Análise de Componentes Principais

Dados: morfologia e sobrevivência de pardais após uma tempestade

Cláudia e Raquel

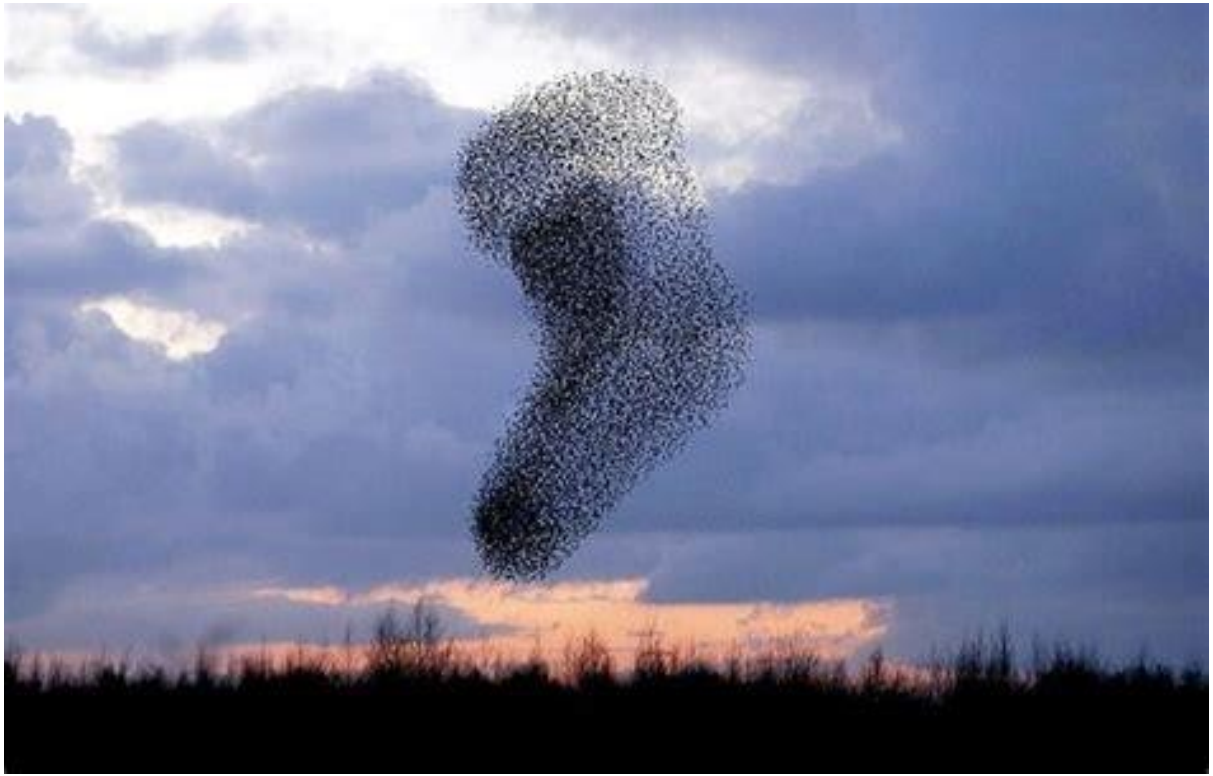
Sparrow = Pardal ou tico-tico?

Sparrow em diferentes ângulos

Tico-tico

pardal

ACP = representação de nuvem de pontos em dimensão reduzida



Aqui temos uma imagem 2D de uma nuvem 3D.

Que plano resulta na melhor foto?

Exemplo: 49 tico-ticos e 6 variáveis

variáveis
nas colunas

ID	sobreviveu	corpo	asa	cabeca	perna	peito
1	1	156	245	31.6	18.5	20.5
2	1	154	240	30.4	17.9	19.6
3	1	153	240	31.0	18.4	20.6
4	1	153	236	30.9	17.7	20.2
5	1	155	243	31.5	18.6	20.3
6	1	163	247	32.0	19.0	20.9
7	1	157	238	30.9	18.4	20.2
8	1	155	239	32.8	18.6	21.2
9	1	164	248	32.7	19.1	21.1
10	1	158	238	31.0	18.8	22.0
11	1	158	240	31.3	18.6	22.0
12	1	160	244	31.1	18.6	20.5
13	1	161	246	32.3	19.3	21.8
14	1	157	245	32.0	19.1	20.0
15	1	157	235	31.5	18.1	19.8

Indivíduos
nas linhas

Pergunta da pesquisa:
Houve alguma relação
Entre o
tamanho/formato
do pássaro e a
sobrevivencia após a
tempestade?

Matematicamente é
Uma matriz de tamanho
49x6.
1 variavel categórica e
5 numéricas

É uma matriz. Podemos estudar do ponto de vista das linhas ou colunas:

Estudo das linhas (dos indivíduos)

Que indivíduos são parecidos quando olhamos todas as variáveis morfológicas juntas?

Indivíduos que morreram são parecidos/diferentes dos que sobreviveram?

Existem grupos de indivíduos parecidos? Podemos identificar grupos de indivíduos parecidos?

ID	sobreviveu	corpo	asa	cabeca	perna	peito
1	1	156	245	31.6	18.5	20.5
2	1	154	240	30.4	17.9	19.6
3	1	153	240	31.0	18.4	20.6
4	1	153	236	30.9	17.7	20.2
5	1	155	243	31.5	18.6	20.3
6	1	163	247	32.0	19.0	20.9
7	1	157	238	30.9	18.4	20.2
8	1	155	239	32.8	18.6	21.2
9	1	164	248	32.7	19.1	21.1
10	1	158	238	31.0	18.8	22.0
11	1	158	240	31.3	18.6	22.0
12	1	160	244	31.1	18.6	20.5
13	1	161	246	32.3	19.3	21.8
14	1	157	245	32.0	19.1	20.0
15	1	157	235	31.5	18.1	19.8

Comparando colunas:

Comparacao de colunas

ID	sobreviveu	corpo	asa	cabeca	perna	peito
1	1	156	245	31.6	18.5	20.5
2	1	154	240	30.4	17.9	19.6
3	1	153	240	31.0	18.4	20.6
4	1	153	236	30.9	17.7	20.2
5	1	155	243	31.5	18.6	20.3
6	1	163	247	32.0	19.0	20.9
7	1	157	238	30.9	18.4	20.2
8	1	155	239	32.8	18.6	21.2
9	1	164	248	32.7	19.1	21.1
10	1	158	238	31.0	18.8	22.0
11	1	158	240	31.3	18.6	22.0
12	1	160	244	31.1	18.6	20.5
13	1	161	246	32.3	19.3	21.8
14	1	157	245	32.0	19.1	20.0
15	1	157	235	31.5	18.1	19.8

As variáveis são parecidas? Correlacionadas?

Isso é, elas são associadas de alguma forma?

Podemos identificar grupos de variáveis relacionadas entre si linearmente (correlacionadas)?

Se são muito relacionadas, podemos usar essa relação para reduzir o número de variáveis em um pno número de indicadores sintéticos?

Pra que reduzir? Para simplificar, ajudar a entender a fonte dessa correlação, para fazer análises que não gostam de variáveis correlacionadas.

Comparação é melhor com dados normalizados (padronizados), mesma escala, mesma unidade de medida.

ID	sobreviveu	corpo	asa	cabeca	perna	peito
1	1	-0.541719129	0.7248615	0.17718246	0.05424955	-0.32937165
2	1	-1.089022992	-0.2617555	-1.33272023	-1.00904159	-1.23720227
3	1	-1.362674923	-0.2617555	-0.57776889	-0.12296564	-0.22850158
4	1	-1.362674923	-1.0510492	-0.70359411	-1.36347197	-0.63198186
5	1	-0.815371061	0.3302147	0.05135723	0.23146474	-0.53111179
6	1	1.373844390	1.1195083	0.68048336	0.94032550	0.07410862
7	1	-0.268067198	-0.6564024	-0.70359411	-0.12296564	-0.63198186
8	1	-0.815371061	-0.4590790	1.68708515	0.23146474	0.37671883
9	1	1.647496321	1.3168318	1.56125993	1.11754069	0.27584876
10	1	0.005584733	-0.6564024	-0.57776889	0.58589512	1.18367938
11	1	0.005584733	-0.2617555	-0.20029321	0.23146474	1.18367938
12	1	0.552888596	0.5275381	-0.45194366	0.23146474	-0.32937165
13	1	0.826540527	0.9221849	1.05795903	1.47197107	0.98193924
14	1	-0.268067198	0.7248615	0.68048336	1.11754069	-0.83372199
15	1	-0.268067198	-1.2483726	0.05135723	-0.65461121	-1.03546213

Xpad = (X-media(X))/dp(X)

Interpretação:

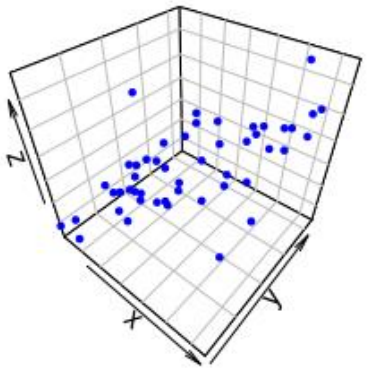
Valor acima de 0 : maior que a media
Valor abaixo de 0: abaixo da média
Valor = 1 -> 1 dp acima da média
Valor = -0.5 -> meio dp abaixo da média

Se a variável é normal,
valores > 2 ou < -2 são extremos, pois
Espera-se que 95% dos pontos estejam
Entre - 1,96 e 1,96.

A transformação não afeta a relação
Entre as variáveis nem entre os indivíduos,
Isso é, não afeta a forma da nuvem.

Nuvem multivariada de pardais: cada indivíduo é um ponto no espaço multivariado 3D

cabeça



corpo

asa

Nesse espaço, indivíduos próximos são indivíduos parecidos.
Indivíduos distantes são diferentes.

Proximidade ou distância pode ser medida como:

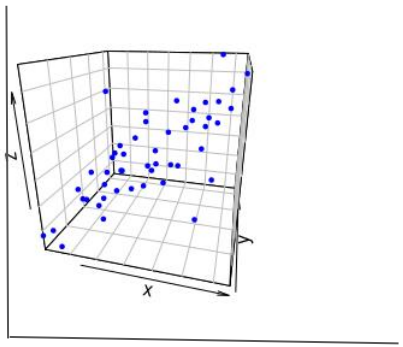
- Distância entre dois pontos i, i’
- Distância de cada um em relação ao centro da nuvem (centro da nuvem é o centróide = media multivariada)

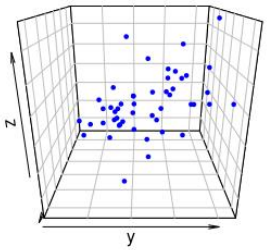
$$d^2(i,i') = \sum_{k=1}^{K} (x_{ik} - x_{i'k})^2$$

scatter3D(d$corpo, d$asa, d$cabeca)

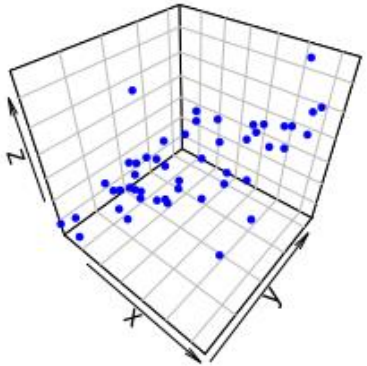
Nuvem multivariada de pardais: buscando o melhor ângulo para representar uma nuvem 3D numa página 2D

O objetivo da Análise de Componentes Principais É encontrar esse plano em que os pontos ficam melhor visualizados. Isso é, que estão mais dispersos entre si. Dispersão = Variância!

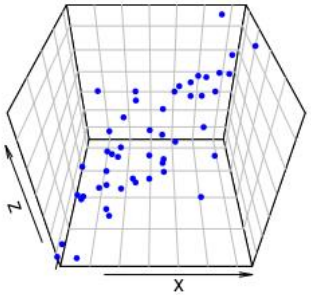


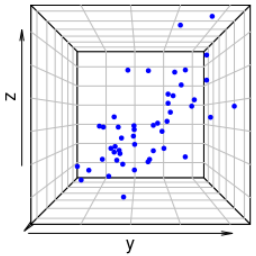


cabeça



corpo





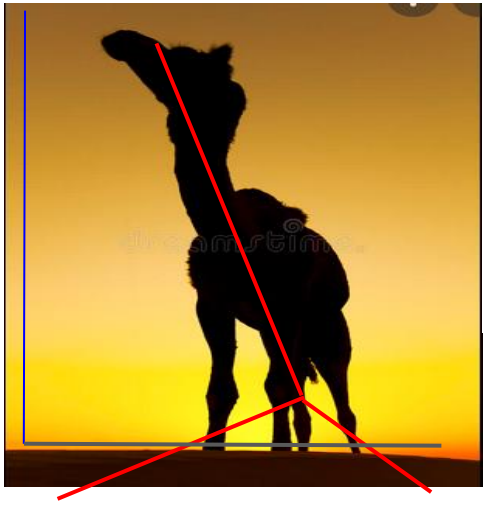
scatter3D(d$corpo, d$asa, d$cabeca, theta = 15, phi = 30)

A nuvem anterior foi convertida de 3D -> 2D. Como visualizar nossa nuvem 5D em 2D?

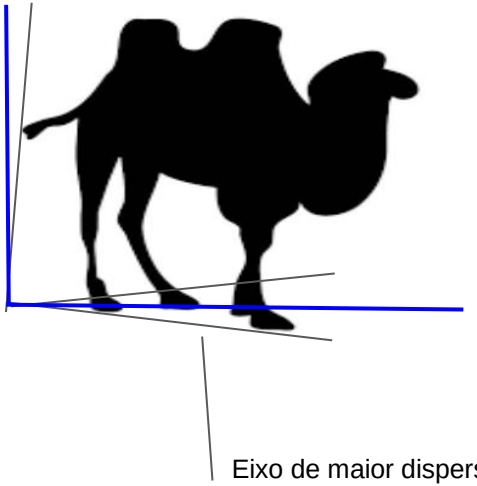
ID	sobreviveu	corpo	asa	cabeca	perna	peito
1	1	-0.541719129	0.7248615	0.17718246	0.05424955	-0.32937165
2	1	-1.089022992	-0.2617555	-1.33272023	-1.00904159	-1.23720227
3	1	-1.362674923	-0.2617555	-0.57776889	-0.12296564	-0.22850158
4	1	-1.362674923	-1.0510492	-0.70359411	-1.36347197	-0.63198186
5	1	-0.815371061	0.3302147	0.05135723	0.23146474	-0.53111179
6	1	1.373844390	1.1195083	0.68048336	0.94032550	0.07410862
7	1	-0.268067198	-0.6564024	-0.70359411	-0.12296564	-0.63198186
8	1	-0.815371061	-0.4590790	1.68708515	0.23146474	0.37671883
9	1	1.647496321	1.3168318	1.56125993	1.11754069	0.27584876
10	1	0.005584733	-0.6564024	-0.57776889	0.58589512	1.18367938
11	1	0.005584733	-0.2617555	-0.20029321	0.23146474	1.18367938
12	1	0.552888596	0.5275381	-0.45194366	0.23146474	-0.32937165
13	1	0.826540527	0.9221849	1.05795903	1.47197107	0.98193924
14	1	-0.268067198	0.7248615	0.68048336	1.11754069	-0.83372199
15	1	-0.268067198	-1.2483726	0.05135723	-0.65461121	-1.03546213

Análise de Componentes Principais

- Objetivo 1: estudo dos indivíduos
  - visualizar a forma da nuvem de pontos em um espaço reduzido (2D ou 3D) , buscando uma representação que seja fiel ao formato dos dados, que mostre quem é distante e quem é próximo.



ACP

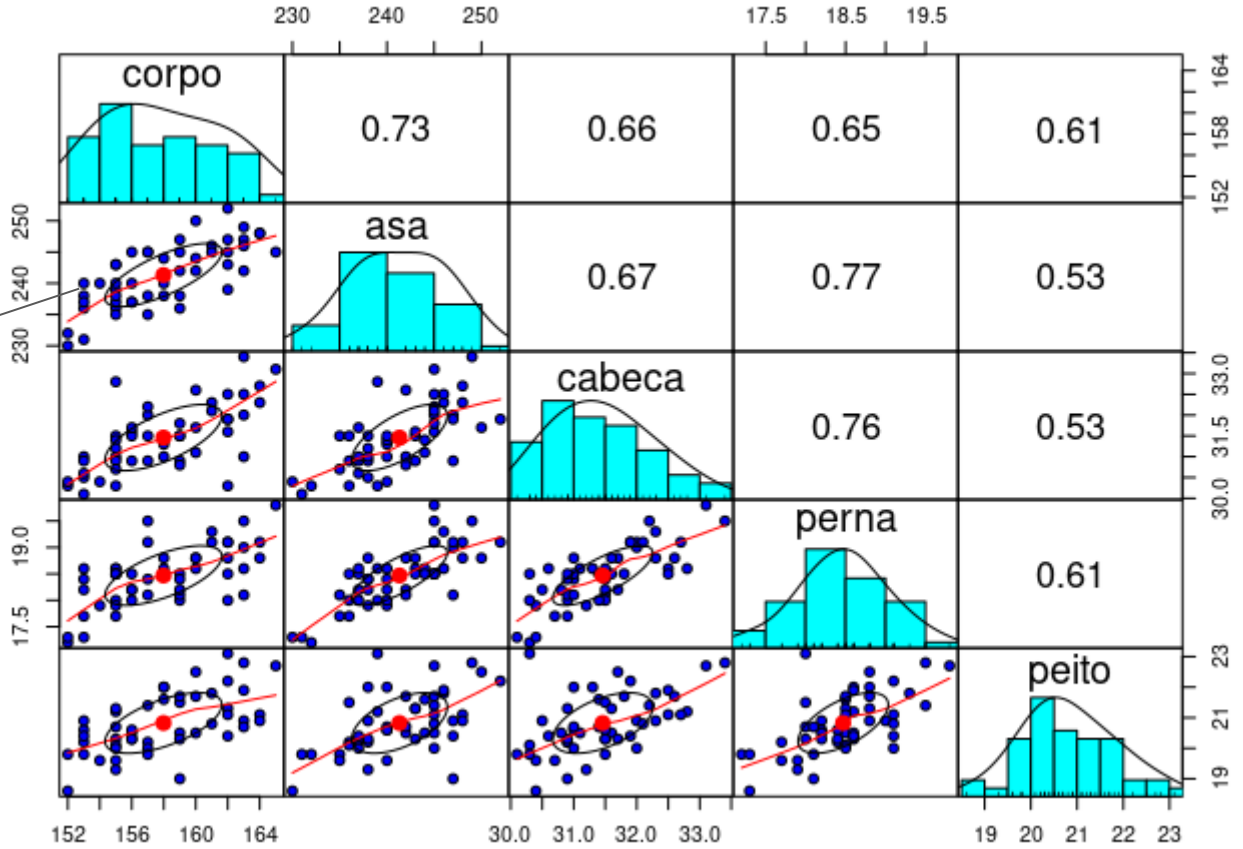


Eixo de maior dispersão dos pontos

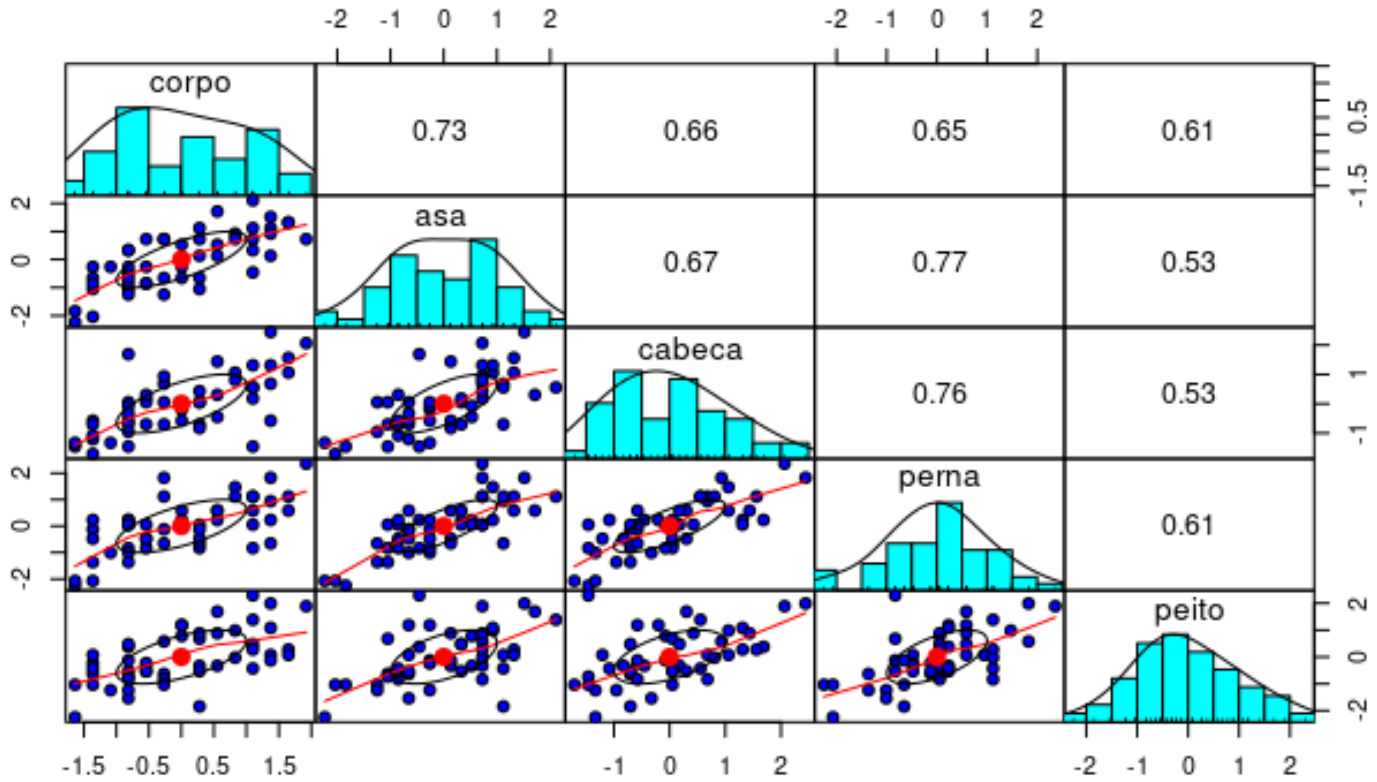
A equação de cada ponto da imagem bidimensional é dada por uma combinação das três variáveis: x,y,z originais.

Comparação da relação entre variáveis morfológicas duas a duas (correlação linear) : sugere um plano pode ser adequado para representar os pontos sem deformar a relação entre as variáveis.

Cada ponto corresponde aos valores do par de variáveis para um indivíduo



Correlação das variáveis padronizadas

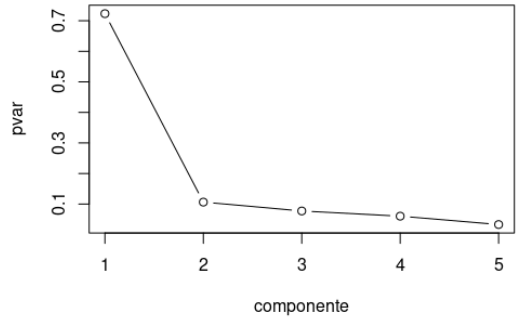


No R

```
## Biblioteca:
library(psych)
pc <- prcomp(d[,3:7],
             center = TRUE,
             scale. = TRUE)
```

```
> # o sdev é desvio padrao explicado pelo componente
> pc$sdev
[1] 1.9015726 0.7290433 0.6216306 0.5491498 0.4056199

# calculando a variancia explicada (na mão)
pvar = (pc$sdev^2/sum(pc$sdev^2))

> pvar
[1] 0.72319567 0.10630082 0.07728491 0.06031310 0.03290550
```

```
plot(pvar, type = "b", xlab = "componente") # screeplot
```



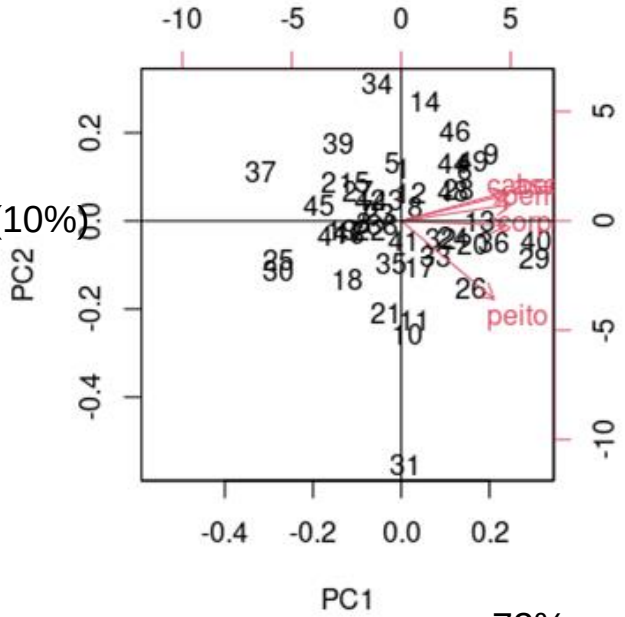
Output (posicao do individuo)

```
> pc$x # novos fatores: posicao de cada ponto nesse espaco
              PC1         PC2          PC3          PC4  -0
 [1,]  0.06428901  0.60083713 -0.171233350  0.515825561 -0
 [2,] -2.18031283  0.44230082  0.400069585  0.645459959 -0
 [3,] -1.14556567 -0.01925412 -0.676126877  0.716298164 -0
 [4,] -2.31106565 -0.17199267 -0.305962098 -0.149289289 -0
 [5,] -0.29504203  0.66520783 -0.474213811  0.545862110 -0
 [6,]  1.91626198  0.59525444  0.620933018 -0.006608669  0
 [7,] -1.05036763  0.11981084  0.074460843  0.088396192  0
 [8,]  0.43854156  0.16397253 -1.648440499 -0.815773999 -0
 [9,]  2.69147373  0.78226687  0.367947609 -0.464857885  0
[10,]  0.18568959 -1.31372223 -0.409088264  0.297345077  0
[11,]  0.37111481 -1.13843986 -0.300603973  0.147077931  0
[12,]  0.26770575  0.31473805  0.730471791  0.397776413  0
[13,]  2.35924685 -0.01109568 -0.376201520  0.155467751  0
[14,]  0.71464741  1.38874532 -0.557970698  0.473746184  0
[15,] -1.39425236  0.44297424 -0.179778374 -0.927861283  0
[16,] -1.55867849 -0.14474686  0.091608997 -0.172950915  0
```

Ficaremos com dois eixos, pois um espaço 2D já é suficiente para discriminar individuos parecidos e diferentes morfometricamente! Já dá conta de 82% da diferenca entre eles

Visualização 2D da nuvem morfométrica dos pardais, no plano de maior dispersão dos dados, obtidos pela ACP

(10%)



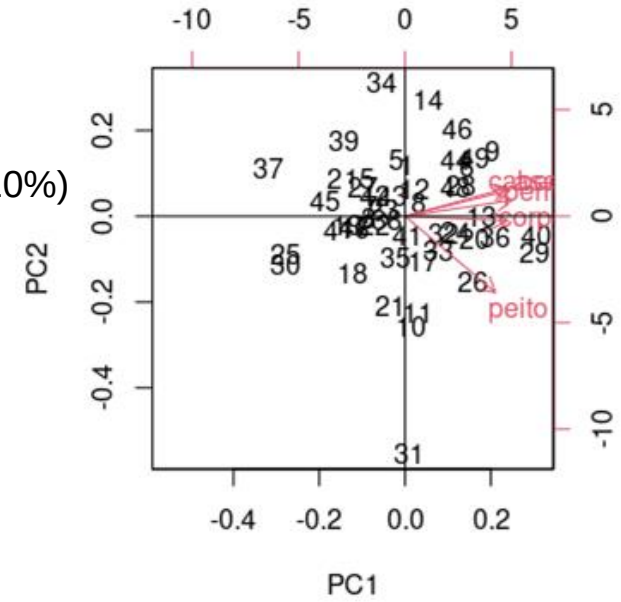
72% variancia

```
> d[c(29,40),]
   ID sobreviveu corpo asa cabeca perna peito
29 29          0   165 245   33.1  19.8  22.7
40 40          0   163 249   33.4  19.5  22.8

> d[c(25,30),]
   ID sobreviveu corpo asa cabeca perna peito
25 25          0   152 232   30.3  17.2  19.8
30 30          0   153 231   30.1  17.3  19.8

> d[c(31,34),]
   ID sobreviveu corpo asa cabeca perna peito
31 31          0   162 239   30.3  18.0  23.1
34 34          0   159 247   30.9  18.1  19.0
```

Distância no eixo PC1 implica em mais diferença do que no eixo 2.

Os eixos podem ser usados como indicadores sintéticos que melhor Separam os indivíduos.

```
> pc$rotation   = coeficientes da combinacao linear
              PC1         PC2         PC3         PC4         PC5
corpo   0.4517989 -0.05072137  0.6904702 -0.42041399  0.3739091
asa     0.4616809  0.29956355  0.3405484  0.54786307 -0.5300805
cabeca  0.4505416  0.32457242 -0.4544927 -0.60629605 -0.3427923
perna   0.4707389  0.18468403 -0.4109350  0.38827811  0.6516665
peito   0.3976754 -0.87648935 -0.1784558  0.06887199 -0.1924341
```



(10%)

72% variancia

Para cada individuo i, podemos calcular:

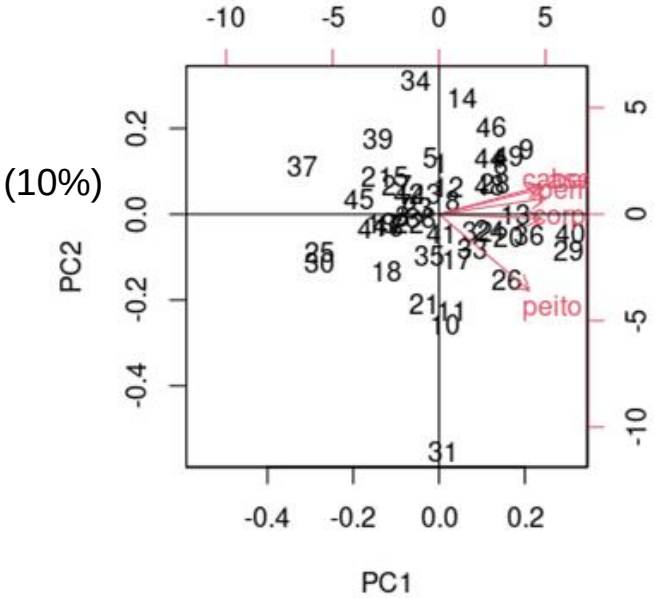
PC1(i) = 0.45 corpo(i) + 0.46 asa(i) + 0.45 cab (i) + 0.47 perna(i) + 0.39 peito(i)

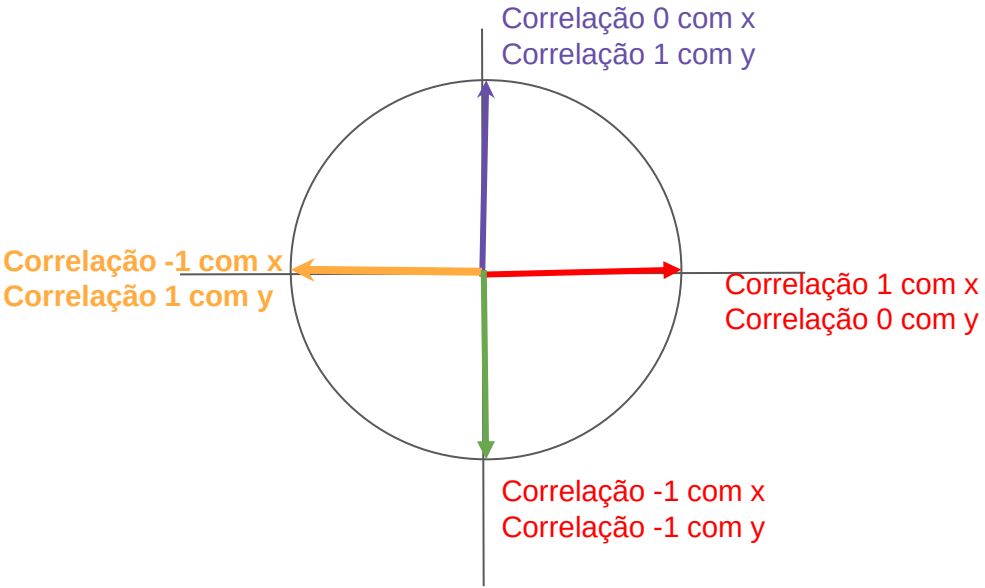
PC2(i) = -0.05 corpo (i) + 0.29 asa(i) + 0.32 cab (i) + 0.18 perna (i) - 0.87 peito (i)

As setas indicam a direção crescente das variáveis individuais

26 e 39 sao os mais diferentes em peito
25 e 9 sao os mais diferentes em cabeça

Tambem indicam a correlação de cada variavel com
Os componentes



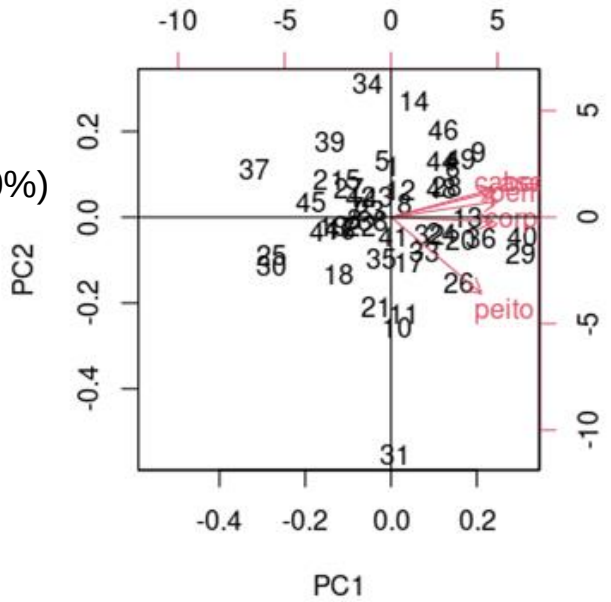


72% variancia

Interpretando os eixos

26 e 39 sao os mais diferentes em peito
25 e 9 sao os mais diferentes em cabeça

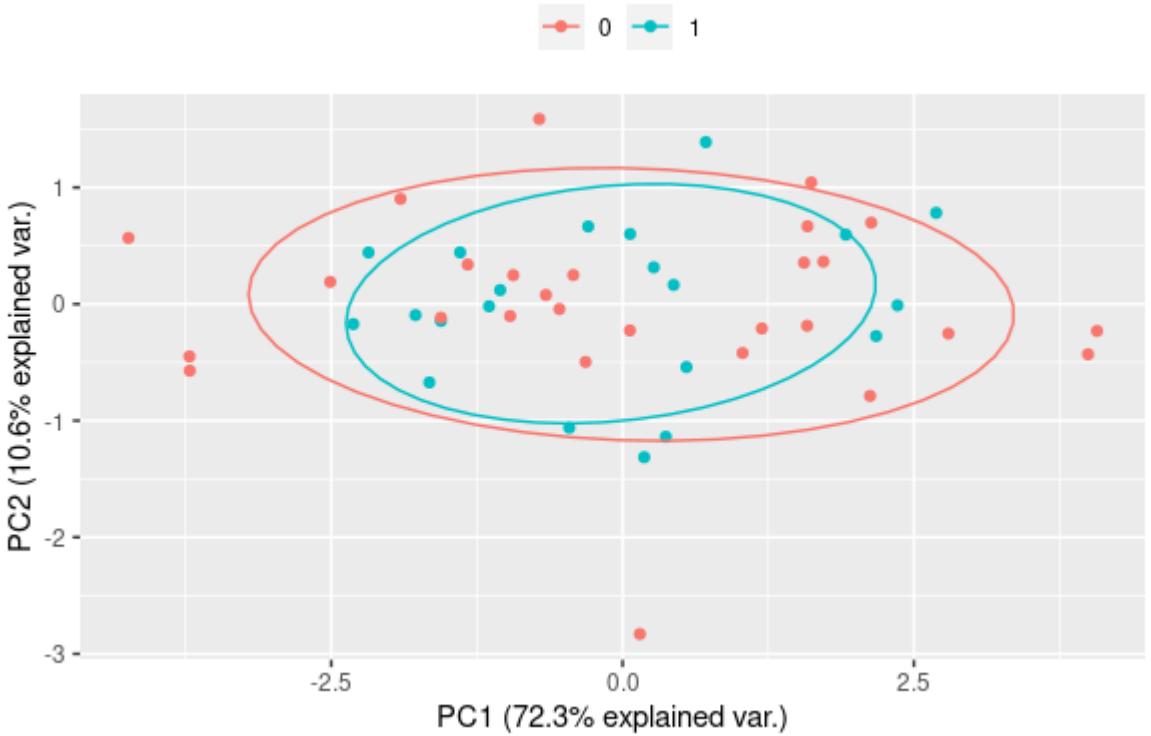
(10%)



72% variancia

PC1 = indicador de tamanho geral do pássaro
PC2 = indicador do formato do corpo (menos peitudo)

Pardais que morreram (0) e sobreviveram (1) na tempestade



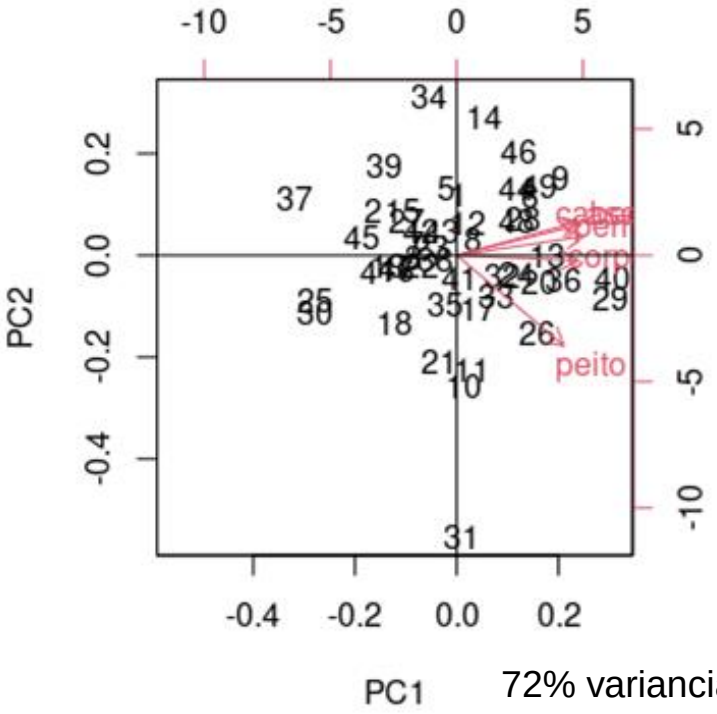
O autor tinha proposto que os pássaros de tamanho
“Fora do normal” teriam mais
Prob de morrer.

Assim, comprovando que
Existe uma seleção natural
Para pássaros de tamanho
Médio.

Os dados confirmam essa hipótese?

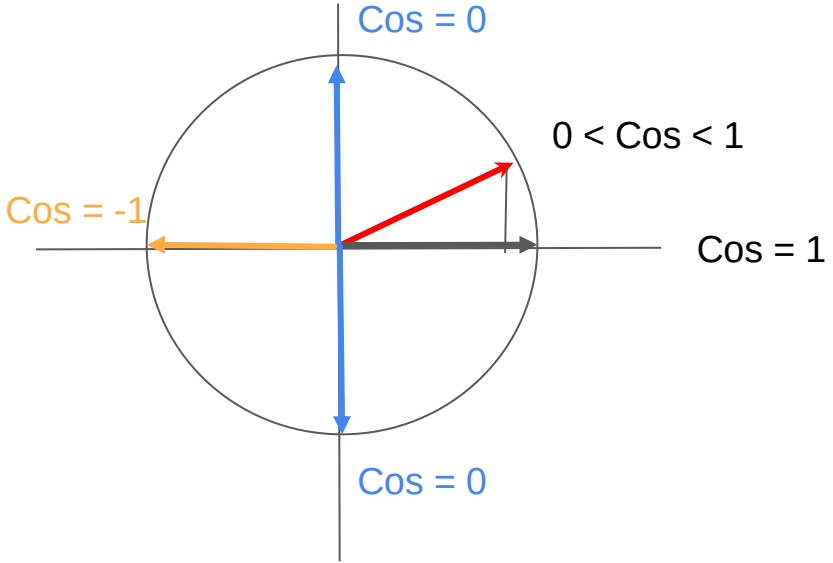
Análise de Componentes Principais

- Objetivo 2: estudo das variáveis



72% variancia

Cosseno = ângulo em relação ao eixo x



Cosseno = cateto adjacente/hipotenusa

Quanto mais próximo de 1, mais associada é a variável com o eixo x

Análise de Componentes Principais

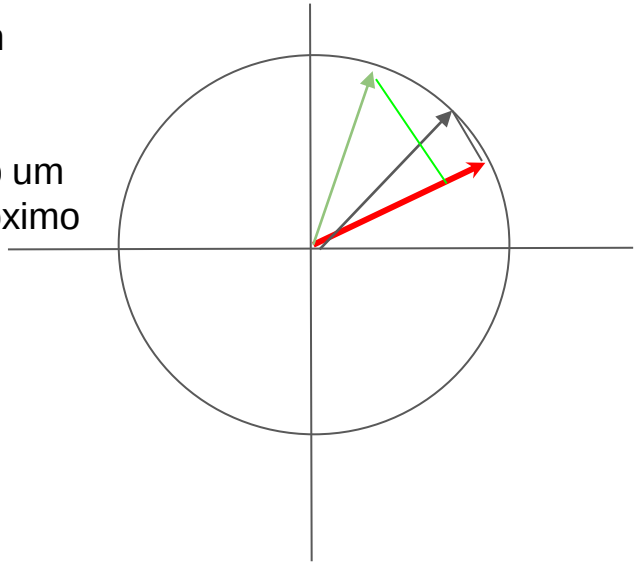
- Objetivo 2: estudo das variáveis

Cosseno = ângulo entre duas variáveis
Variáveis mais próximas **(correlacionadas)**, tem
Menor ângulo, e cosseno proximo de 1

Variáveis distantes **(não correlacionadas)** terão um maior angulo entre elas. E o cosseno menor, proximo de 0.

Logo o cosseno é uma forma geométrica de
Medir a correlação entre as variáveis no espaço multivariado,

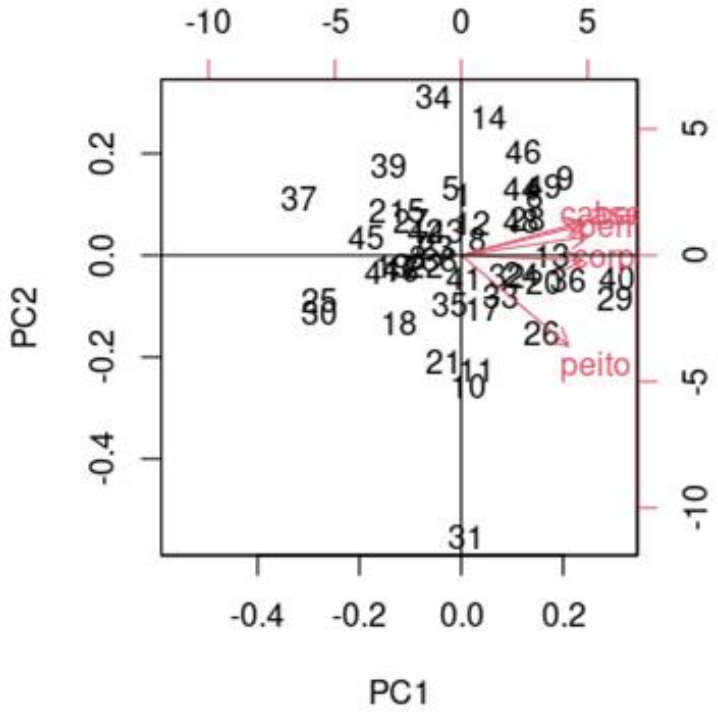
Quao maior o cosseno, melhor é
A representacao da variável pela componente.
Medida de qualidade!



Cosseno = cateto adjacente/hipotenusa

Quanto mais próximo de 1, mais associada é a variável com o eixo x

As setas são uma representação da matriz de correlação!



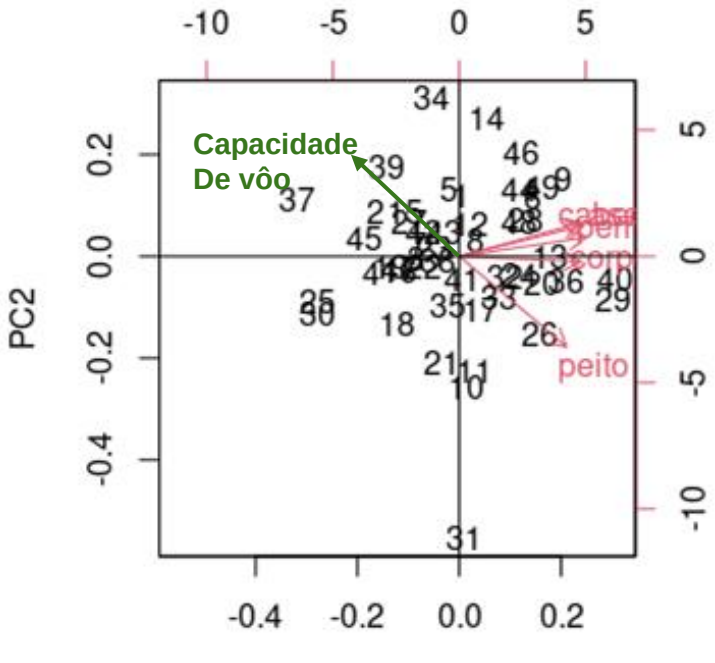
PS. o **tamanho da seta tbm** é
Informativo. Pois setas curtas indicam
Que as variáveis não estão bem represen-
tadas no gráfico (elas estão apontando para
Outra direção no espaço multidimensional).

Aqui, todas as setas tem o mesmo tamanho,
Entao elas estão igualmente representadas
Nesse espaço.

Numa outra situação, poderiamos **rotacionar**
O gráfico para favorecer outras
Variáveis!

Variáveis suplementares quantitativas

Podemos plotar nesse gráfico, variáveis que não foram usadas na análise
para ver como elas se distribuem nesse espaço.

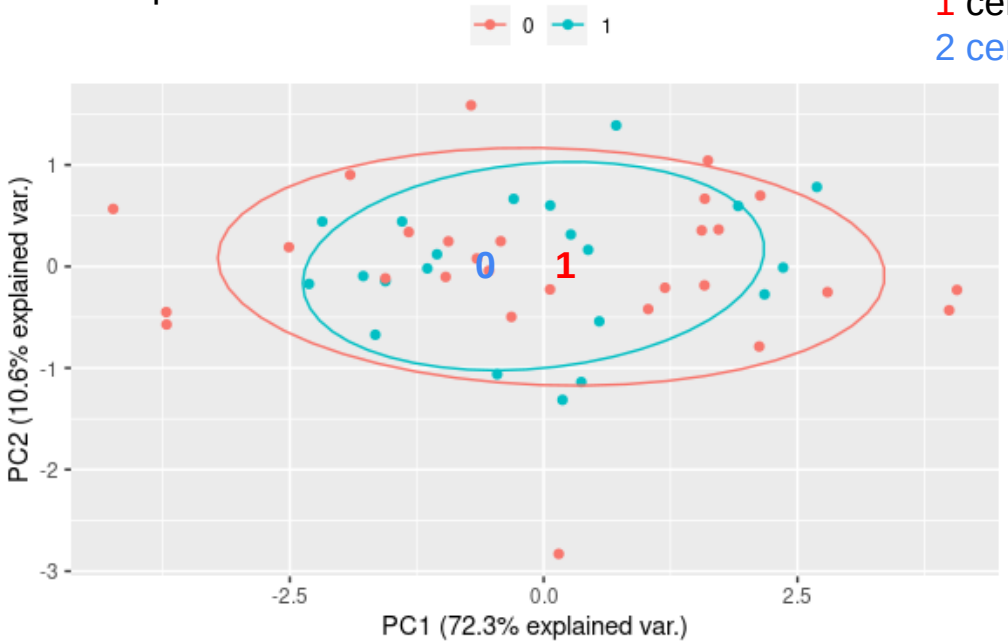


Exemplo: capacidade de vôo nao
Foi usado na ACP.
Aqui ela se mostra associada com pássaros
Menores e menos peitudos.

O cosseno do ângulo da nova seta é
a correlacao com cada eixo do PCA

Variáveis suplementares qualitativas

As variáveis qualitativas podem ser representadas pela cor dos pontos. O círculo mostra onde estão concentrados os pontos.



X é o centróide (centro Geométrico de cada nuvem de pontos)

1 centro da nuvem dos que morreram

2 centro nuvem dos que sobreviveram

Análise da Contribuição

Das variáveis: quanto maior a correlação com o eixo, maior sua contribuição para a construção desse.

Dos indivíduos: quanto maior o valor da coordenada do ponto, maior a contribuição para sua construção (mais longe o ponto está do centro da nuvem).

Pontos isolados e variáveis isoladas devem ser inspecionadas, podem ser outliers ou variáveis candidatas a serem vistas de forma univariada.

A contribuição de variáveis e de indivíduos é medida em termos de cosseno2, no fundo é o ângulo dos vetores ou dos pontos em relação ao eixo.