MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

IME – 1o. Sem/2022

Análise Multivariada

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

Já vimos ☺

- ✓ Estatísticas Descritivas Multivariadas
- ✓ Distribuição Normal Multivariada, Distribuições Amostrais
- ✓ Regiões de Confiança, Testes Multivariados, MANOVA, IC Simultâneos, Correções para Múltiplos Testes
- ✓ Análises Multivariadas Clássicas (*n>p, iid,* $\Re^{nxp}$): CP, CoP, AC, AF, AD, ACC

Análises Multivariadas em Dados de Alta Dimensão (Big-Data)

Razão n/p?

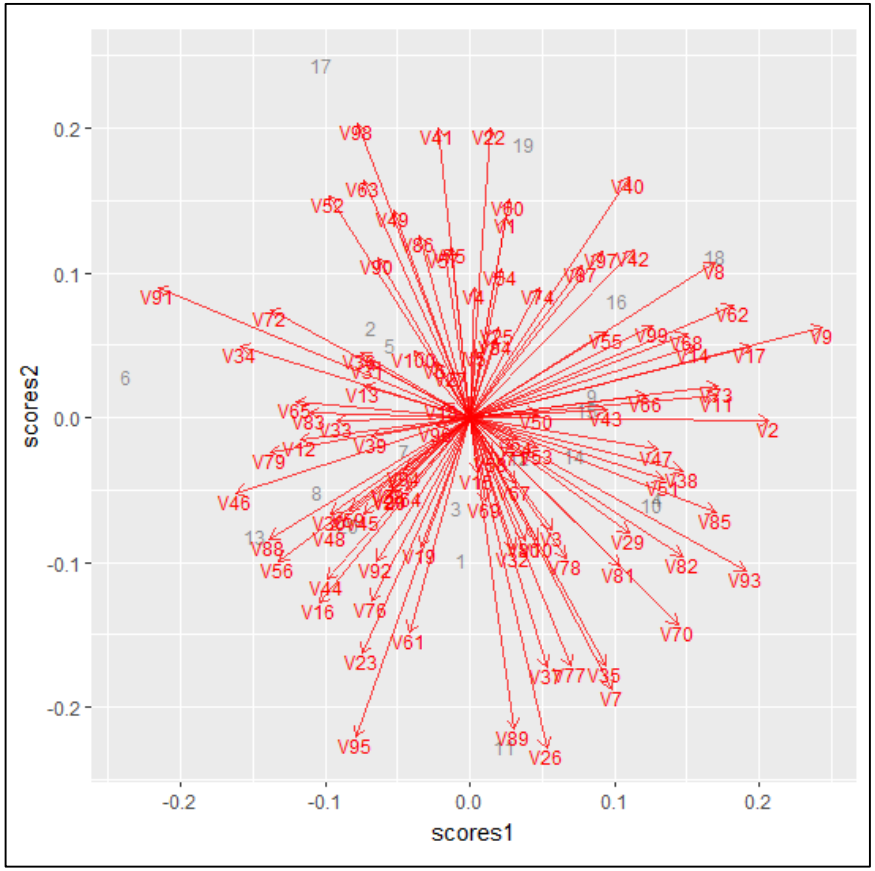
- ✓ Análises Multivariadas em Big-p (*n<<p*): Soluções duais e Soluções regularizadas e esparsas (CP, AD, ACC)

⇒ Análises Multivariadas em Big-n (n >> p)

Dados Multivariados: $Y_{n\times p}$
Dimensionalidade dos Dados

Big-p (n << p) "wide data"

Biplot: n=20 p=100

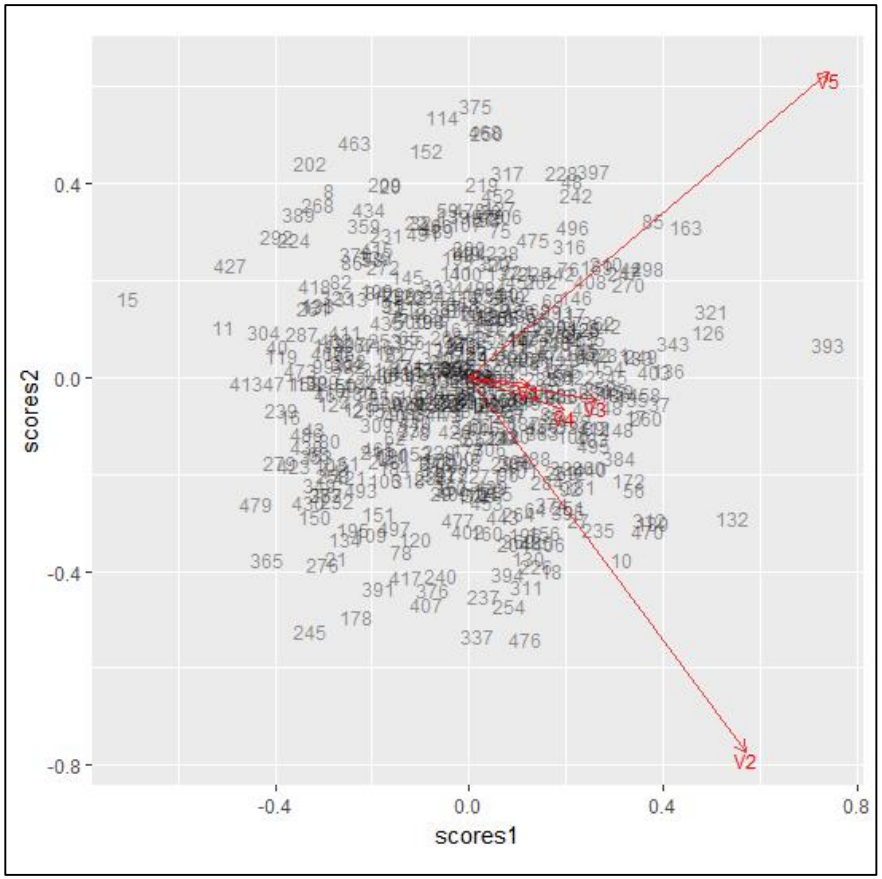


Soluções em espaços duais

Soluções regularizadas e/ou penalizadas

Big-n (n >> p) "long data"

Biplot: n=500 p=5



Sumarização e visualização ?

(Black Screen Problem: *R_alpha blending*)

Não confundir com outras definições!

Estrutura de Dados

Dados: Medidas Repetidas no formato "wide"

Alternativa de análise: MANOVA

```
> library(reshape2)
> dat.wide <- dcast(dat.long, Subj +
Grup + Staph~ Time, value.var="O2")
> dat.wide
```

	Subj	Grup	Staph	6	12	18
1	1	P	1	1.48	2.81	3.56
2	2	P	0	1.04	2.07	2.81
3	3	P	1	1.48	2.52	3.41
4	4	P	0	1.04	1.93	2.89
…						
21	21	P	1	1.50	2.85	3.12
22	22	P	0	1.65	2.70	3.40
23	23	P	1	1.80	2.15	3.90
24	24	P	0	1.20	2.25	3.30
25	25	V	1	1.78	2.96	4.00
26	26	V	0	1.48	2.81	3.85
27	27	V	1	1.33	2.52	3.84
28	28	V	0	1.03	2.07	2.96
…						
45	45	V	1	1.65	3.00	4.05
46	46	V	0	1.20	2.70	3.90
47	47	V	1	1.35	2.55	3.67
48	48	V	0	1.20	2.70	3.60

Dados: Medidas Repetidas no formato "long"

Alternativa de análise: Modelos lineares mistos

```
> library(MANOVA.RM)
> dat.long<-o2cons
> dat.long
> dat.long
```

	O2	Staph	Time	Grup	Subj
1	1.48	1	6	P	1
2	2.81	1	12	P	1
3	3.56	1	18	P	1
4	1.04	0	6	P	2
5	2.07	0	12	P	2
6	2.81	0	18	P	2
…					
67	1.80	1	6	P	23
68	2.15	1	12	P	23
69	3.90	1	18	P	23
70	1.20	0	6	P	24
71	2.25	0	12	P	24
72	3.30	0	18	P	24
73	1.78	1	6	V	25
74	2.96	1	12	V	25
75	4.00	1	18	V	25
76	1.48	0	6	V	26
77	2.81	0	12	V	26
78	3.85	0	18	V	26
…					
139	1.35	1	6	V	47
140	2.55	1	12	V	47
141	3.67	1	18	V	47
142	1.20	0	6	V	48
143	2.70	0	12	V	48
144	3.60	0	18	V	48

Dados Multivariados - O que é Big-Data ?

- Hair et al., 2005, 2018: Tamanho amostral $\Rightarrow$ n>100, n > 5p
- Grau de correlação entre as variáveis $\Rightarrow$ |r| > 30%

Análises Clássicas

- **Fokoué, E. (2015):** apresenta uma taxonomia para dados de "Alta Dimensão"

  $\rightarrow$ Grandes Bancos de Dados: n>1000 ou p>50

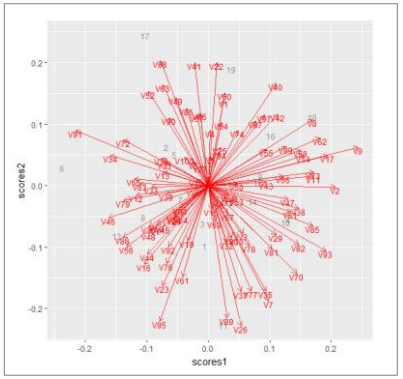
Razão n/p !

	$\frac{n}{p} < 1$ Information Poverty ($n \lll p$)	$1 \leq \frac{n}{p} < 10$ Information Scarcity	$\frac{n}{p} \geq 10$ Information Abundance ($n \ggg p$)
$n > 1000$	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

In this taxonomy, A and D pose a lot of challenges.

Já vimos!

Dados Multivariados – *Big-p*



Redução de dimensionalidade: Soluções duais, soluções regularizadas e penalizadas

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

Componente Principal Regularizado e Penalizado: (Elastic Net)

(Zou, Hastie and Tibshirani, 2006)

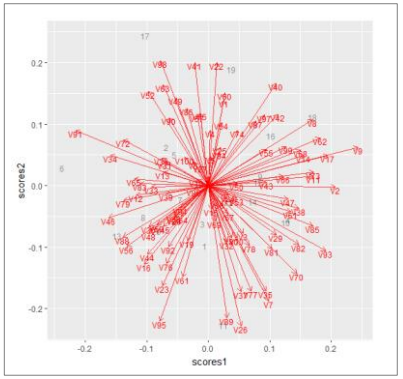
CoP (solução dual)

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

$$\arg\min_{\alpha,\beta} \sum_{i=1}^{n} \left\| Y_{i_{p\times 1}} - \alpha_{p\times 1}\beta'_{1\times p}\, Y_i \right\|_2^2 + \lambda \left\| \beta \right\|_2^2; \quad \left\| \alpha \right\|_2^2 = 1 \quad \Rightarrow \hat{\beta} \propto \hat{v}_1$$

Já vimos!

Dados Multivariados – *Big-p*



Redução de dimensionalidade: Soluções duais, soluções regularizadas e penalizadas

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

Decomposição em valores singulares de matrizes

(Witten et al., 2009; Tibshirani et al., 2015)

$$\max_{U_k,V_k} U_k' Y_{n\times p} V_k \; ; \quad \begin{cases} \|U_k\|_2^2 \le 1, \|U_k\|_1 \le c_1 \\ \|V_k\|_2^2 \le 1, \|V_k\|_1 \le c_2 \end{cases} \qquad \text{PMD}(L_1,L_1)$$

⇒ **CP:** $\max_{V_k} V_k' (Y'Y)_{p\times p} V_k ; \quad \|V_k\|_2^2 \le 1, \; \|V_k\|_1 \le c_2$ PMD(· , L_1)

⇒ CoP: $\max_{U_k} U_k' (YY')_{n\times n} U_k ; \quad \|U_k\|_2^2 \le 1, \; \|U_k\|_1 \le c_1$ PMD(L_1 , ·)

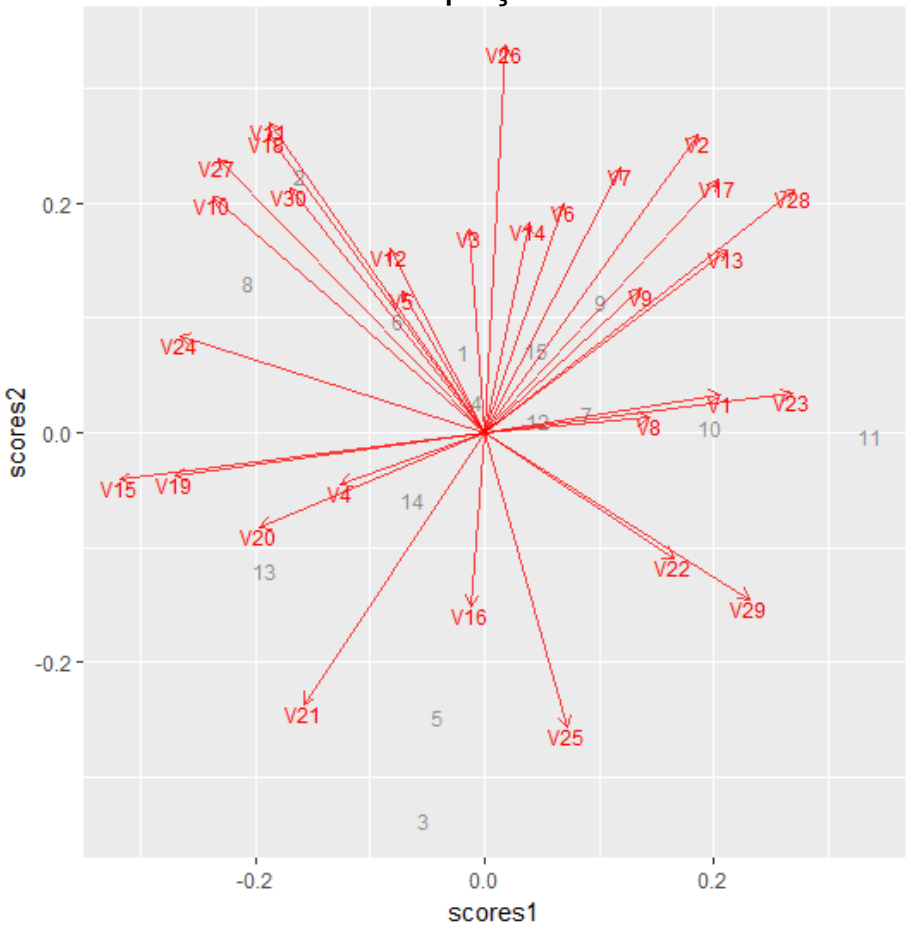
⇒ **ACC:** $\max_{a_k,b_k} a_k' \left(Y_1'Y_2\right)_{p\times q} b_k \; ; \quad \begin{cases} a_k'Y_1'Y_1a_k \le 1, \; \|a_k\|_1 \le c_1 \\ b_k'Y_2'Y_2b_k \le 1, \; \|b_k\|_1 \le c_2 \end{cases}$ PMD(L_1,L_1)

Já vimos!

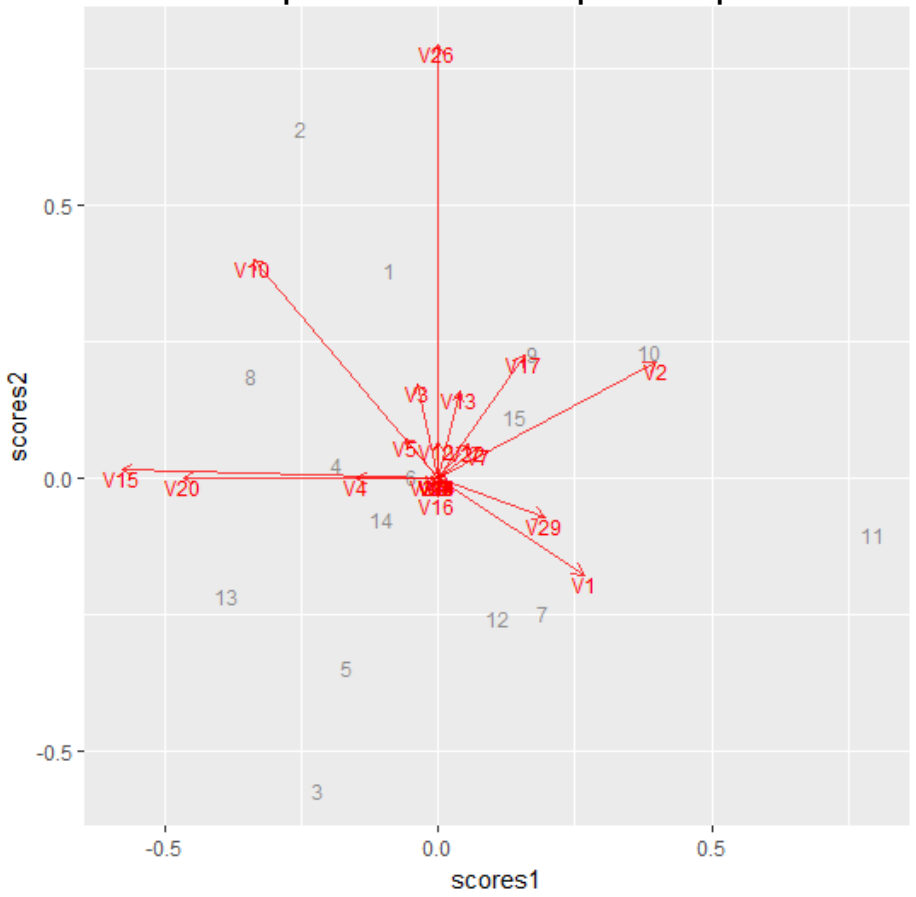
Componentes Principais – n<<p

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 dos estimadores tendem a zero

$\frac{s}{\sqrt{n}} \to 0$ $\notin$ problema inferencial, somente análise descritiva de dados ??

Não é um consenso!

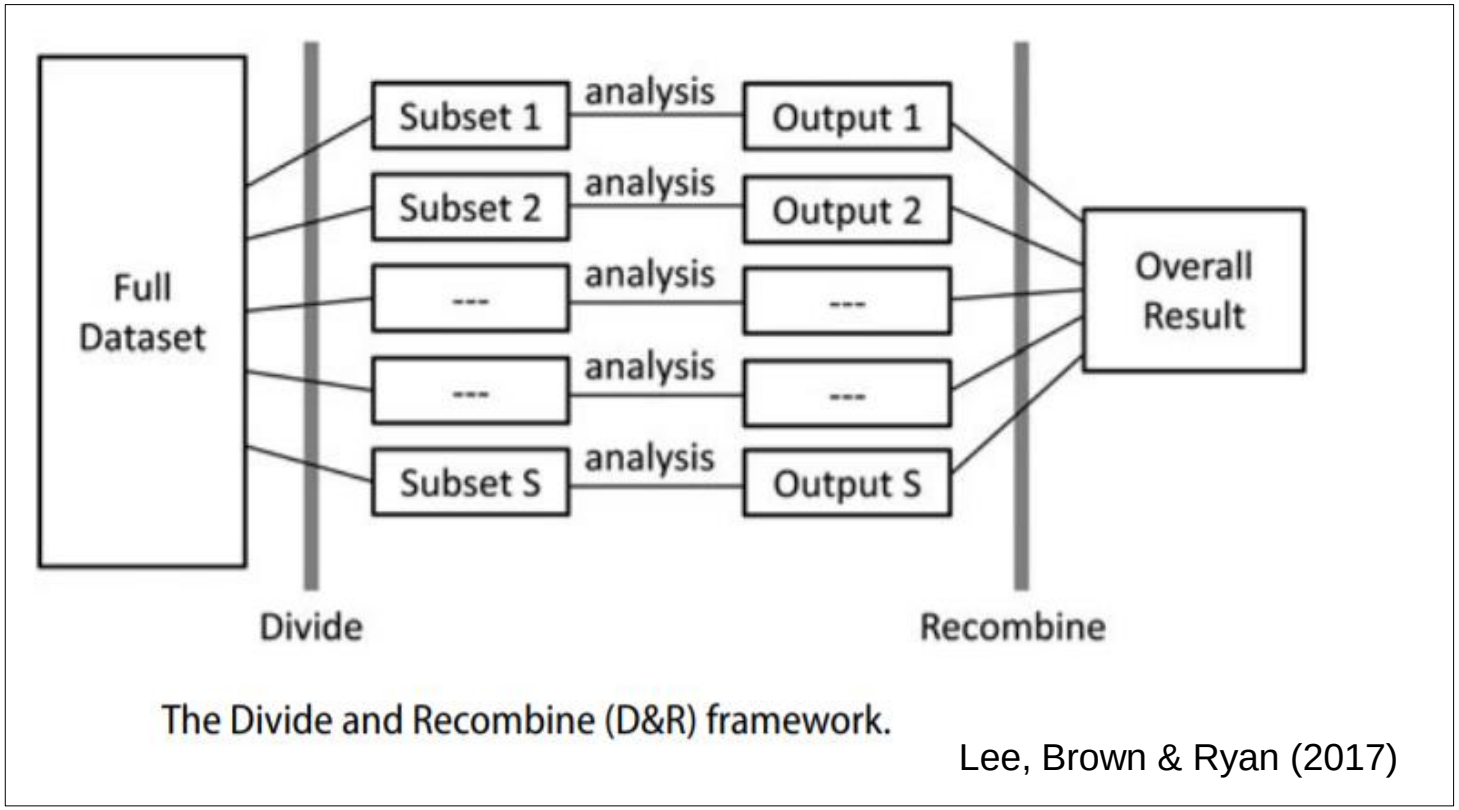
- Muitas análises multivariadas são realizadas no espaço $\Re^{nxn}$: memória e tempo computacional para leitura, armazenamento e análise dos dados é um problema

No R: `read.table()` $\Rightarrow$ `read.big.matrix()` `library(bigmemory)` `library(biganalytics)`

- Análise de Dados em Big-n $\Rightarrow$ Soluções Split-Apply-Combine ou Divide-Recombine
  - SAS: `by`
  - Google: `MapReduce`
  - Apache: `Hadoop`
  - Bibliotecas do R: `biganalytics`
    `partools` (**Software Alchemy-Matloff, 2016**)
    `foreach, doMC` (multiple Cores)

Dados Multivariados – *Big-n*

Racional das Soluções Split-Apply-Combine ou Divide-Recombine:



The Divide and Recombine (D&R) framework.

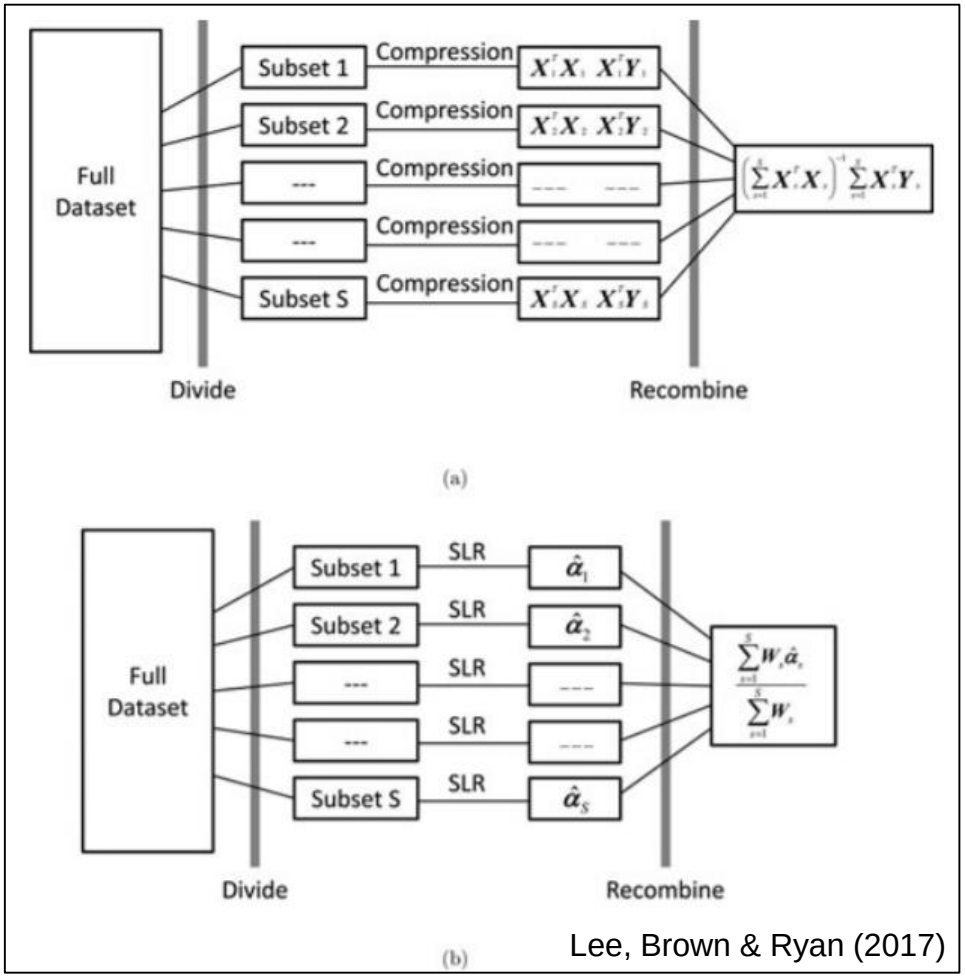
Lee, Brown & Ryan (2017)

Soluções Paralelizadas (de cálculos estatísticos em amostras "*iid*"):

- Particionar os dados via Aleatorização (com repos.) ou condicional a uma variável
- Em cada sub-amostra calcular a estatística de interesse
- Obter a medida "agregada" das estatísticas nas subamostras

Dados Multivariados – *Big-n*

Soluções Paralelizadas Splt-Apply-Combine ou Divide-Recombine (D&R):



Lee, Brown & Ryan (2017)

(a) D&R via estatísticas resumo (estat. suficiente)
(b) D&R horizontal: método clássico

Propriedades assintóticas dos EMVS na família exponencial (Portnoy, 1988)

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

Limite inferior de Kramer-Rao

$$n = \sum_{s=1}^{S} n_s; \quad n_s = \frac{n}{S}$$

D&R horizontal em Ajustes lineares

$$\Rightarrow \bar{\alpha}_n = \frac{1}{S}\sum_{s=1}^{S}\hat{\alpha}_s = \frac{\sum_{s=1}^{S} W_s \hat{\alpha}_s}{\sum_{s=1}^{S} W_s}$$

$$W_s = X_s{}' X_s$$

$$\Rightarrow Cov\left(\bar{\alpha}_n\right) = \frac{1}{S} Cov\left(\hat{\alpha}_s\right)$$

$$\bar{\alpha}_n \overset{\substack{S\ fixo \\ n\to\infty, n_s\to\infty}}{\to} \hat{\alpha}_{EMVS}$$

Para análises lineares (*lm*) tais soluções (D&R horizontal e via suficiência) são equivalentes. Para *glm* isto não ocorre!

Dados Multivariados – *Big-n*

Matloff (2015): solução paralelizada

```
require(partools)
cls <- makeCluster(2)
setclsinfo(cls)
#Gerar dados (n>>p) para ajustar Modelo Linear
n <- 10000
p <- 2
```

n/p=5000

```
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
# D&R horizontal
# (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
```

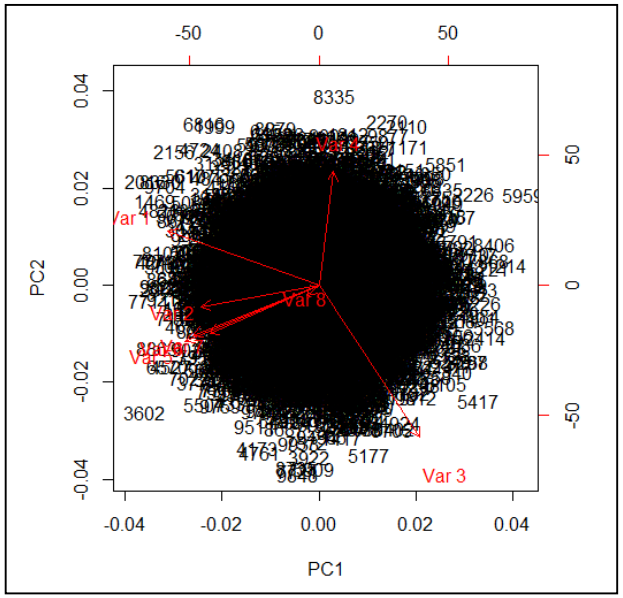
```
> head(u)
              X1          X2          Y
[1,] -0.2056215 -0.2229407 -1.712522
[2,] -0.5722810 -1.2393545 -1.965369
[3,]  0.5898422 -0.4974834  1.028857
[4,]  0.6045709 -0.9882614 -1.649882
[5,]  0.4346162  0.8333716  1.488213
[6,]  0.1783079 -0.6069011 -1.326253
```

Dados Multivariados – *Big-n*

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

#gerar dados multivariados
library(clusterGeneration)
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))
```

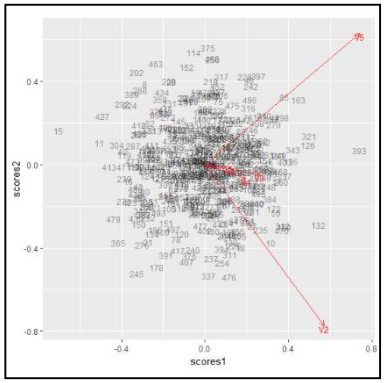
Biplot: Dados completos




```
#Rodar análise de CP via solução paralelizada (N. Matloff, 2015)
#SA: Sottware Alchemy
distribsplit(cls,"dados")
caprcomp(cls,'dados,scale=TRUE',8)$sdev
[1] 1.4707854 1.0525281 1.0046278 0.9884619 0.9051253 0.8954517
[7] 0.8450872 0.6382045
prcomp(dados,scale=TRUE)$sdev
[1] 1.4707662 1.0516048 1.0024852 0.9883842 0.9063294 0.8971400
[7] 0.8458021 0.6382948
```

Desvio padrão dos 8 primeiros CP são "iguais"

Dados Multivariados – *Big-n* Visualização



Visualização de Dados de Alta Dimensão – Big_n (Norman Matloff)

- Dificuldade: “Borrão” (BSP)

Construção do Gráfico em Coordenadas Paralelas

Representação dos pontos (perfis) mais frequentes: Padrões *TopFrequence*

Dados Contínuos: Estimar a densidade (método dos *k*-vizinhos mais próximos) e Representar os padrões TopFrequence ou os Outliers

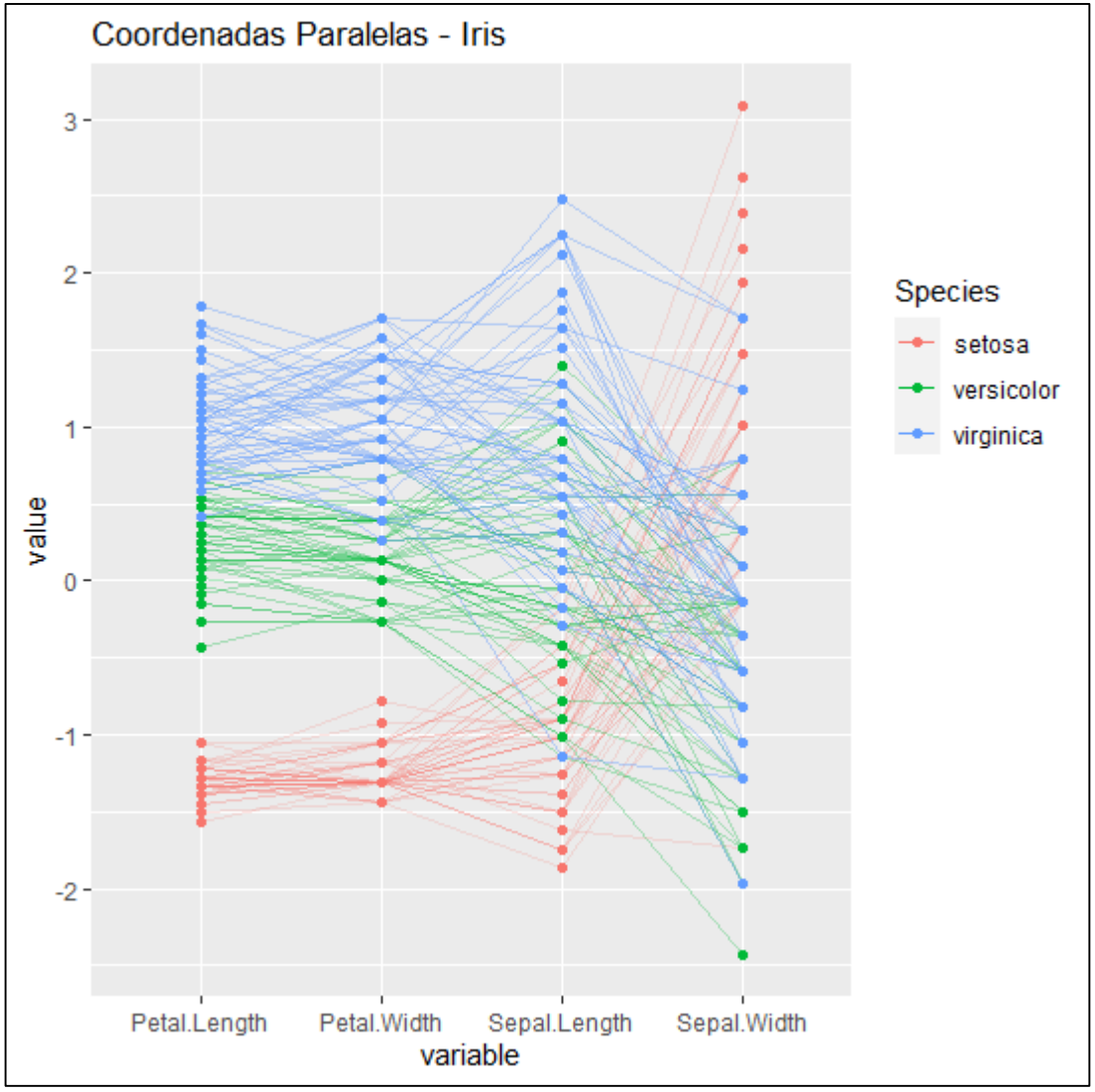
Dados Heterogêneos: Categorizar os dados contínuos e Representar os Padrões TopFrequence ou os Outliers

Gráfico em Coordenadas Paralelas

Dados – Iris

$Y_{150x(4+1)}$

Valores: escore z

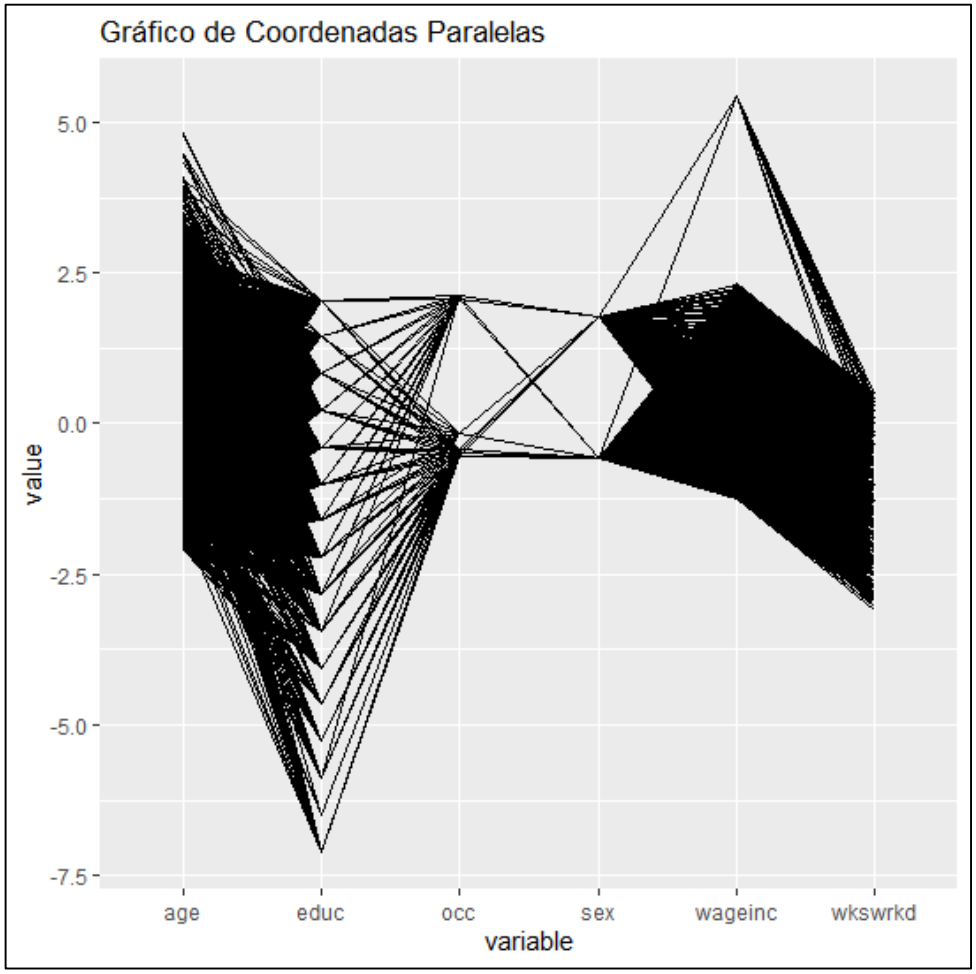


Boa visualização ("borrão" não aparece!)

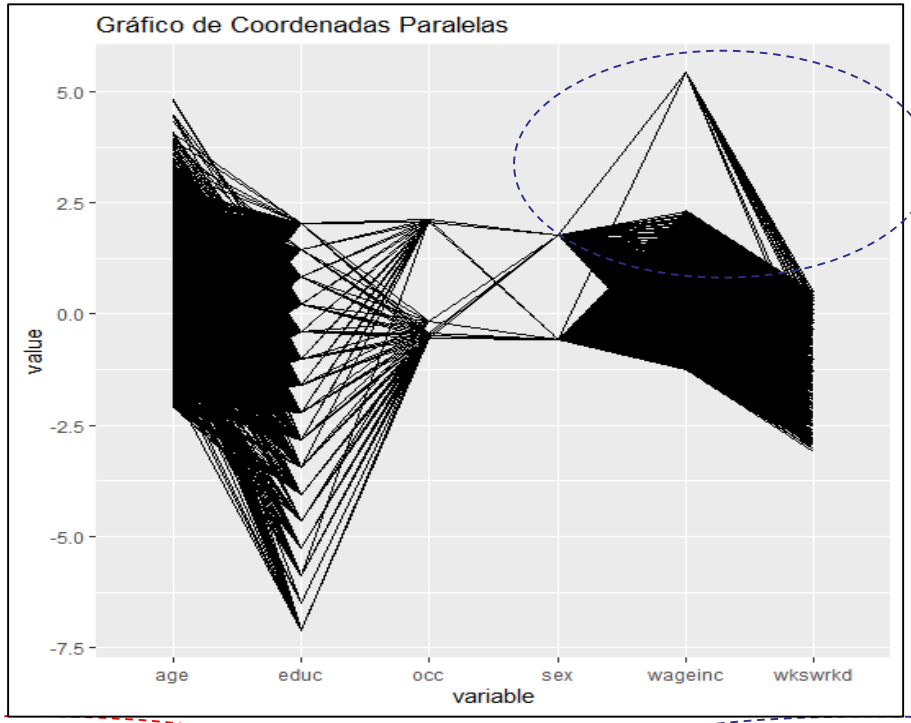
Big-n

Visualização

```
> data(prgeng)
> pe <- prgeng[,c(1,3,5,7:9)]  # extrair 6 vars de interesse
> dim(pe)
[1] 20090     6
> head(pe)
       age educ occ sex wageinc wkswrkd
1 50.30082   13 102   2   75000      52
2 41.10139    9 101   1   12300      20 …
```

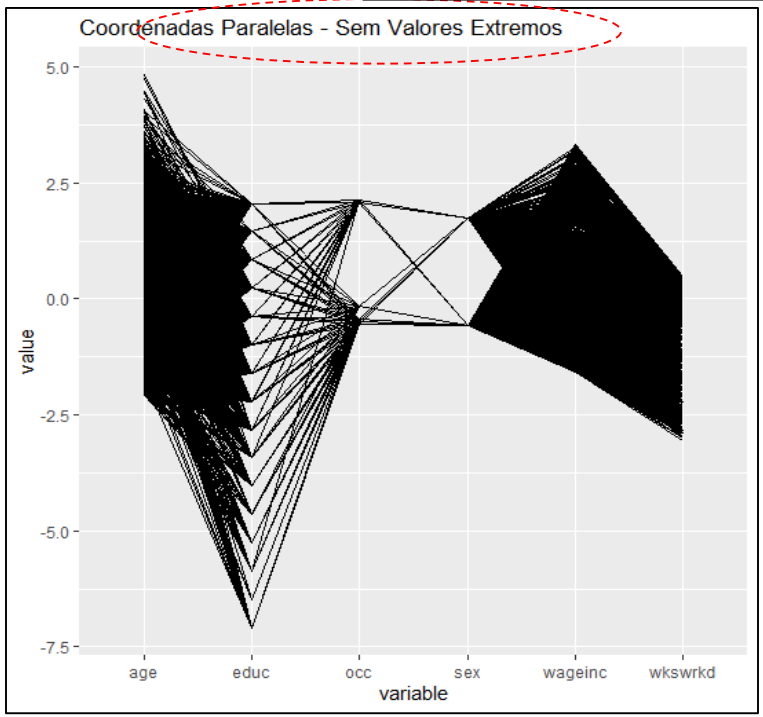


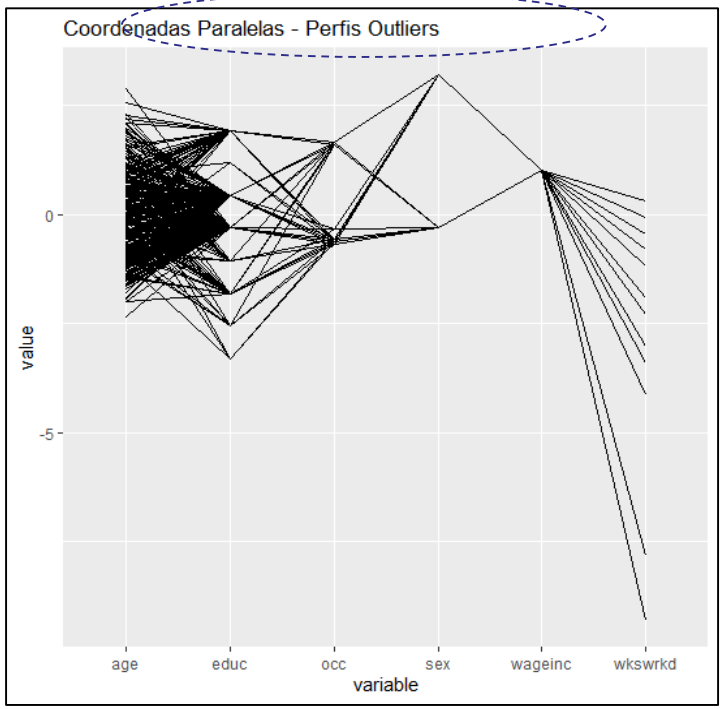
Problema: "borrão"



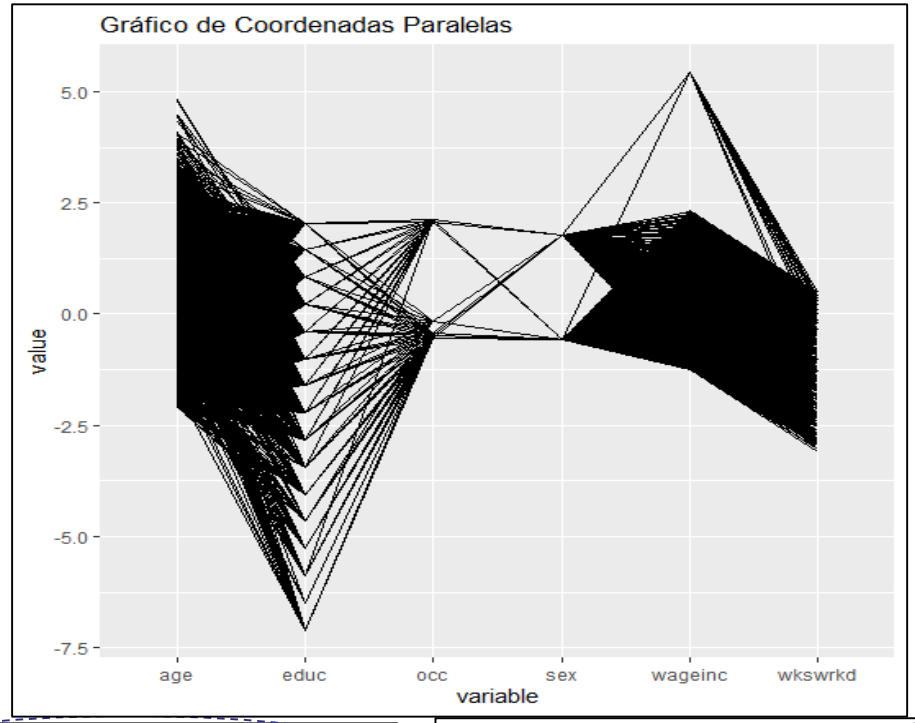
Valores extremos (outliers)

Visualização de partições dos dados:

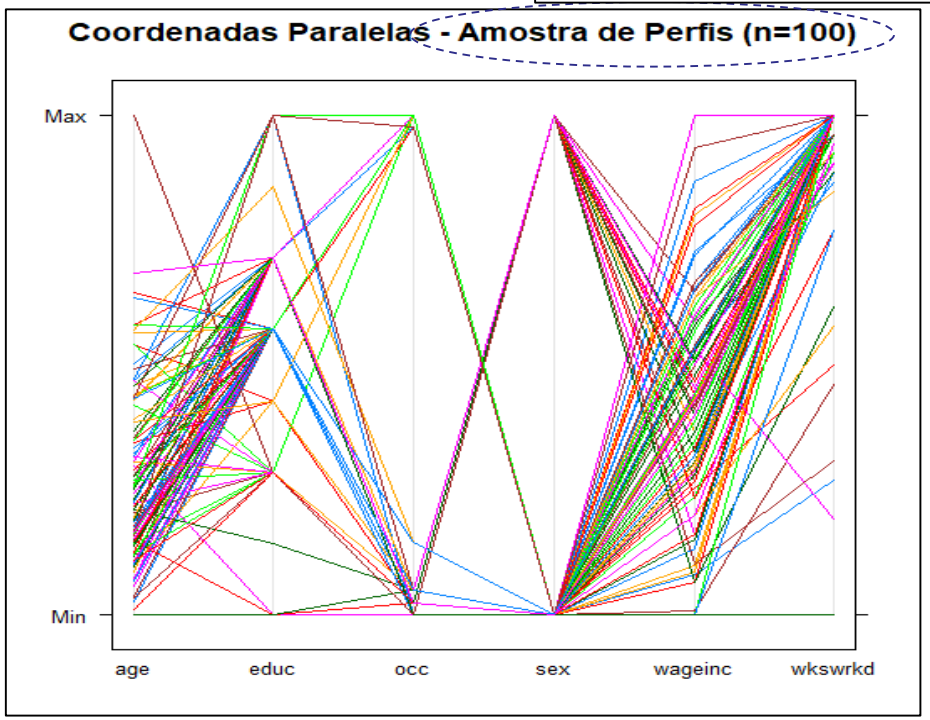


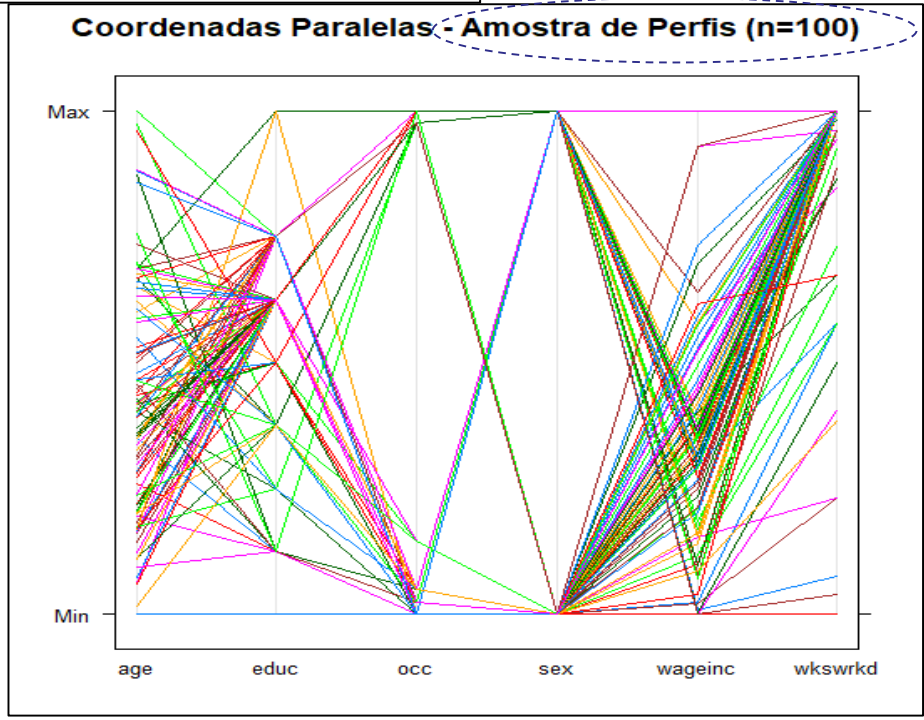


Visualização de partições dos dados:

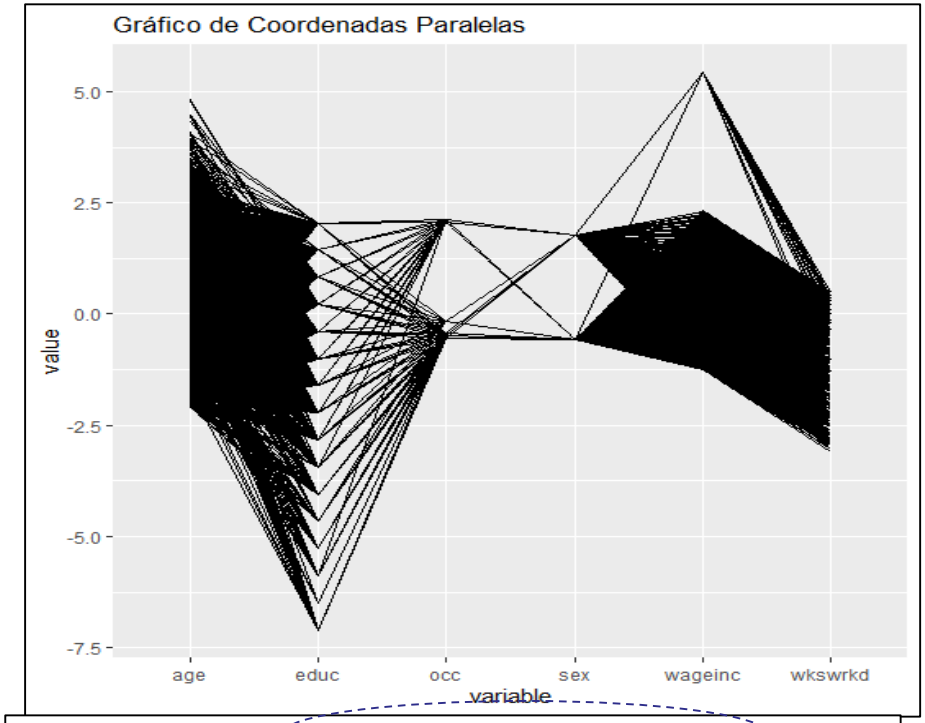


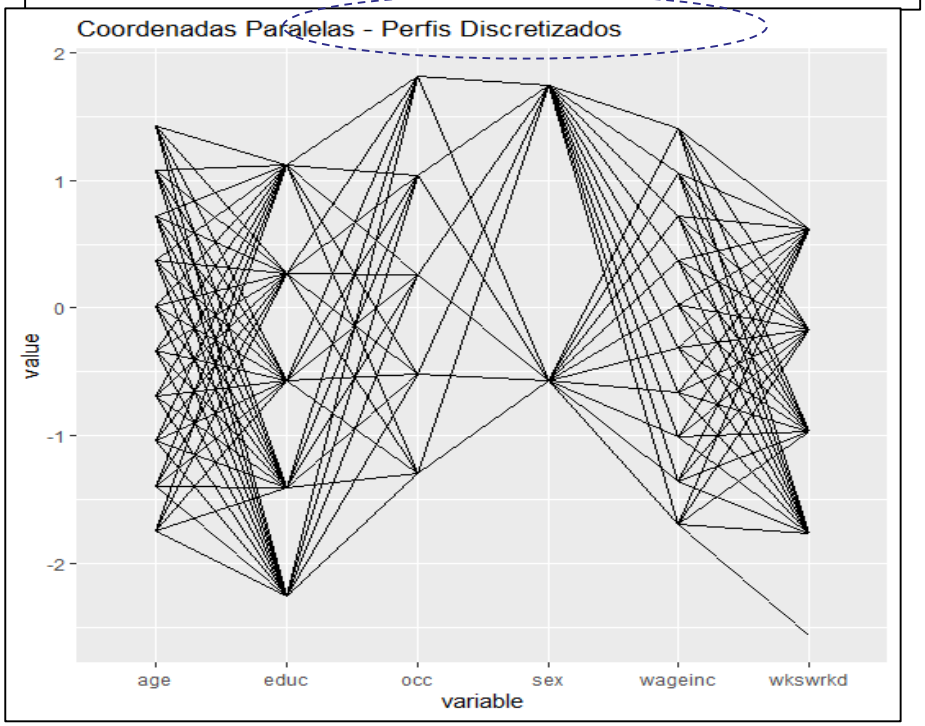
Duas Amostras aleatórias de 100 perfis



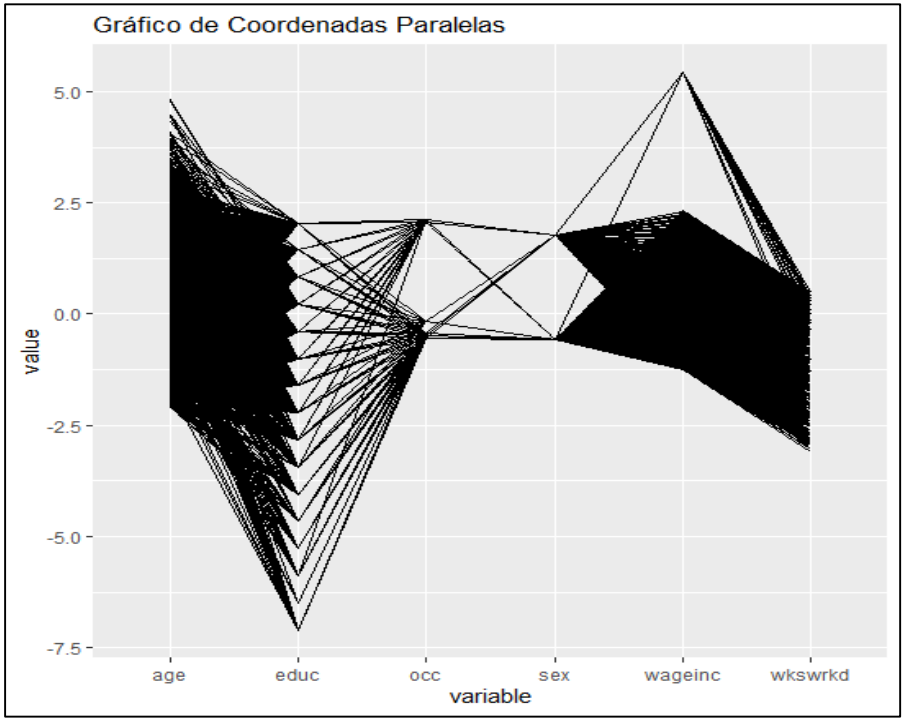


Visualização dos Dados Discretizados:

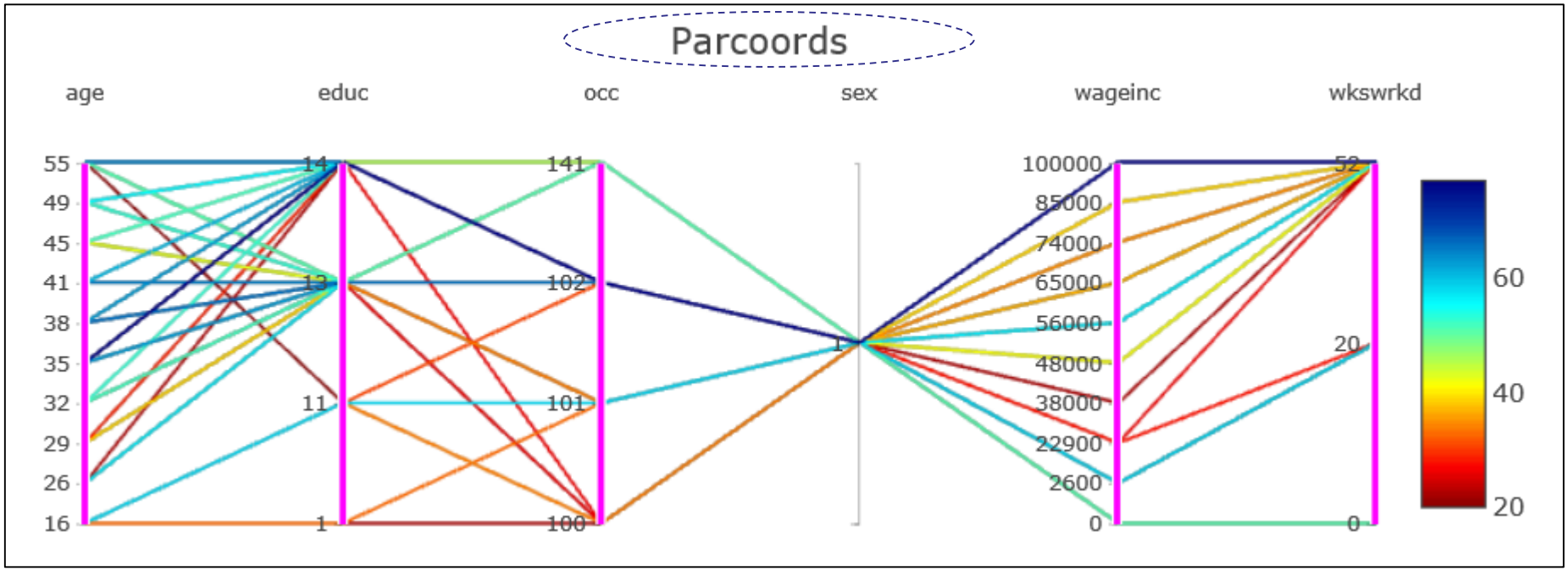




Discretizar: Dados quantitativos foram categorizados



Top 100 Perfis mais Frequentes



Dados Multivariados – *Big Data*

Principais Classes de Problemas (Galeano e Peña, 2019):

- Quantificar Dados provenientes de diferentes fontes (textos, imagens, etc.) e Realizar a Anotação
- Integrar Dados heterogêneos (diferentes escalas)
- Visualização de Dados em Alta Dimensão
- Teste de Múltiplas Hipóteses (α corrigido)
- Métodos automatizados de Seleção de Modelos
- Ajustes de Modelos via soluções Esparsas
- Teoria de Redes (Grafos para Resumir o Padrão de Relação entre Variáveis)

MAE5776-Análise Multivariada

Objetivo: Fornecer base teórica de inferência estatística e fatoração de matrizes na análise de dados multivariados. Apresentar e discutir técnicas de redução de dimensionalidade, integração de bancos de dados e **aprendizado de estruturas**, sob contextos supervisionados e não-supervisionados. Serão abordados casos clássicos (n>p; observações independentes) e soluções em espaços estruturados (n<<p; n muito grande; **observações dependentes**)

Conteúdo (geral):

1. Introdução: estrutura de dados, medidas resumo multivariadas, propriedades em espaços duais.
2. Distribuição Normal Multivariada: propriedades, estimação, distribuições amostrais, testes de hipóteses para vetores de médias e matrizes de covariância.
   Elipses de concentração, regiões de confiança, *outliers* multivariados, gráficos multivariados.
3. Técnicas (clássicas) de redução da dimensionalidade (n>p e observações independentes). Teoria de Fatoração de Matrizes na redução de dimensionalidade e integração de bancos de dados.
4. Técnicas de redução da dimensionalidade em espaços mais gerais (n<<p, n muito grande, observações não independentes): soluções em espaços duais, soluções regularizadas e penalizadas, reamostragem.
5. **Temas adicionais**: Modelos de Equações Estruturais; Teoria de Grafos Probabilísticos e Aprendizado de Estruturas; Análise de Dados heterogêneos.

Componentes Principais em Dados Heterogêneos

$$Y_{n\times p} = \underbrace{U_{n\times n}\Lambda_r^{1/2}}_{A}\underbrace{V_{p\times r}}_{B} \quad \overset{\text{(Gabriel, 2008)}}{\Rightarrow} \quad \min_{a,b}\sum_{i=1}^{n}\sum_{j=1}^{p}(Y_{ij} - a_i b_j)^2$$

A: matriz de Escores
B: Matriz de Cargas

GRLM (Generalized Low Rank Model): Extensão da Análise de Componentes Principais (Udell et al., 2016)

$$\min_{a_i,b_j} \sum_{(i,j)\in\Omega} L\left(Y_{ij} - a_i b_j\right) + \lambda\sum_{i=1}^{n} r(a_i) + \lambda\sum_{j=1}^{p}\tilde{r}(b_j)$$

Funções Perda Regularizadores

Obter Arquétipos: "Escores = a_i" e "Cargas = b_j"

Flexibilidade da solução: combinar diferentes funções Perda e Regularizadores ((L1, L2), bem como realizar a imputação de dados

Modelo	$L_j(u,v)$	$r(a)$	$\tilde{r}(b)$
CP	$(u-v)^2$	0	0
CPs Regularizados	$(u-v)^2$	$\|\|a\|\|_2^2$	$\|\|b\|\|_2^2$
NNMF	$(u-v)^2$	$\mathbb{I}(a \geq 0)$	$\mathbb{I}(b \geq 0)$
CP esparso	$(u-v)^2$	$\|\|a\|\|_1$	$\|\|b\|\|_1$
CP Robusto	$\|u-v\|$	$\|\|a\|\|_2^2$	$\|\|b\|\|_2^2$
CP Logístico	$log(1+esp(-vu))$	$\|\|a\|\|_2^2$	$\|\|b\|\|_2^2$
CP Binário	$(1-vu)_+$	$\|\|a\|\|_2^2$	$\|\|b\|\|_2^2$
K-means	$(u-v)^2$	$\mathbb{I}(card(a)=1)$	0

Imputação em Dados *Missing*:

$$\hat{Y}_{ij} = a_i\, b_j; \quad (i,j) \notin \Omega$$

$$\hat{Y}_{ij} = \arg\min_x L_{ij}\left(a_i b_j, y\right); \quad (i,j) \notin \Omega$$

Implementação no Python (glrm.py), Julia and H2O.
Ana G Vasconcelos (Mestrado-IME, 2020)