MBA Big Data e Business Intelligence

PCA -Principal Component Analsysis

Prof. Abraham Laredo Sicsu

Ideia

- Notas dos alunos de uma classe
  - His ➔ história
  - Mat ➔ matemática
  - Geo ➔ geografia
  - Qui ➔ química
  - Fis ➔ física

3

Matrizes de correlação

FGV EDUCAÇÃO EXECUTIVA

R=

	Hist	Mat	Geo	Qui	Fis
Hist	**1,00**				
Mat	**0,02**	**1,00**			
Geo	**0,96**	**0,13**	**1,00**		
Qui	**0,42**	**0,71**	**0,50**	**1,00**	
Fis	**0,01**	**0,85**	**0,11**	**0,79**	**1,00**

reordenando

R=

	Hist	Geo	Mat	Qui	Fis
Hist	**1,00**				
Geo	**0,96**	**1,00**			
Mat	**0,02**	**0,13**	**1,00**		
Qui	**0,42**	**0,50**	**0,71**	**1,00**	
Fis	**0,01**	**0,11**	**0,85**	**0,79**	**1,00**

FGV EDUCAÇÃO EXECUTIVA

$X_1, X_2, \ldots\ldots, X_P$

Variáveis originais (intervalares / razão)

$CP_1, CP_2, \ldots, CP_P$

Componentes principais:

Variáveis não correlacionadas tais que

$CP_j = w_{ji} X_1 + \ldots\ldots + w_{jp} X_p \qquad j = 1, \ldots\ldots, p$

$CP_1, \ldots\ldots, CP_k$

k < p → redução de dimensionalidade: menor número de variáveis que conservem quase toda a informação contida nas variáveis originais

5

Ideia

$CP_j = w_{ji}\,AV_1 + \ldots\ldots + w_{jp}\,AV_p$

PROF	AV1	AV2	AV3	AV4	AV5	AV6
1	6,4	5,9	6,9	6,9	5,0	8,2
2	8,3	6,1	7,6	9,4	3,0	8,7
3	8,0	4,9	4,3	4,0	5,0	6,2
---	---	---	---	---	---	---
14	10,0	3,9	5,3	8,0	5,0	3,2
15	10,0	5,8	6,6	6,9	4,0	9,5

1

'PROF	CP1	CP2	CP3	CP4	CP5	CP6
1	0,444	-0,428	0,399	-0,206	0,544	0,003
2	1,433	1,123	-1,059	-0,963	0,124	-0,015
3	-0,897	-0,853	0,129	-0,162	-0,405	-0,955
---	---	---	---	---	---	---
13	-2,677	-0,487	0,342	-0,348	0,697	-0,012
14	-0,529	-1,624	-1,231	-0,609	-0,995	0,308
15	1,241	0,152	-0,504	-0,274	-0,097	-0,998

2

REGRAS DE REDUÇÃO

'PROF	CP1	CP2	CP3	CP4	CP5	CP6
1	0,444	-0,428	0,399	-0,206	0,544	0,003
2	1,433	1,123	-1,059	-0,963	0,124	-0,015
3	-0,897	-0,853	0,129	-0,162	-0,405	-0,955
---	---	---	---	---	---	---
13	-2,677	-0,487	0,342	-0,348	0,697	-0,012
14	-0,529	-1,624	-1,231	-0,609	-0,995	0,308
15	1,241	0,152	-0,504	-0,274	-0,097	-0,998

1: como obter as CP's

2: quantas CP's reter

3: como interpretar as CP´s

Objetivos

- Trabalhar com menos variáveis (**redução de dimensionalidade**) sem perda significativa da informação contida nos dados.
- Trabalhar com novas variáveis que contenham **boa parte da informação** original e que sejam **não correlacionadas**
- Entender a **estrutura dos dados** através da análise das componentes principais: como interpretar cada componente principal?
  - Que dimensão dos dados **Yj** representa ?

Interpretação geométrica I

- Caso extremo: colinearidade perfeita

x1	x2
1	23
2	41
3	59
4	77
5	95
6	113
7	131
8	149
9	167
10	185
11	203
12	221
13	239
14	257
15	275
16	293
17	311
18	329

Interpretação geométrica I

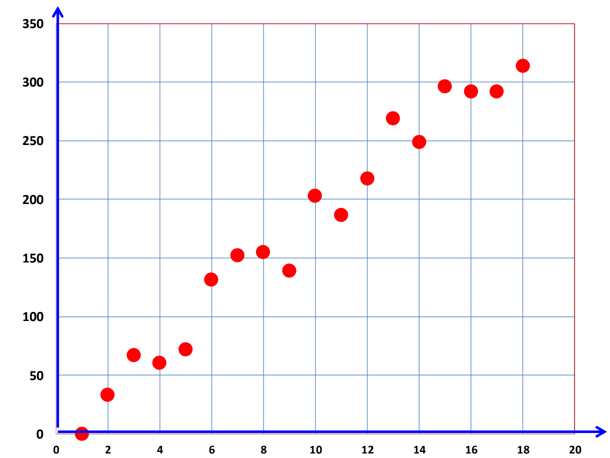
- Caso extremo: colinearidade perfeita

x1	x2
1	23
2	41
3	59
4	77
5	95
6	113
7	131
8	149
9	167
10	185
11	203
12	221
13	239
14	257
15	275
16	293
17	311
18	329

Interpretação geométrica II

FGV EDUCAÇÃO EXECUTIVA

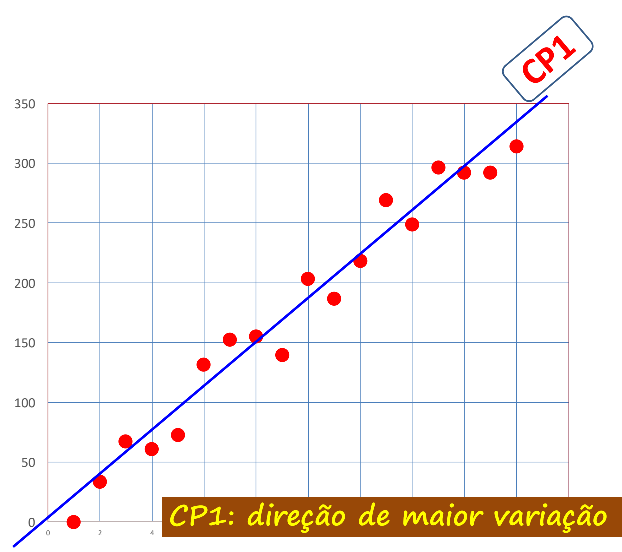
x1	x2
1	0
2	33,5
3	67,2
4	61
5	72,5
6	131,6
7	152,5
8	155,1
9	139,5
10	203,5
11	186,6
12	218,2
13	269,3
14	249,1
15	296,5
16	292,1
17	292,4
18	314,2



Interpretação geométrica II

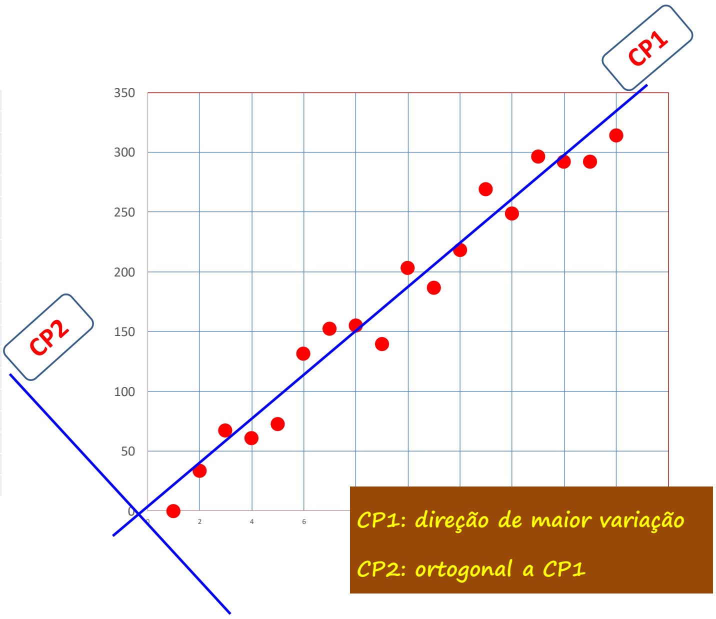
FGV EDUCAÇÃO EXECUTIVA

x1	x2
1	0
2	33,5
3	67,2
4	61
5	72,5
6	131,6
7	152,5
8	155,1
9	139,5
10	203,5
11	186,6
12	218,2
13	269,3
14	249,1
15	296,5
16	292,1
17	292,4
18	314,2



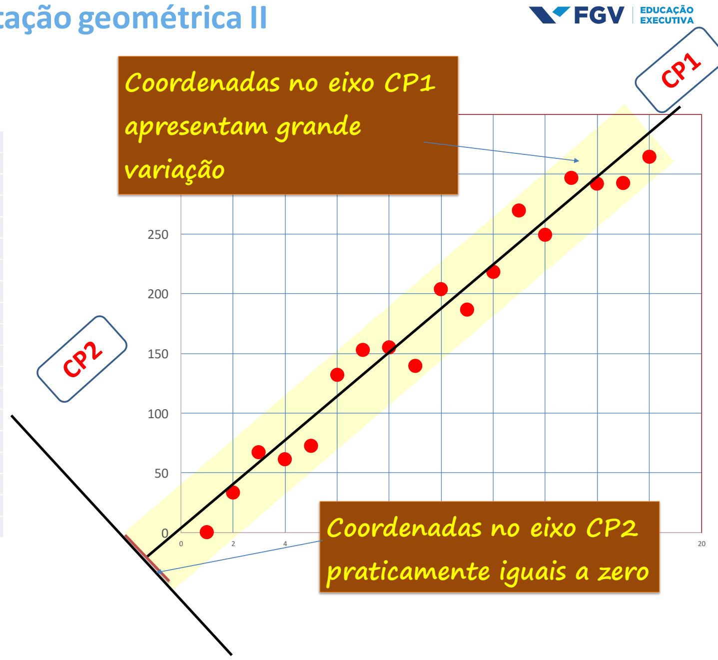
Interpretação geométrica II

x1	x2
1	0
2	33,5
3	67,2
4	61
5	72,5
6	131,6
7	152,5
8	155,1
9	139,5
10	203,5
11	186,6
12	218,2
13	269,3
14	249,1
15	296,5
16	292,1
17	292,4
18	314,2



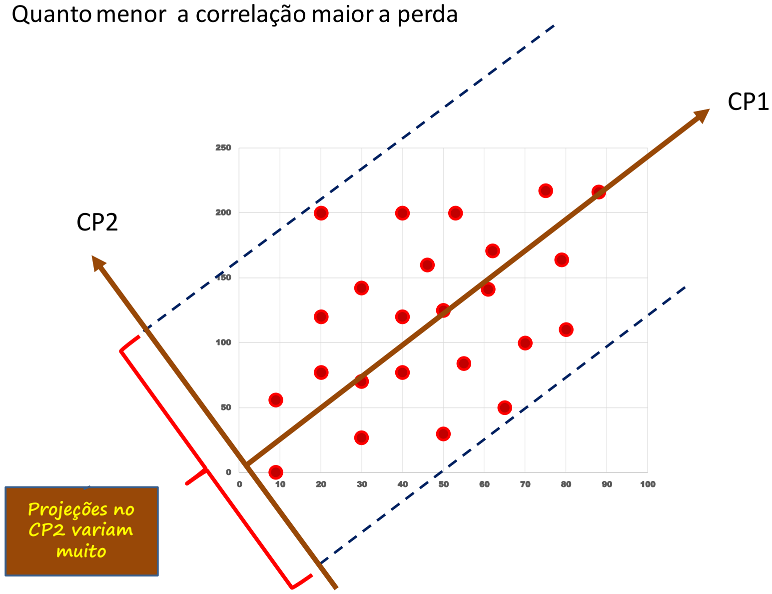
Interpretação geométrica II

x1	y
1	0
2	33,5
3	67,2
4	61
5	72,5
6	131,6
7	152,5
8	155,1
9	139,5
10	203,5
11	186,6
12	218,2
13	269,3
14	249,1
15	296,5
16	292,1
17	292,4
18	314,2



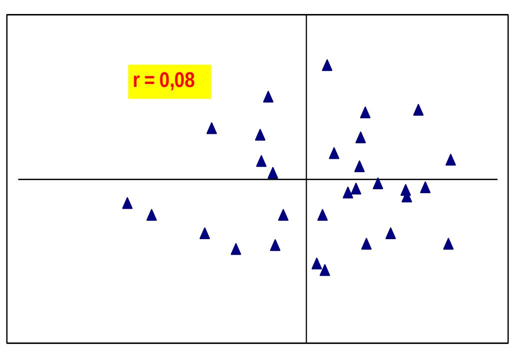
Interpretação geométrica III

- Quanto menor a correlação maior a perda



14

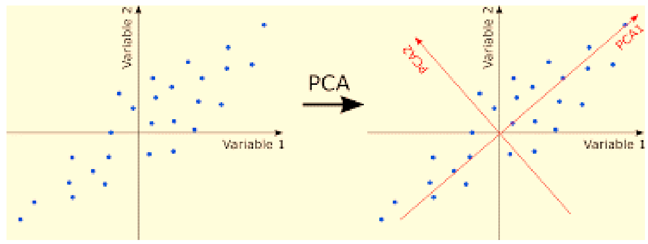
Correlação ≈ 0



Se variáveis forem independentes, as CPs serão as próprias variáveis originais.(Ver demonstração em J&W)

Há testes apropriados (Bartlett e outros – ver referências) para avaliar correlação significativa.

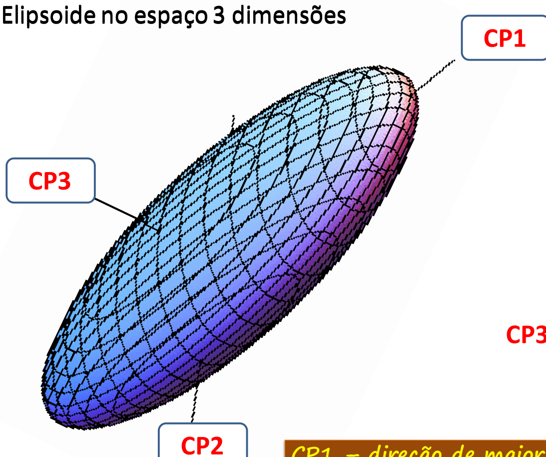
Álgebra matricial

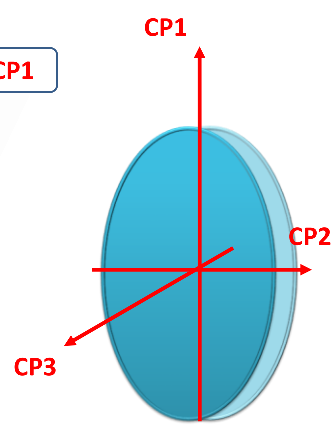


- **Terminologia (R: matriz de correlação das variáveis)**
  - O vetor unitário na direção PCA1 → 1º autovetor (eigenvector) de R
  - O vetor unitário na direção PCA2 → 2º autovetor (eigenvector) de R
  - Variação das projeções em PCA1 → medida pelo 1º autovalor (eigenvalue) de R
  - Variação das projeções em PCA1 → medida pelo 2º autovalor (eigenvalue) de R

Interpretação geométrica IV – no espaço

- Elipsoide no espaço 3 dimensões





CP1 – direção de maior variação

CP2: maior variação perpendicular a CP1

CP3: maior variação perpendicular a CP1 e CP2

17

Dados originais ou dados padronizados ?

Qual usar ?

- ✓ As soluções com dados originais e dados padronizados diferem. Não há uma fórmula para passar de uma solução para a outra.
- ✓ Decisão depende dos objetivos do problema.
- ✓ **Em geral, quando as variáveis são medidas em escalas diferentes recomenda-se a padronização.** ***(não faz sentido combinar linearmente variáveis de diferentes dimensões)***
- ✓ Se uma (ou mais) variável tiver variância significativamente maior que a das outras, ela poderá dominar uma componente principal (o que é intuitivo, por definição das CPs). Neste caso, se não quisermos ter uma direção associada a essa única variável, devemos padronizar ou transformar essa variável para redução da variância
- ✓ Se variáveis forem medidas nas mesmas unidades e forem de magnitudes similares recomenda-se (por motivos que não serão aqui explicados – ver Morrison) utilizar as variáveis originais, ou seja, não trabalhar com variáveis padronizadas.

Cuidado

software dá os scores padronizados; é diferente de trabalhar com os scores das variáveis padronizadas !

18

Questões em ACP

- Como medir informação e perda de informação ?
- Devemos trabalhar com variáveis originais ou variáveis padronizadas?
- **Como determinar os componentes principais ? Ou seja, os pesos w_{ij} ?**

$$Y_j = w_{j1} X_1 + \dots\dots + w_{jp} X_p$$

- Como interpretar as CPs ?
- Quantas CPs reter sem "grande" perda de informação ? Qual o critério a utilizar ?
- Como utilizar as CPs posteriormente ?

19

Variância total – medida de informação

- Variância Total das variáveis Xi

$$\boldsymbol{vartot}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_p) = \sum_i \boldsymbol{var}(\boldsymbol{X}_i)$$

Utilizando **variáveis padronizadas**

$$\boldsymbol{var}(\boldsymbol{X}_i) = \mathbf{1} \ \rightarrow \boldsymbol{vartot}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_p) = \mathbf{p}$$

Variância das componentes principais

$$\boldsymbol{vartot}(\boldsymbol{PC}_1,\ldots,\boldsymbol{PC}_p) = \sum_i \boldsymbol{var}(\boldsymbol{PC}_i)$$

Demonstra-se que **para variáveis Xi padronizadas**

$$\boldsymbol{vartot}(\boldsymbol{PC}_1,\ldots,\boldsymbol{PC}_p)\text{=}\boldsymbol{vartot}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_p) = \boldsymbol{p}$$

Cuidado: as variâncias das componentes principais não são necessariamente iguais a 1

20

Variância total – medida de informação

$$vartot(PC_1,\ldots,PC_p)\text{=}vartot(X_1,\ldots,X_p) = p$$

$$vartot(PC_1,\ldots,PC_k) = var(PC_1)+\ldots+var(PC_k) \quad k < p$$

Variância explicada pelas k primeiras componentes principais a partir das variáveis padronizadas

$$\%explicada = \frac{vartot(PC_1,\ldots,PC_k)}{p}$$

v

21

Variância total – medida de informação

$$vartot(PC_1, \ldots, PC_p) = vartot(X_1, \ldots, X_p) = p$$

Exemplo:

- $X_1, X_2, \ldots, X_{40}$ padronizadas . Vimos que

  $vartot(X_1, X_2, \ldots, X_{40}) = 40$ & $vartot(CP_1, CP_2, \ldots, CP_{40}) = 40$

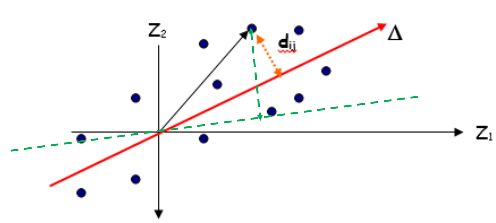
- Suponhamos que com 5 CP .... $vartot(CP_1, CP_2, \ldots, CP_5) = 32$
  - Variância explicada = 80%
  - Perda de informação = 20%

22

Inércia – determinação das CP

1. Direção do **primeiro eixo principal Δ_1** (Y_1) não é a da reta de mínimos quadrados. Eixo **Δ_1** é a reta de menor inércia para a nuvem N* de pontos definida pelos 15 professores:
   Inércia (N*, Δ_1) : **medida da dispersão** da nuvem N* ao redor da reta Δ_1

Inércia (N* , Δ) = $\Sigma_{i,j} (d_{ij})^2$ (Δ_1 é a reta que minimiza essa somatória)



- Δ_1 passa pela origem (estamos trabalhando com variáveis padronizadas)
- O vetor diretor de Δ_1 é u_1, auto vetor da matriz de correlações R associado ao maior auto valor dessa matriz

Δ_1 a reta que "melhor" se ajusta à nuvem pontos dentro dessa definição de distância à reta. Δ_2 é a reta ortogonal a Δ_1 que "melhor" se ajusta aos pontos (no caso p = 2, solução é única para Δ_2)

Determinação das CP

CP_1 : Combinação das variáveis X_i determinada de forma que

- Var CP_1 seja máxima
- CP1 define "direção" de variação máxima

CP_2 :Combinação das variáveis X_i determinada de forma que:

- CP_2 seja ortogonal a CP1 (CP1 **e** CP2 não correlacionadas)
- Var (CP2) máxima
  - limitada por Vartot(X) - Var(CP1)

CP_3 : Combinação das variáveis X_i determinada de forma que:

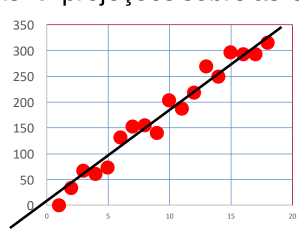
- seja ortogonal a CP1 e CP2
- Var CP3 máxima
  - limitada por Vartot(X) - Var(CP1) - - Var(CP2)

etc, ...

24

Variâncias (CP's a partir de variáveis padronizadas)

novas variáveis → projeções sobre as CP



Var (CP_1) = λ_1 : maior auto valor ("*eingenvalue*") da matriz **R**

Var (CP_2) = λ_2 : segundo maior auto valor da matriz **R**

....

Var (CP_p) = λ_p : menor auto valor da matriz **R**

25

Correlação entre as variáveis Z_i a CP_j

$$CP_j = w_{j1} Z_1 + \ldots + w_{jp} Z_p \qquad Var(CP_j) = \lambda_j$$

- Demonstra-se que

  Correlação (Z_i, CPj) = $w_{ji}\sqrt{\lambda_j}$ ($\sqrt{\lambda_j}$: raiz quadrada de λ_j)

- Correlação (Z_i, CPj) é denominada ***Component loading***
  - **facilitam interpretação as componentes principais**.
- Correlação (Z_i, CPj) = Correlação (X_i, CPj)

26

Arquivo BANCO MCL

CLIENTE	**Identificação do cliente**
REGIAO	**Região da agencia**
LIMCRED	**satisfação com o limite de crédito**
SATHOR	**satisfação com horários de atendimento (podem diferir nas diferentes agencias)**
PARKING	**qualidade do estacionamento**
INSTAL	**avaliação das instalações da agência do banco**
ATENDM	**satisfação com atendimento recebido nas agências do banco**
GLOBAL	**nível de satisfação global**

CLIENTE	REGIAO	LIMCRED	SATHOR	PARKING	INSTAL	ATENDM	GLOBAL
1001	norte	7,0	9,5	7,1	5,9	5,6	8,0
1002	norte	8,3	7,4	5,8	5,0	7,2	8,1
1003	centro	8,5	7,9	6,2	6,0	8,7	7,8
...							

27

Análise das variáveis & correlações

- Simplificando notação

> bb=BANCO_MCL

- Com auxílio dos box plots não detectamos outliers
- Matriz de correlações

> bb.num=bb[,3:8]

> print(cor(bb.num),digits = 2)

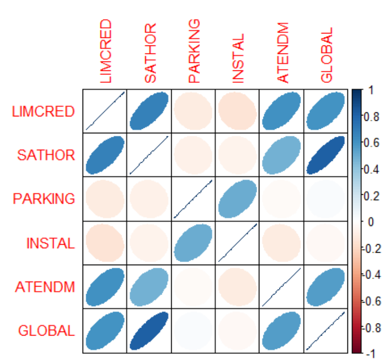
> library(corrplot)

> corrplot(RR, method="ellipse",addgrid.col = 1) #meu preferido

28

Análise das variáveis

	LIMCRED	SATHOR	PARKING	INSTAL	ATENDM	GLOBAL
LIMCRED	1.00	0.678	-0.110	-0.149	0.602	0.598
SATHOR	0.68	1.000	-0.078	-0.065	0.477	0.810
PARKING	-0.11	-0.078	1.000	0.496	-0.022	0.024
INSTAL	-0.15	-0.065	0.496	1.000	-0.100	-0.034
ATENDM	0.60	0.477	-0.022	-0.100	1.000	0.555
GLOBAL	0.60	0.810	0.024	-0.034	0.555	1.000



29

Cálculo das componentes principais

FGV EDUCAÇÃO EXECUTIVA

```
> bbscale=scale(bb.num)
> pca=prcomp(bbscale)
> tudo=cbind(bb, pca$x ) #matriz com  as variáveis originais e as CP
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4    PC5     PC6
Standard deviation     1.7012 1.2189 0.77396 0.70109 0.6099 0.39714
Proportion of Variance 0.4823 0.2476 0.09984 0.08192 0.0620 0.02629
Cumulative Proportion  0.4823 0.7300 0.82979 0.91171 0.9737 1.00000
```

Desvio padrão de cada CP

% da variância total explicada pelas k primeiras PC

30

Determinação do número m de CP

FGV EDUCAÇÃO EXECUTIVA

Em geral seleciona-se **m** tal que as variâncias das CP sejam

a) Superiores a 1,0 **ou**

b) Superiores a 0,75 (Standard deviation > 0,87)

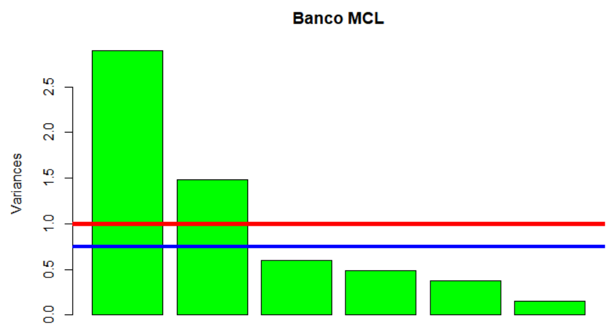
```
Importance of components:
                          PC1    PC2     PC3     PC4    PC5     PC6
Standard deviation     1.7012 1.2189 0.77396 0.70109 0.6099 0.39714
Proportion of Variance 0.4823 0.2476 0.09984 0.08192 0.0620 0.02629
Cumulative Proportion  0.4823 0.7300 0.82979 0.91171 0.9737 1.00000
```

Vamos selecionar m=2 componentes principais

31

Determinação do número m de CP

```
> plot(pca, col="green", main="Banco MCL")
> abline(h=1, col="red", lwd=5)
> abline(h=.75, col="blue", lwd=4)
```



32

Interpretação das CP

#as **correlações** entre as variáveis e os fatores são dadas por

#correlação (Zi, CPj) = wij * desvio padrao (CP)

#são chamadas **components loadings** ou **'cargas das componentes**"

#Vamos considerar apenas as 2 primeiras componentes principais

```
> cargas=cor(bb.num,pca$x[,1:2])
> round(cargas,3)
```

	PC1	PC2
LIMCRED	0.854	-0.0193
SATHOR	0.881	0.0913
PARKING	-0.133	0.8553
INSTAL	-0.188	0.8413
ATENDM	0.760	0.0631
GLOBAL	0.871	0.1839

Como podemos batizar cada componente principal?

Notas sobre interpretação

Cuidado com a interpretação das componentes principais

- Critério sugerido é baseado apenas no bom senso. Pode não funcionar.
- **Interpretação nem sempre é viável !**
- Com pequenas amostras, flutuações amostrais entre amostras da mesma população podem provocar diferentes interpretações.
- Quando a 1ª componente tem caráter de "desempenho global" as demais componentes é que podem revelar dimensões interessantes dos dados.
- As últimas CPs podem revelar informações importantes :
  - Se a Variância da CP for muito próxima de zero, significa que essa combinação linear é praticamente constante, ou seja podemos expressar uma variável como combinação linear das demais. Isto é importante para identificar multicolinearidade em regressão múltipla.
  - Everitt & Dunn sugerem que os gráficos das ultimas CPs podem auxiliar ao determinar possíveis outliers que só apareceriam nesses gráficos (pois "criariam" uma dimensão própria, fictícia na realidade)

34

Coeficientes w_{ji} das CP

Apenas a título de ilustração. Não utilizamos!

Coeficientes das componentes principais (lembre que as variáveis X foram padronizadas)

$$CP_j = w_{ji}\, z_1 + \ldots\ldots + w_{jp}\, z_p$$

Não é o melhor para interpretar as CP!!!!! (correlações são mais interessantes)

```
> print(pca$rotation[,1:2], digits=2)
           PC1    PC2
LIMCRED  0.502 -0.016
SATHOR   0.518  0.075
PARKING -0.078  0.702
INSTAL  -0.111  0.690
ATENDM   0.447  0.052
GLOBAL   0.512  0.151
```

35

Escores fatoriais

Já salvamos no arquivo “tudo”; são as coordenadas dos pontos nos eixos principais

```
> print(head(pca$x[,1:2],10), digits = 2)

# escores dos 10 primeiras observações
        PC1   PC2
 [1,]  0.85 -0.31
 [2,]  1.27 -1.52
 [3,]  1.77 -0.73
 [4,] -2.19  0.30
 [5,]  2.89  0.95
 [6,] -0.63 -0.32
 [7,]  2.58 -0.29
 [8,]  0.42 -0.78
 [9,]  2.11 -0.19
[10,]  3.01  0.31
```

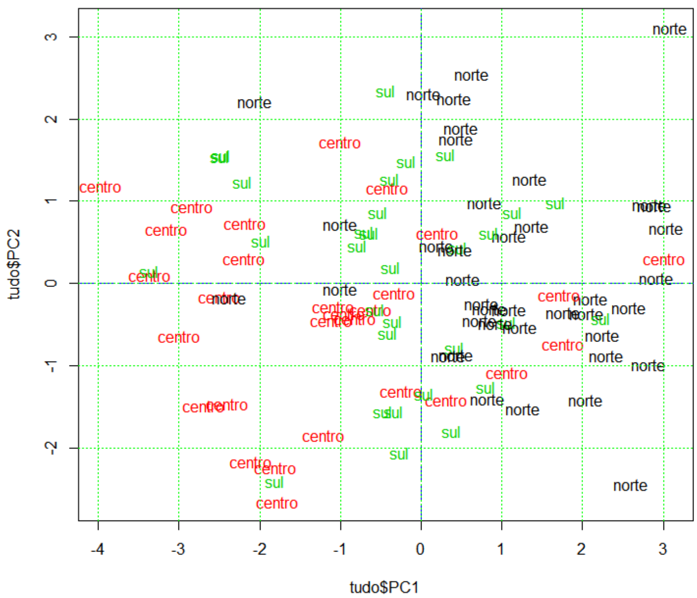
36

REPRESENTAÇÃO GRÁFICA

```
#representação gráfica
#criamos variável cor para melhorar representação gráfica

> cor=ifelse(bb$REGIAO=="norte",1,ifelse(bb$REGIAO=="centro",2,3))

> plot(tudo$PC1,tudo$PC2,type = "n")

> text(tudo$PC1,tudo$PC2, labels=tudo$REGIAO, col=cor)

> abline(h=0,v=0,col="blue",lty=5)

> grid(col="green")
```

Região da agência do cliente

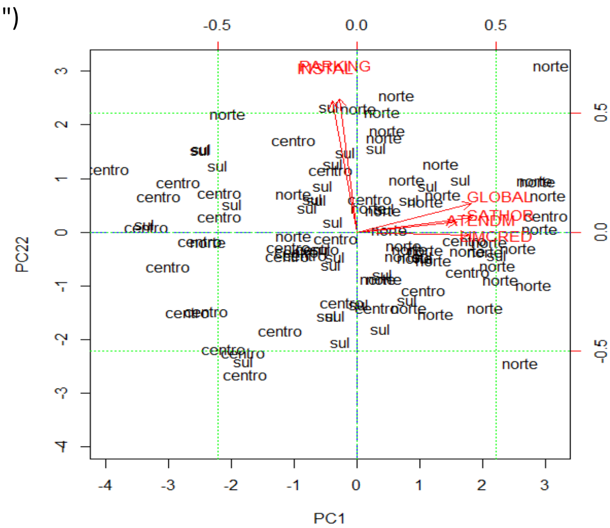
EDUCAÇÃO EXECUTIVA



Biplot

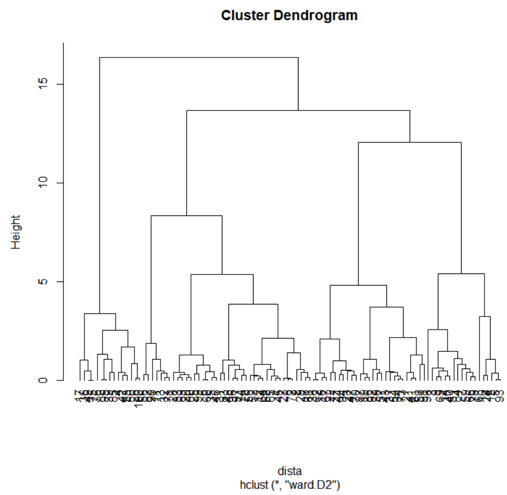
FGV EDUCAÇÃO EXECUTIVA

```
biplot(pca, xlab="PC1",ylab="PC2",xlabs= tudo$REGIAO, scale=0)
abline(h=0,v=0,col="blue",lty=5)
grid(col="green")
```



Agrupando com CP

```
> dista=dist(tudo[,9:10])
> grupo=hclust(dista, method = "ward.D2")
> plot(grupo, hang = -1)
> grupx=cutree(grupo,4)
```



Agrupamentos

```
> par(mfrow=c(1,2))
> boxplot(tudo$PC1 ~grupx, col=rainbow(5))
> boxplot(tudo$PC2~grupx, col=rainbow(5))
```