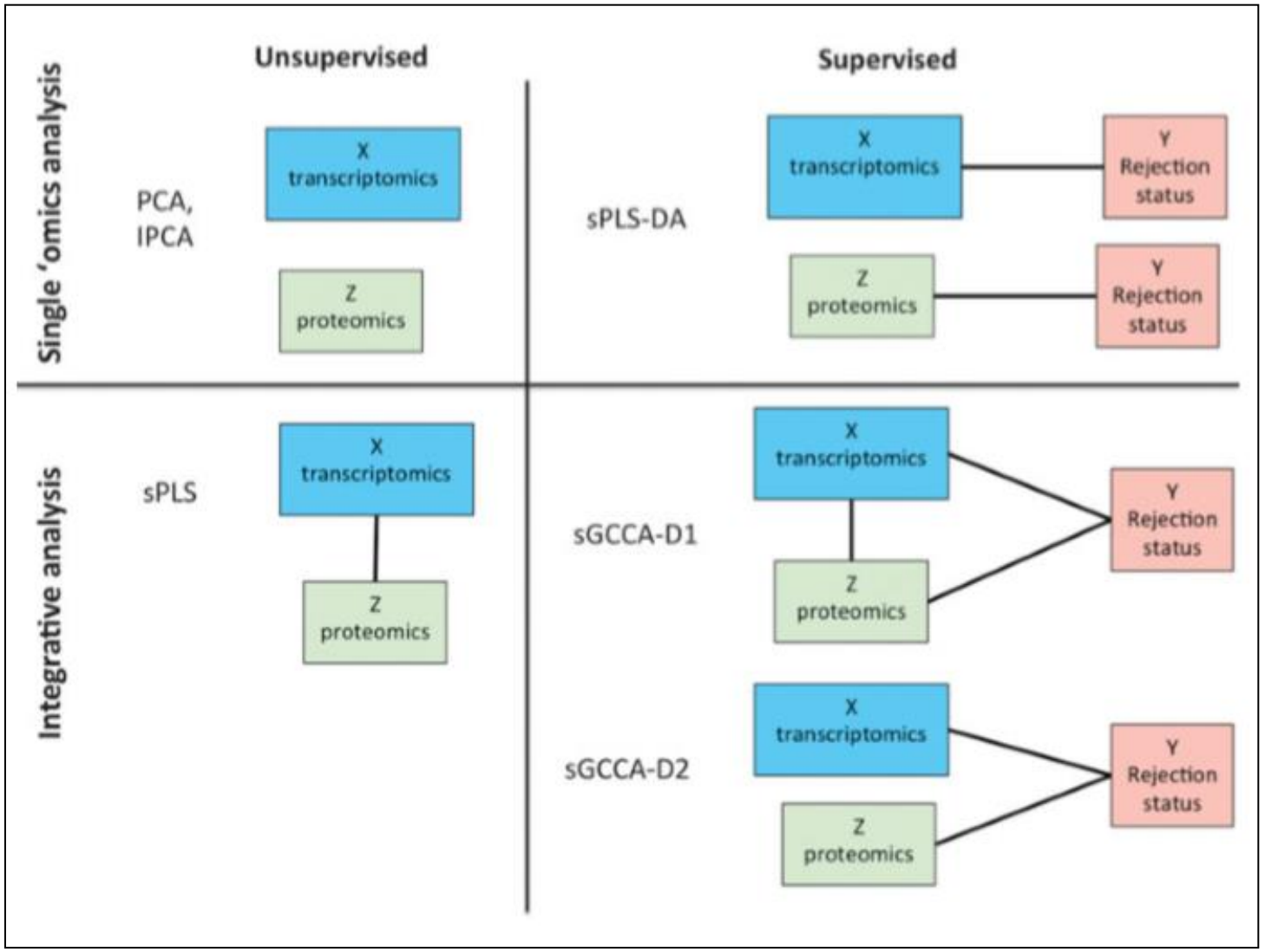
Y: Matriz de Indicadores da Doença

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

$C = (c_{jk})$: atribui pesos às conexões entre BD (Teoria de Grafos)

Integração de Banco de Dados – MixOmics (R)

Recursos



Integração de Banco de Dados – MixOmics (R)

(Lê Cao et al., 2018)

Estudo dos Marcadores Moleculares na Resposta de Rejeição ao Transplante Renal (University of British Columbia, Canada)

Estudo Prospectivo (Dados Pareados):

n=20 Pacientes AR: rejeição antes de 6 meses após transplante

n=20 Pacientes NR: sem resposta de rejeição até 6 meses

Coleta de Sangue no momento do diagnóstico de rejeição:

≤ 5 dias: Grupo AR-early (n=15)

> 5 dias: Grupo AR-late (n=5)

AR-early AR-late NR-early NR-late

X1: expressão de 54.613 transcripts

X2: quantificação de 444 proteínas

Y: matriz de especificação dos 4 grupos

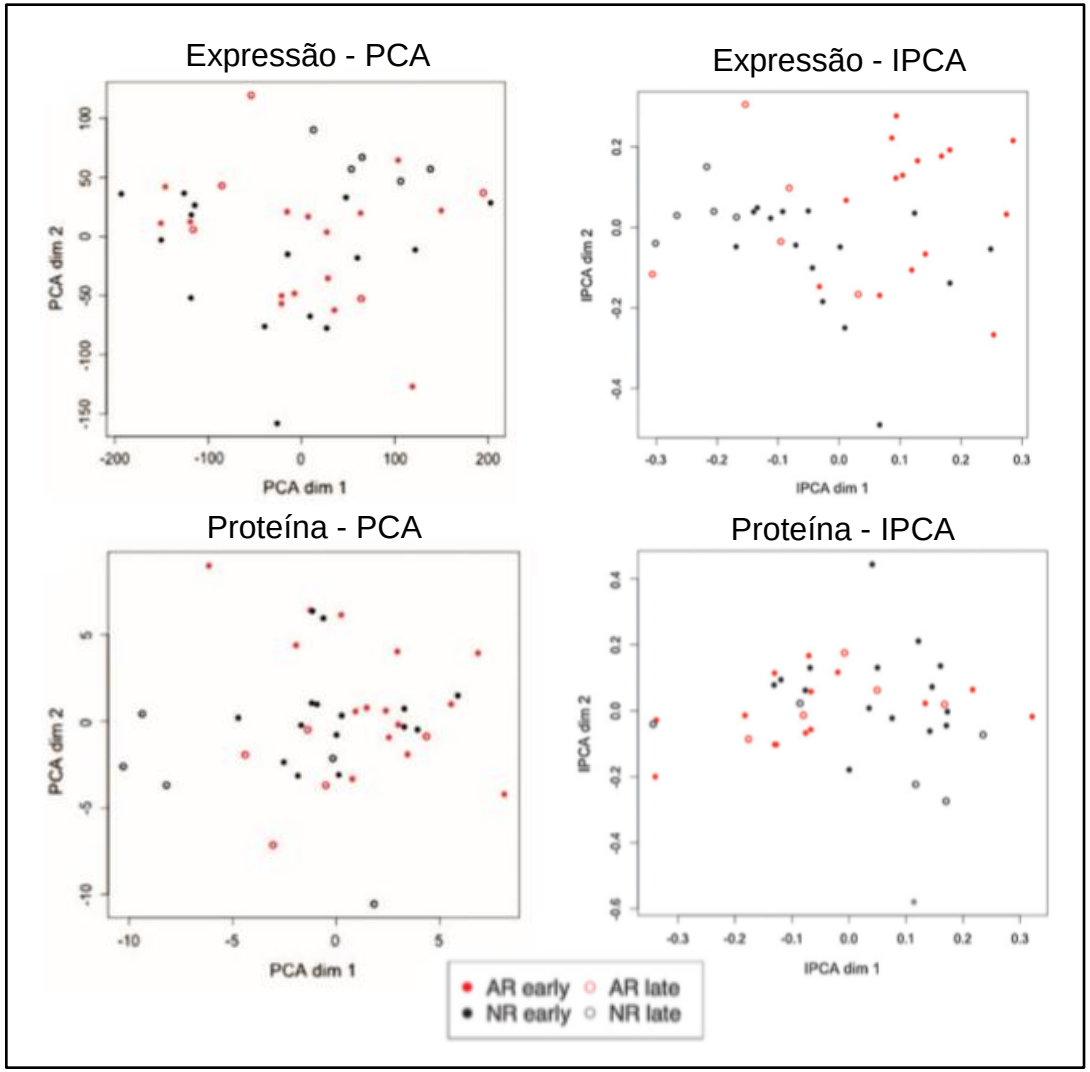
Integração supervisionada de dois bancos de dados,

$X1_{40\times54.613}$ $X2_{40\times444}$

para “aprender” sobre $Y_{40\times4}$

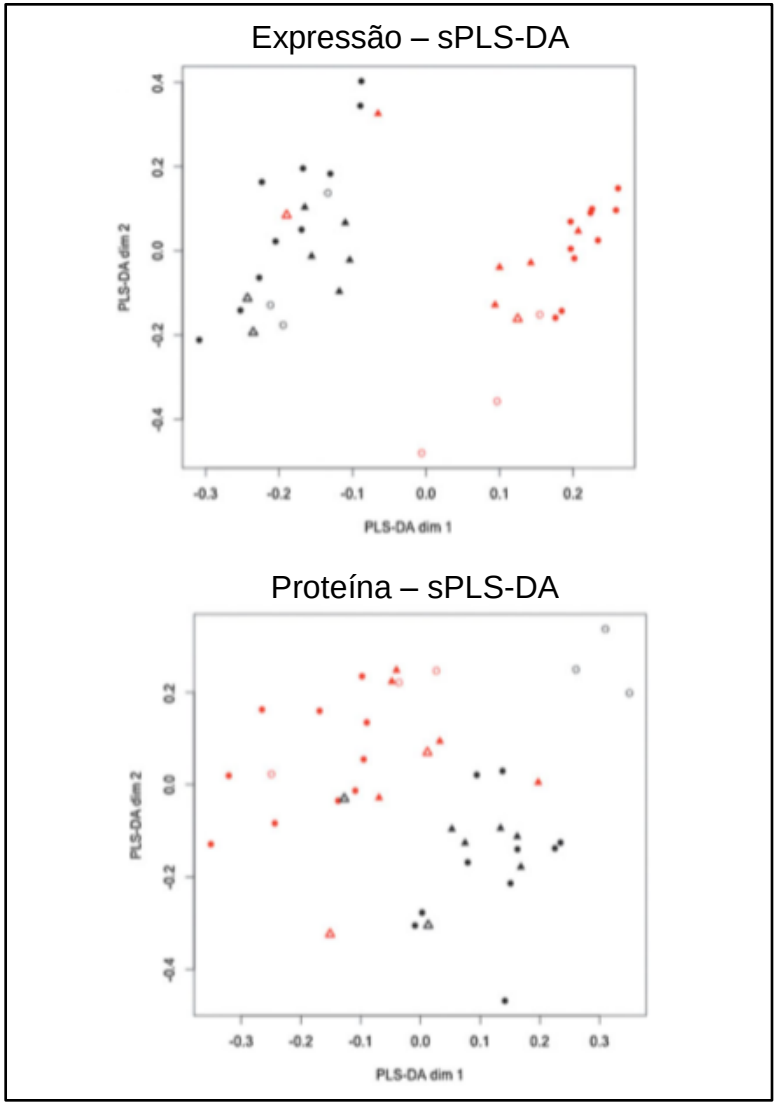
Análise UniOmics Não-Supervisionada

Escores de $X1_{40\times54.613}$ e de $X2_{40\times444}$



Análise Uni-Omics Supervisionada

Escores de $X1_{40\times54.613}$ e de $X2_{40\times444}$ para aprender $Y_{40\times4}$

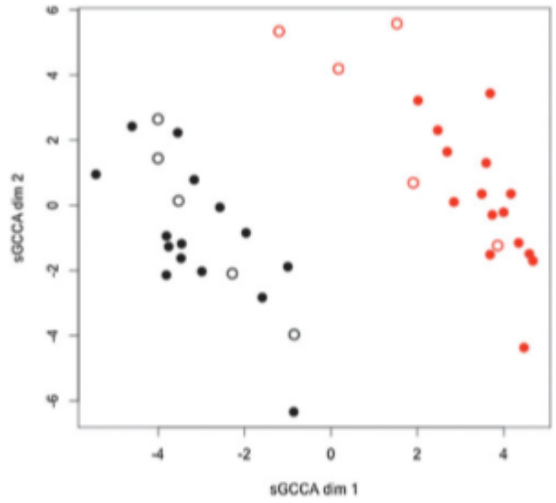


Análises UniOmics: Dados de Expressão com melhor separação do que os dados de proteína

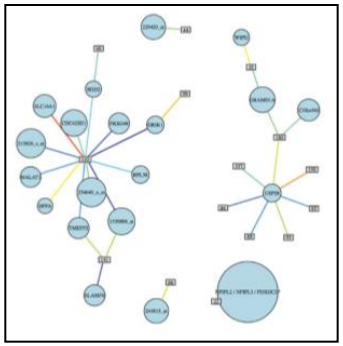
Análise MultiOmics Supervisionada: GCCA

(Lê Cao et al., 2018)

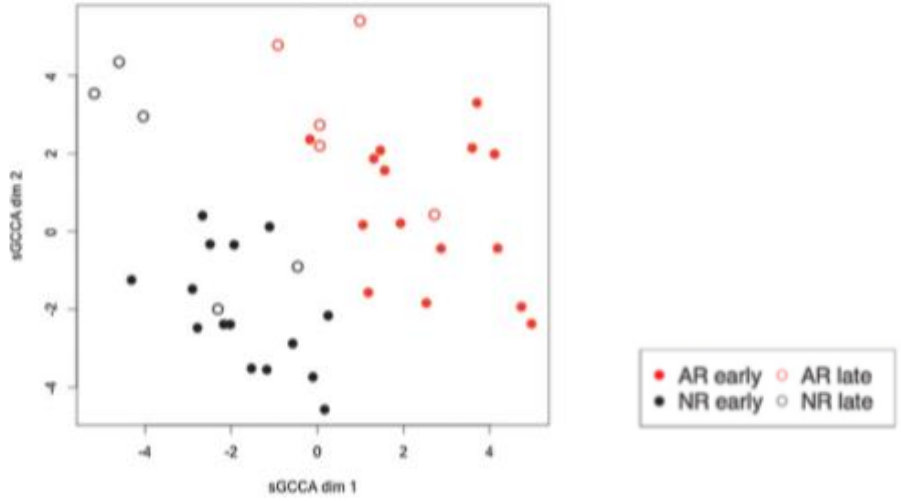
Expressão – sGCCA



Grafo: estabele conexões entre os BD



Proteína – sGCCA

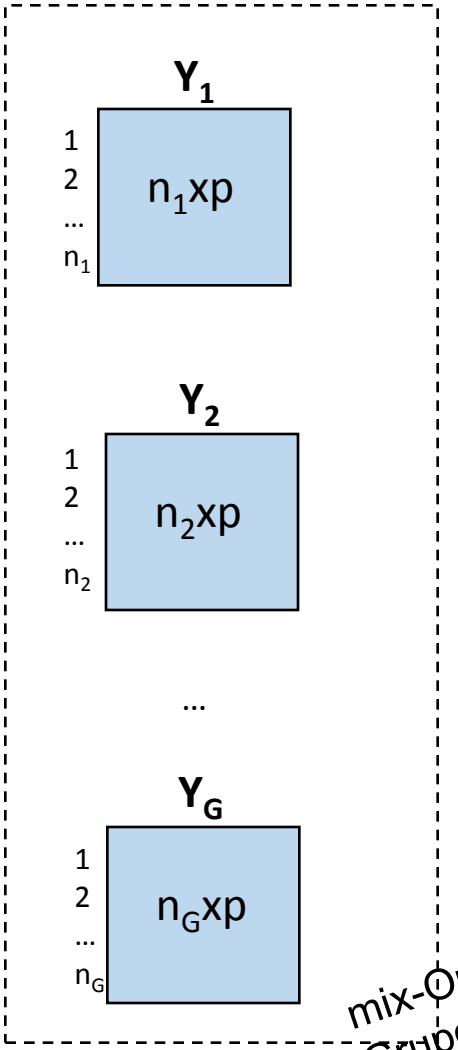


A integração dos BD permitiu melhor separação dos grupos do que a análise UniOmics

Dados de Expressão discriminam melhor que os dados de proteína

P-Integração de Bancos de Dados

ASCA: ANOVA-Simultaneous Component Analysis (Smilde et al., 2005)



$$Y_{n\times p};\quad n=\sum_{g=1}^{G} n_g$$ Matriz de dados

$$Y_{n\times p} = Y_{Médias\ n\times p} + Y_{Grupo\ n\times p} + Y_{Residual\ n\times p}$$

Redução de dimensionalidade em componentes da decomposição de Y

Obter variáveis latentes ("componentes principais") que atendam a estrutura de Y (estratificada)

mix-Omics_R
Grupo de pesquisa da Lê Cao

Redução de Dimensionalidade e Integração de BD
Decomposição de Matrizes

$Y_{n\times p}$; $n = \sum_{g=1}^{G} n_g$: Matriz de dados

Soluções via Decomposição do Espaço do espaço $\Re^{n \times p}$

Tabela de MANOVA: Decomposição em $\Re^{p\times p}$

T = H + E

T/(n-1)=S: Matriz de covariância (total)

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

Modelo de MANOVA

Decomposição em $\Re^{n\times p}$

$$Y_{n\times p} = Y_{Média} + Y_{Grupo} + Y_{Resíduo}$$

Realizar as análises (de redução de dimensionalidade ou integração de BD) em components de Y

Obter CP de T, H, E ou $E^{-1}H$

Redução de Dimensionalidade e Integração de BD via Decomposição de Matrizes

$Y_{n\times p};\ n=\sum_{g=1}^{G} n_g$: Matriz de dados $\qquad Y_{ig\ p\times 1} = \bar{Y} + (\bar{Y}_g - \bar{Y}) + (Y_{ig} - \bar{Y}_g)$

$$Y_{n\times p} = Y_{M\ n\times p} + Y_{Gr\ n\times p} + Y_{E\ n\times p}$$

Análises nos Componentes de Y
- Matriz rectangular (nxp)

Análises nos Componentes da matriz S=T/(n-1)
- Matrizes quadradas (pxp)

Fonte de Variação	Número de g.l.	Matriz de SQPC
Grupo	**G-1**	$H_{p\times p} = \sum_{g=1}^{G} r\left(\bar{Y}_g - \bar{Y}\right)\left(\bar{Y}_g - \bar{Y}\right)'$
Resíduo	**G(r-1)**	$E_{p\times p} = \sum_{g=1}^{G}\sum_{i=1}^{r}\left(Y_{ig} - \bar{Y}_g\right)\left(Y_{ig} - \bar{Y}_g\right)'$
Total	**Gr-1**	$T_{p\times p} = \sum_{g=1}^{G}\sum_{i=1}^{r}\left(Y_{ig} - \bar{Y}\right)\left(Y_{ig} - \bar{Y}\right)' = H + E$

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)					

$$Y_{8\times 2} = M_{8\times 2} + T_{8\times 2} + E_{8\times 2}$$

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

Dependendo do problema pode haver interesse no componente T ou no componente E (residual ou resposta normalizada)

Pense no fator T como ***Multicentros*** e no problema de ***p-Integração*** de bancos de dados!

MAE5776 – Análise Multivariada de Dados
1º Semestre/2019 IME/USP

Considero que
cumprimos o seguinte
Programa da disciplina:

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

Agradeço ☺

- Estatísticas Descritivas Multivariadas, Elipsóides de Concentração
- Distribuição Normal Multivariada: propriedades e Distribuições Amostrais
- Regiões de Confiança, Testes Multivariados, MANOVA, IC Simultâneos, Correções para Múltiplos Testes
- Análises Multivariadas Clássicas (*n>p, iid*): CP, CoP, AC, AF, AAgr, AD, CC e PLS
- Análises Multivariadas: Soluções Esparsas (*n<<p, iid*) - CP, AD e CC
- Componentes Principais em Observações Correlacionadas
- Aprendizado de Estruturas (SEM, Propriedades de Markov, Teoria de Grafos)
- Problemas em Alta dimensão: Big-p (n << p) e Big-n (n >> p)
- Integração de Bancos de Dados

E ainda vamos ouvir cada um nos
seguintes temas complementares →

MAE5776 – Análise Multivariada de Dados
1º Semestre/2019 IME/USP

Temas de Seminários:

- Cleber: Componentes Principais em Dados Composicionais
- Francisco: Componentes Principais em Dados Intervalares
- Rafael: Componentes Principais Independentes
- Jonatas: Validação Cruzada em Modelos Mistos
- Felipe: BiCluster
- Ana Gabriela: Integração de BD Heterogêneos
- Luiza: Visualização em Big-n: Gráfico de Coordenadas Paralelas
- Ana Carolina: Sumarização de um BD em Alta Dimensão (Big-p)
- Piero: Redução de Dimensionalidade em Séries Temporais
- Alia: Algoritmo: Decaimento de Coordenadas
- Désirée: Modelo de Equações Estruturais (Confirmatório, Exploratório)

Apresentação Oral do Tema (1/2 h)
Entrega de um Relatório: Título,
Caracterização dos Dados e do Problema,
Uma Solução, Ilustração da Análise,
Referência Bibliográfica

E ainda vamos revisar o conteúdo básico da disciplina →

MAE5776 – Análise Multivariada de Dados
1º Semestre/2019 IME/USP

Conteúdo a ser revisado:

- Regiões de Concentração de Dados, Regiões de Confiança, Intervalos de Confiança Simultâneos
- Solução Dual para a Redução de Dimensionalidade, Biplots
- Conexão entre PCA e o Método de Regressão
- Correlação Canônica e Integração de Bancos de Dados
- Aprendizado via Análise Discriminante, Validação Cruzada e Classificação
- Soluções Clássicas e para dados em alta dimensão (big-p)

Boa sorte ☺

N-Integração de Bancos de Dados