MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

Verão IME/2020

Análise Multivariada

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

Programa da disciplina ☺

- 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*) e Esparsas (*n<<p, iid*): CP, AD, CC
- Componentes Principais em Observações Dependentes (base familiar)
- Redução de Dimensionalidade em Dados Heterogêneos
- Aprendizado de Estruturas: Teoria de Grafos, Modelos de Equações Estruturais

Caracterização dos Dados Multivariados:

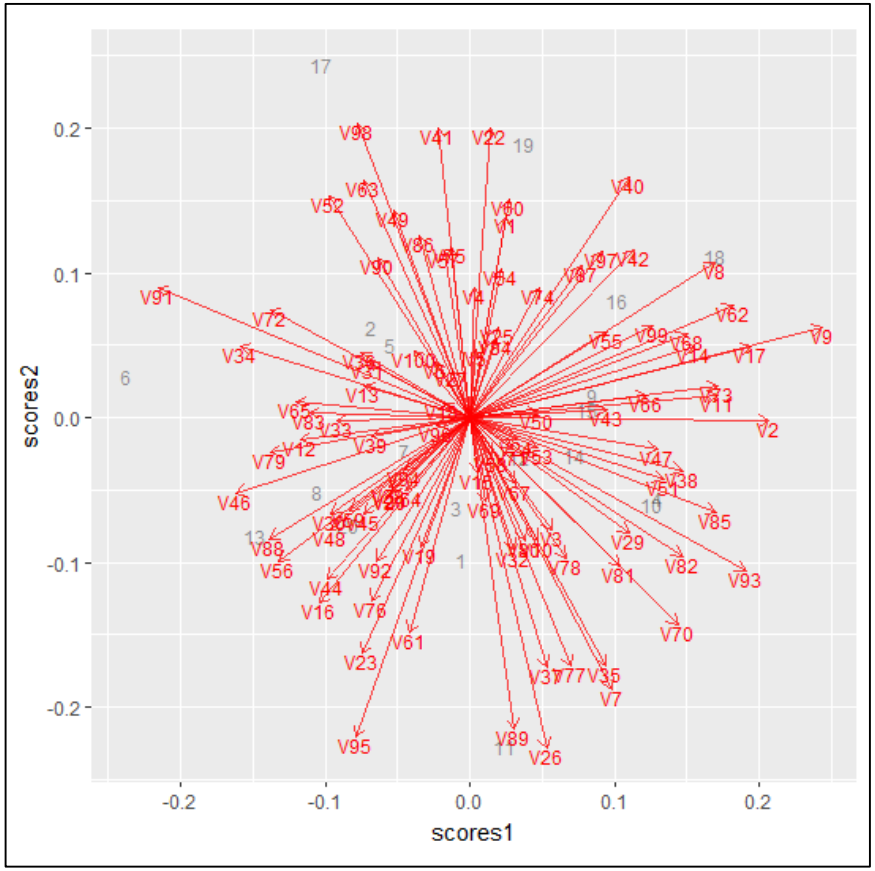
Razão n/p?

- Dimensão do Espaço das Unidades Amostrais: Big-n (n >> p)?
- Dimensão do espaço das Variáveis: Big-p (n << p)?
- Classificação das Variáveis (quantitativas → dados heterogêneos)
- Estrutura de Dependência entre Variáveis
- Estrutura de Dependência entre as Unidades Amostrais

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

Big-p (n << p)

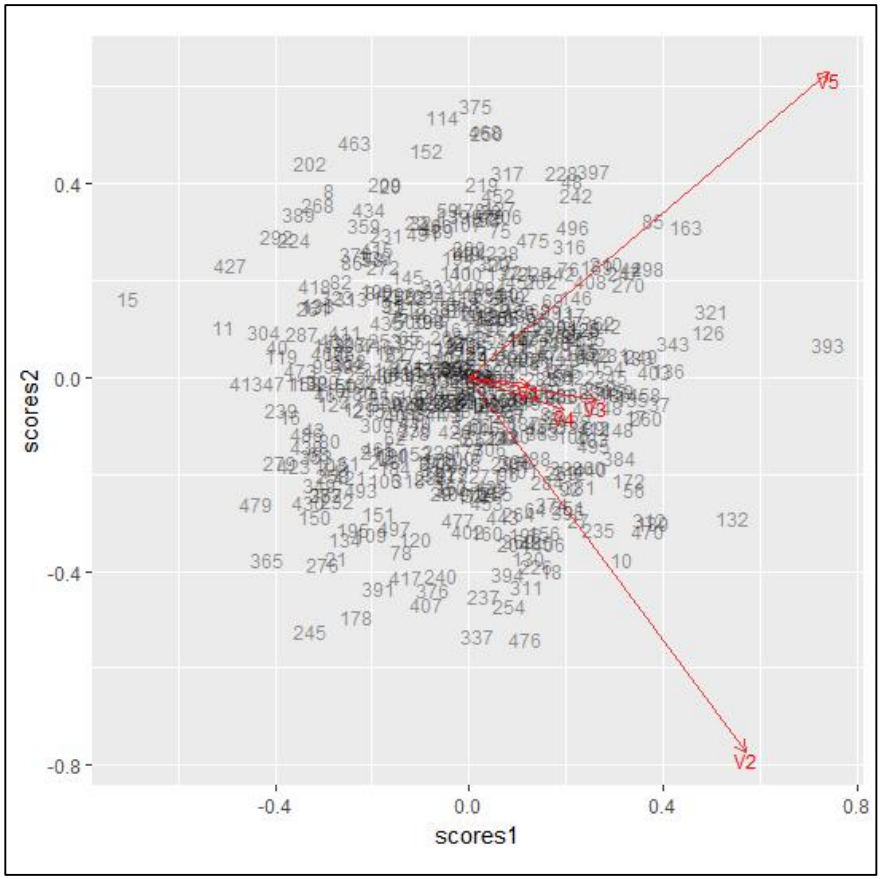
Biplot: n=20 p=100



Seleção de variáveis: Soluções regularizadas e penalizadas!

Big-n (n >> p)

Biplot: n=500 p=5



Sumarização e visualização ?

(Black Screen Problem: *R_alpha blending*)

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 ($n \lll p$)	$1 \leq \frac{n}{p} < 10$ 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

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

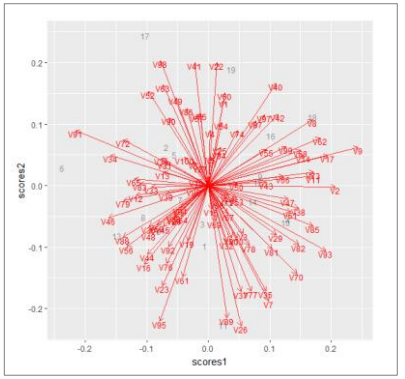
$$\hat{\theta}_{EMVS} \overset{\frac{p^2}{n} \to 0;\, p,n \to \infty}{\sim} N\left(\theta; v(\theta)\right)$$

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

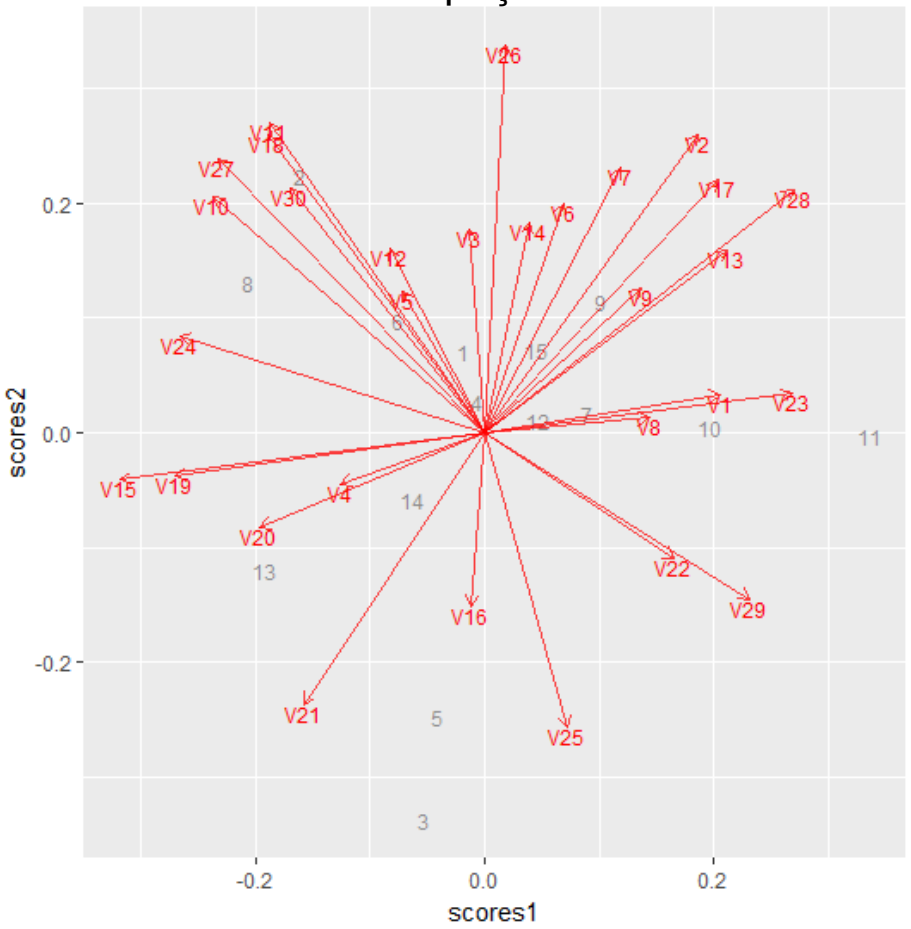
$$\max_{U_k, V_k} U_k'\, Y\, V_k; \quad \begin{cases} \left\| U_k \right\|_2^2 \le 1 \\ \left\| V_k \right\|_2^2 \le 1, \ \left\| V_k \right\|_1 \le c_1 \end{cases}$$

Algoritmo de maximização (Witten et al., 2009)

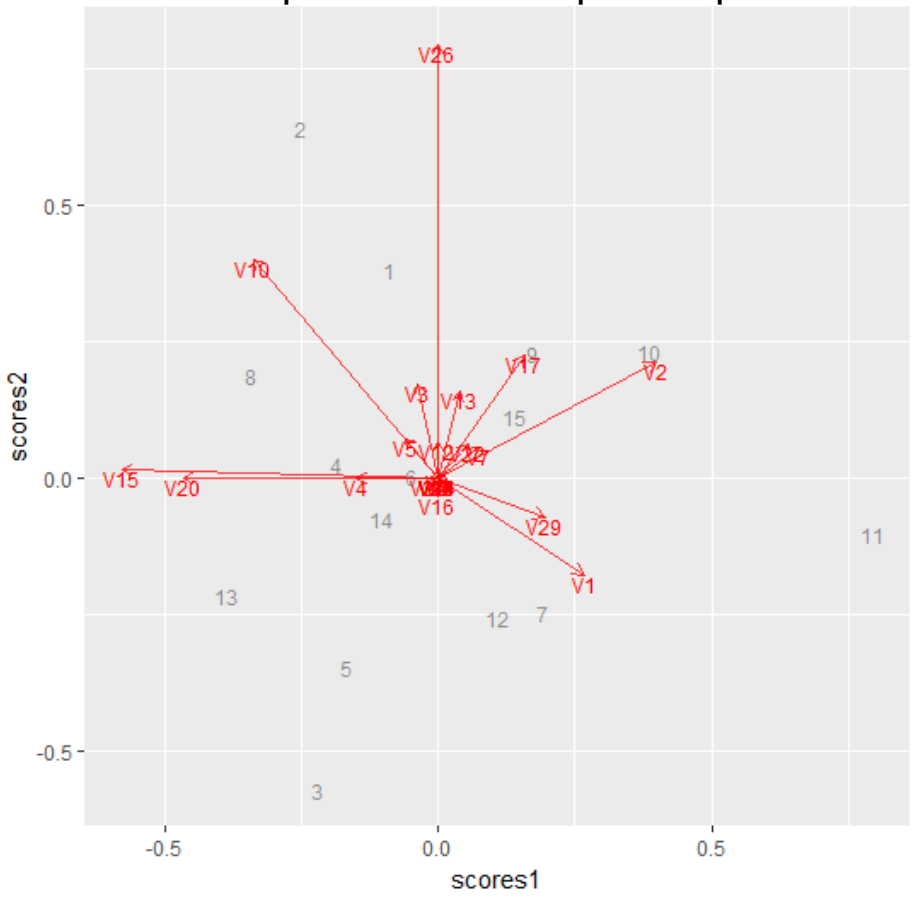
Componentes Principais – n<<p

Representação Biplot: n=15 p=30

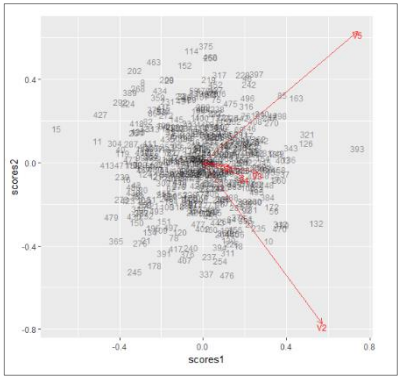
R-prcomp: Componentes Principais em Espaços Duais



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*

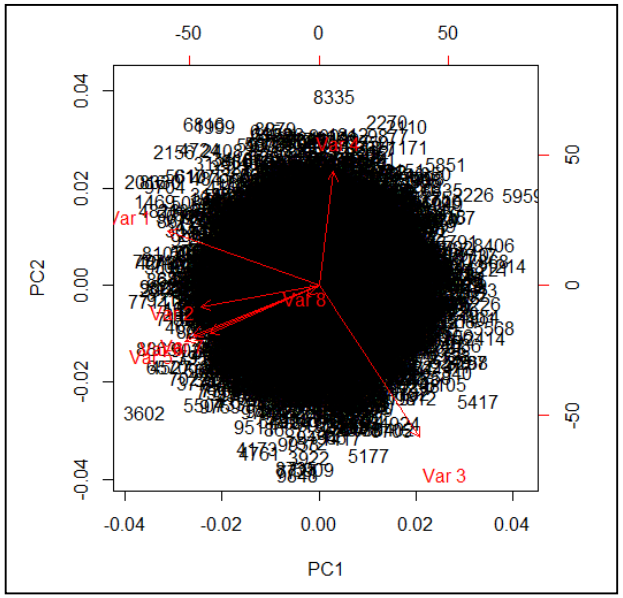
```
require(partools)
cls <- makeCluster(2)
setclsinfo(cls)
#Gerar dados (n>>p) para ajustar Modelo Linear
n <- 10000
p <- 2
set.seed = 1099
tmp <- matrix(rnorm((p+1)*n),nrow=n)
u <- tmp[,1:p] # gerar valores "X"
# adicionar coluna de valores "Y"
u <- cbind(u,u %*% rep(1,p) + tmp[,p+1])
colnames(u) = 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
```

```
> 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*

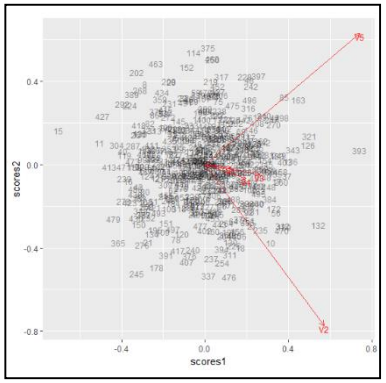
```
require(partools)
cls <- makeCluster(2)
setclsinfo(cls)

#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))
```




```
#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



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

- Dificuldade: “Borrão” (BSP)
- Construção do Gráfico de Coordenadas Paralelas
- Representação dos pontos (perfis) mais frequentes: Padrões *TopFrequence*

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

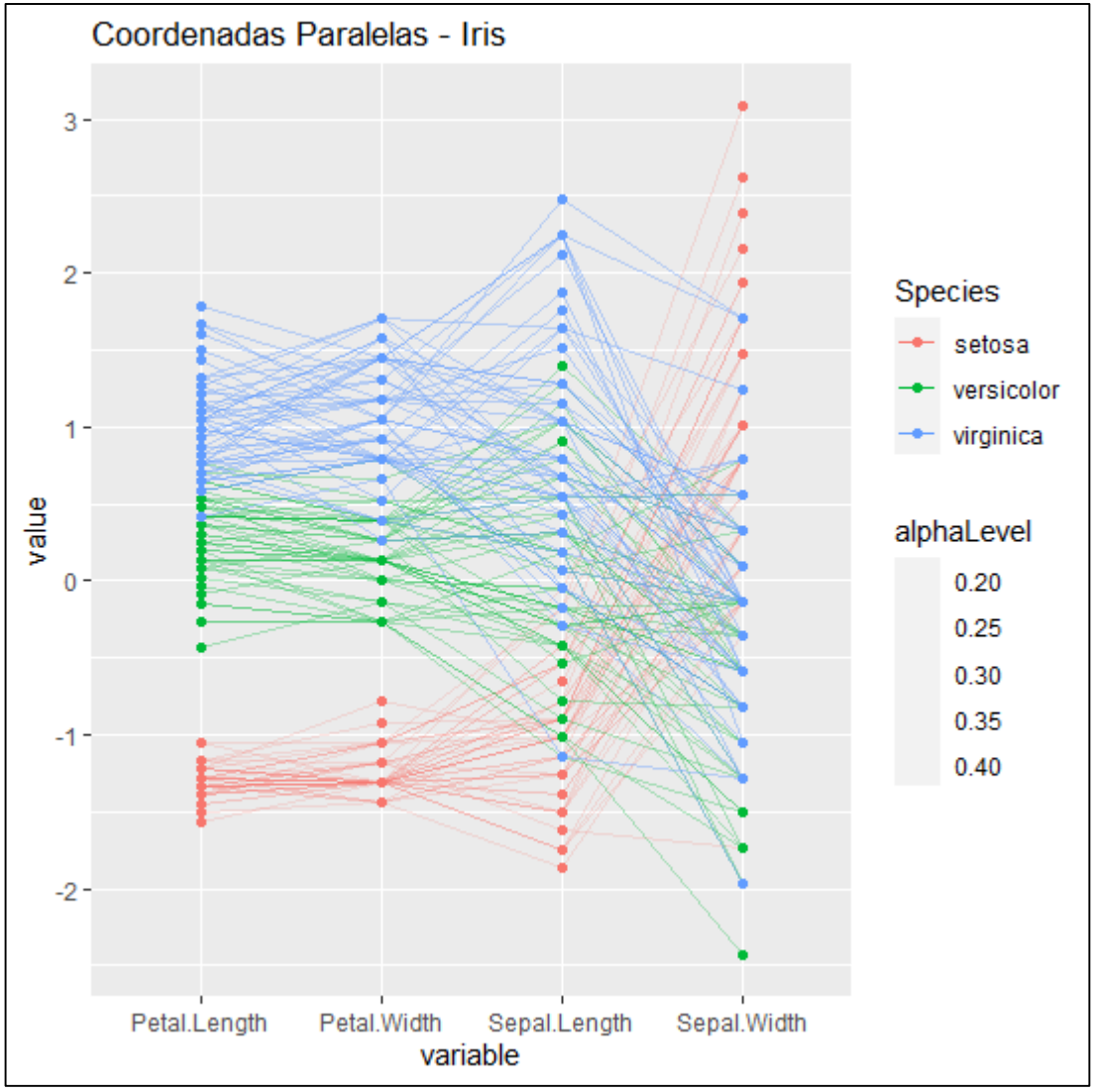
Dados Heterogêneos: Discretizar Dados

Representar Padrões TopFrequence

Representar Padrões Outliers

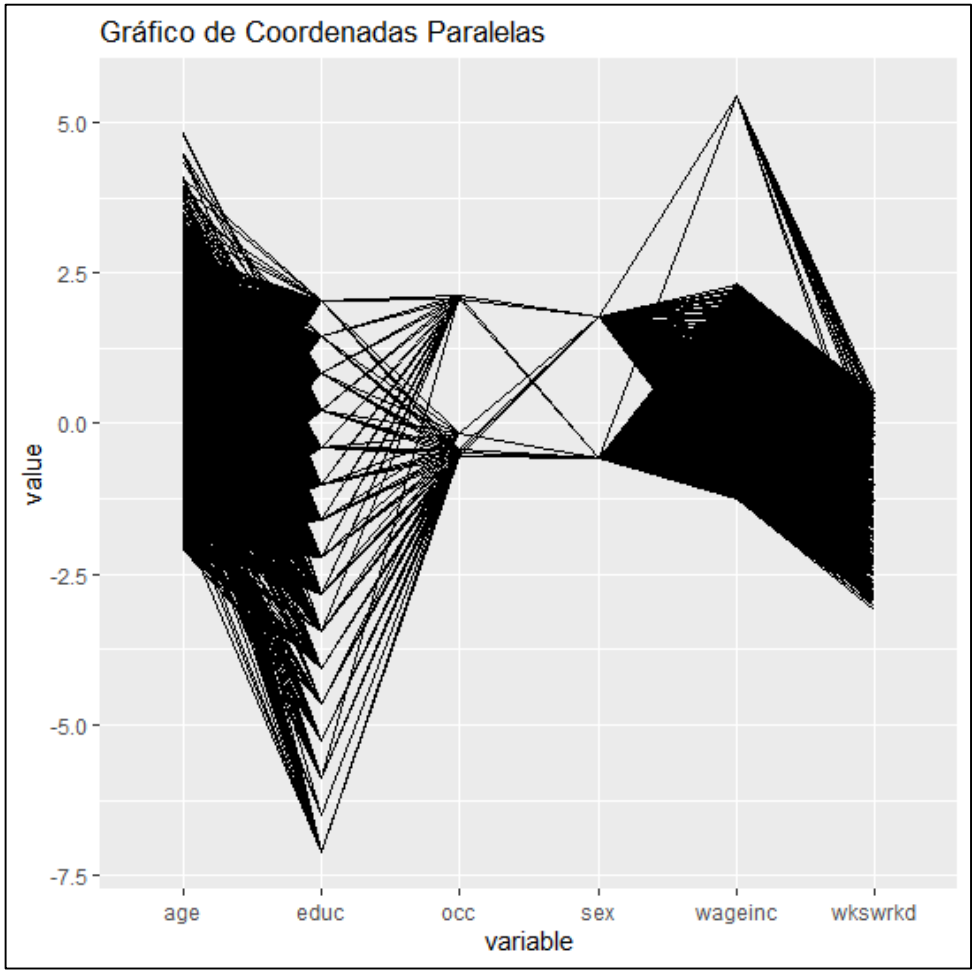
Gráfico de Coordenadas Paralelas

Dados - Iris



Big-n
Visualização

```
> data(prgeng)
> pe <- prgeng[,c(1,3,5,7:9)]  # extrair 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”

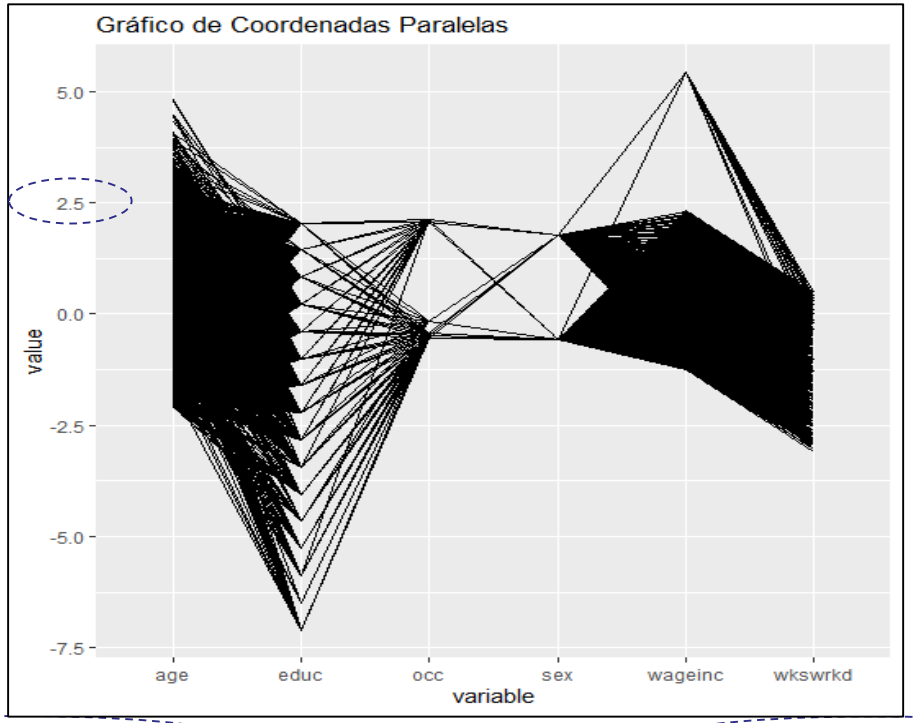
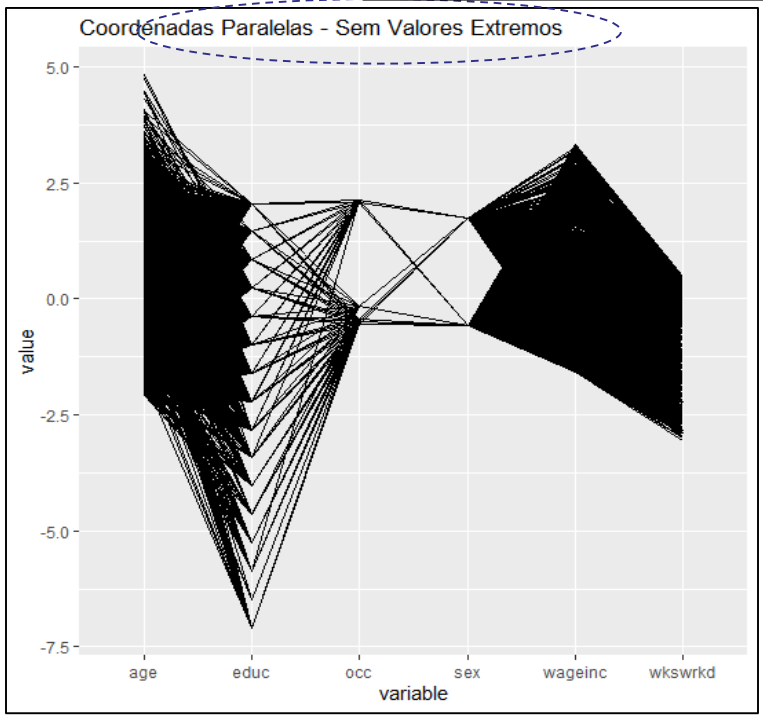
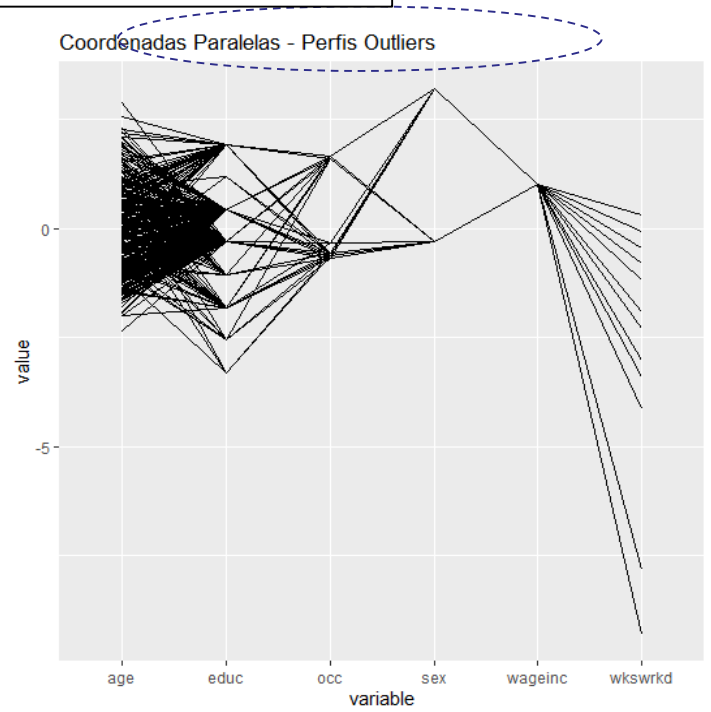
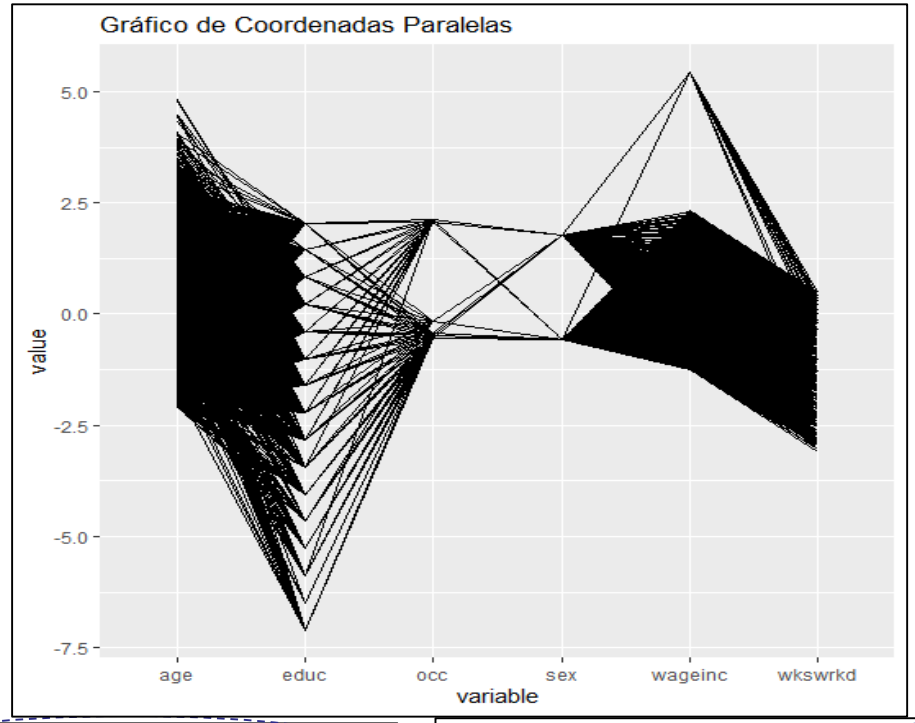


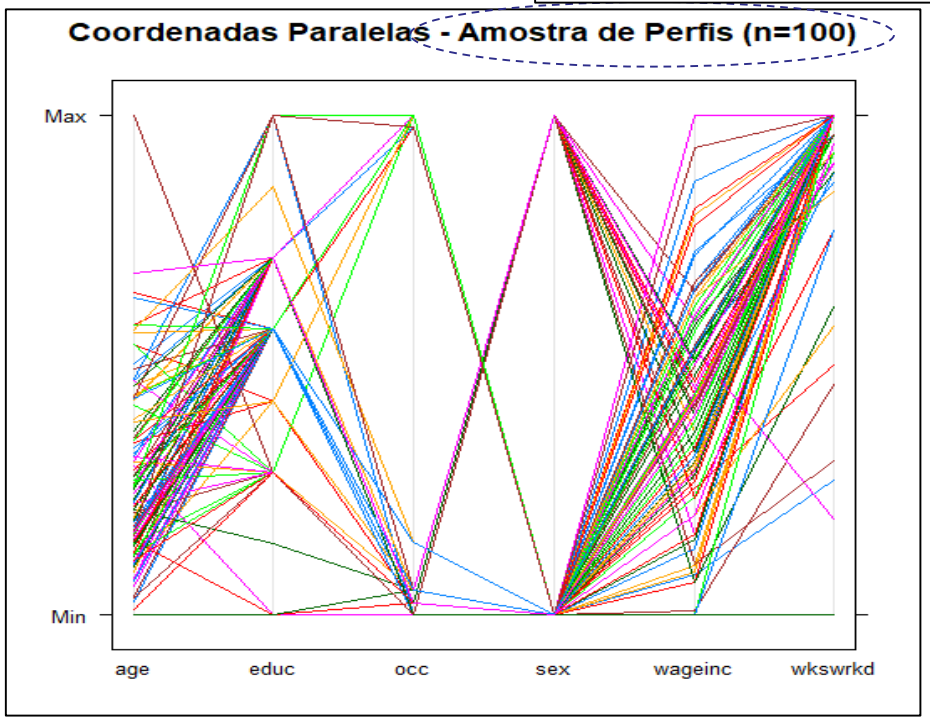
Gráfico de Coordenadas Paralelas
5.0
2.5
0.0
-2.5
-5.0
-7.5
value
age
educ
occ
sex
wageinc
wkswrkd
variable

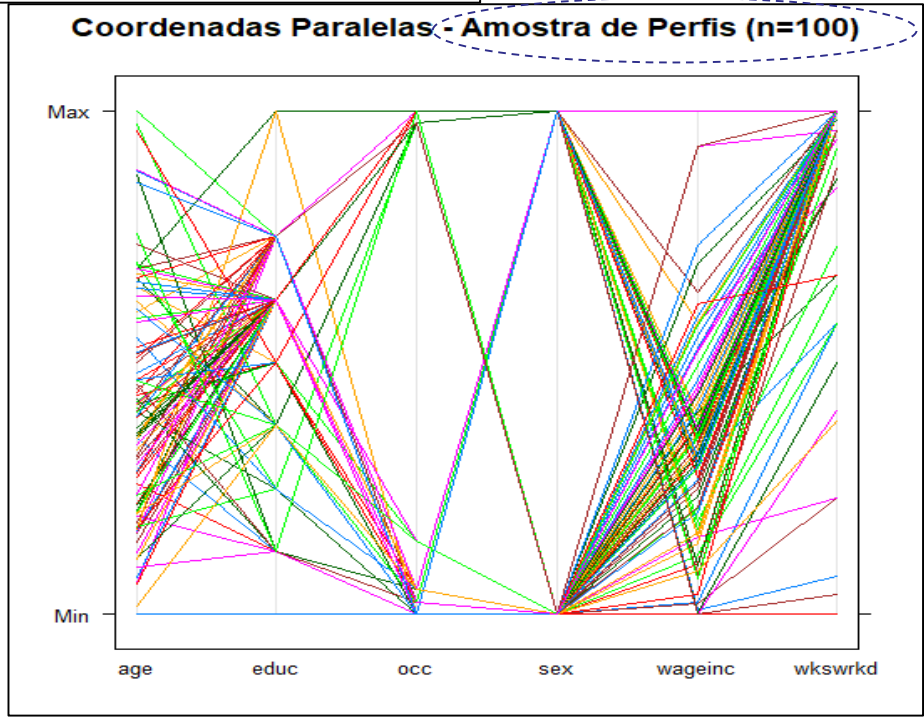

Coordenadas Paralelas - Sem Valores Extremos
5.0
2.5
0.0
-2.5
-5.0
-7.5
value
age
educ
occ
sex
wageinc
wkswrkd
variable


Coordenadas Paralelas - Perfis Outliers
0
-5
value
age
educ
occ
sex
wageinc
wkswrkd
variable



Amostra aleatória de 100 perfis





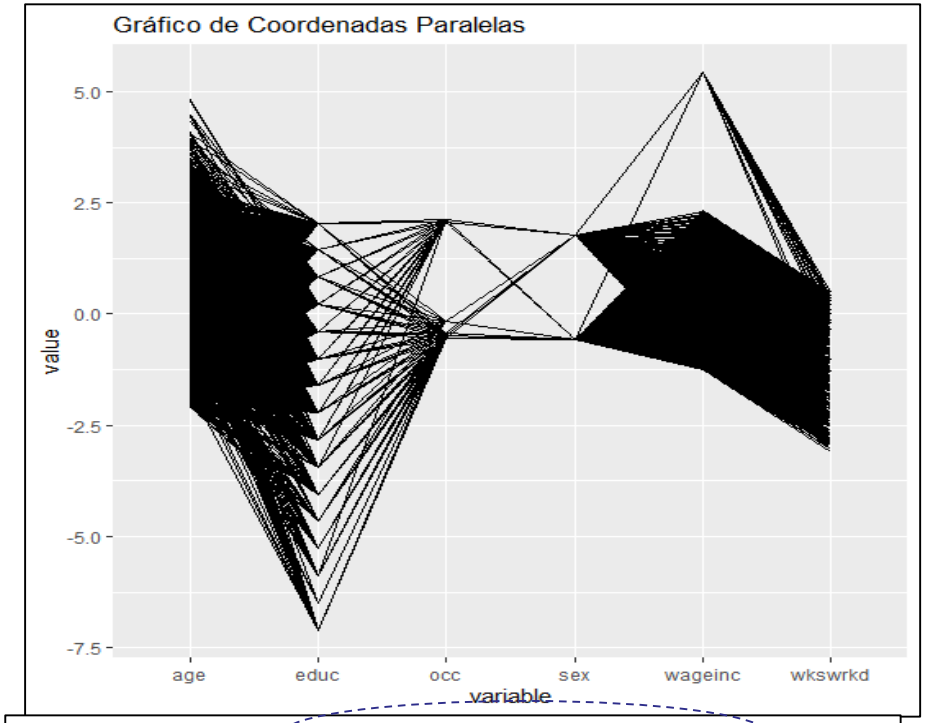
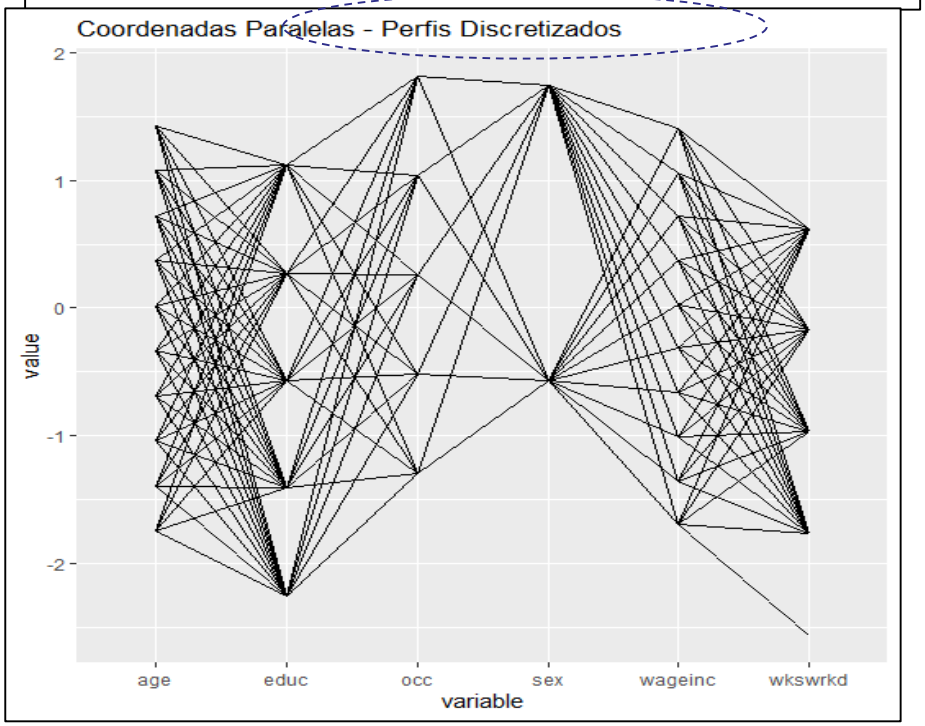
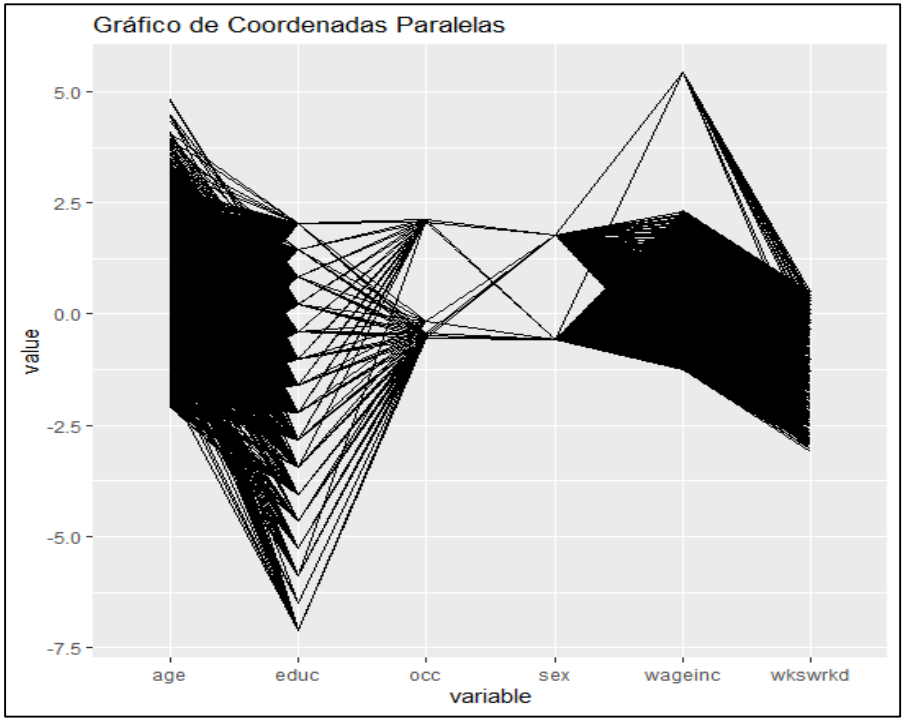


Gráfico de Coordenadas Paralelas
5.0
2.5
0.0
-2.5
-5.0
-7.5
value
age
educ
occ
sex
wageinc
wkswrkd
variable


Coordenadas Paralelas - Perfis Discretizados
2
1
0
-1
-2
value
age
educ
occ
sex
wageinc
wkswrkd
variable



Top 100 Perfis mais Frequentes