Análise de Componentes Principais

Adilson dos Anjos

Sensometria 2

Objetivo

- O objetivo dessa aula é apresentar a Análise de Componentes Principais.
- As análises serão realizadas com uso do R;

Pacotes utilizados nessa aula

- readxl
- FactoMineR
- bpca

Dados

- Café

Análise de Componentes Principais (ACP): Introdução

- PCA: Principal Component Analysis (Pearson, 1901)
- Objetivo: explicar a estrutura de variância/covariância por meio de combinações lineares das variáveis originais;
- As combinações lineares são chamadas de Componentes Principais;
- Essas combinações são não correlacionadas;

ACP

- Busca-se uma redução do número de p variáveis para k componentes principais;
- Em geral, a análise de componentes principais serve como um método intermediário de avaliação;
- Por exemplo: construção de agrupamentos em análise de segmentação (cluster).

ACP

- Na análise de componentes principais, busca-se a informação contida em p variáveis por meio de k componentes principais não correlacionadas ($k < p$);
- A qualidade da informação pode ser medida por meio da proporção de variância total explicada pelas k componentes principais;
- A suposição de normalidade não é necessária para utilização de ACP.

ACP

- Para a obtenção dos componentes principais utiliza-se a matriz de covariâncias dos vetores aleatórios das variáveis originais;
- É comum utilizar-se uma transformação das variáveis;
- Os componentes principais podem ser obtidos a partir da matriz de covariância das variáveis originais padronizadas ou,
- . . . de maneira equivalente, a partir da matriz de correlação das variáveis originais
- Os componentes principais dependem apenas da matriz de covariância (ou correlação);

ACP

- Dado o vetor $X' = [X_1, X_2, \ldots, X_p]$ e a matriz de covariâncias Σ com autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$;
- As combinações lineares são dadas da forma:

$$\begin{aligned}
Y_1 = l_1'X &= l_{11}X_1 + l_{21}X_2 + \ldots + l_{p1}X_p \\
Y_2 = l_2'X &= l_{12}X_1 + l_{22}X_2 + \ldots + l_{p2}X_p \\
\vdots\vdots & \qquad\qquad\qquad\qquad \vdots \\
Y_p = l_p'X &= l_{1p}X_1 + l_{2p}X_2 + \ldots + l_{pp}X_p
\end{aligned}$$

- A primeira combinação linear é a que possui a maior variância.

ACP

- A variância total é:
  $\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \lambda_1 + \lambda_2 + \ldots + \lambda_p$
- Assim, a proporção da variância total explicada pela k-ésima componente principal é

$$\text{Var devida a k-ésima CP} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}$$

ACP

- Se Σ é a matriz de covariância associada com o vetor $X' = [X_1, X_2, \ldots, X_p]$ e Σ tem os pares de autovalores e autovetores (λ_p, e_p) com $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, o i-ésimo componente principal é dado por:

$$Y_i = e_{1i}X_1 + e_{2i}X_2 + \ldots + e_{pi}X_p$$

- Com a condição:

$$Var(Y_i) = e_i'\Sigma e_i = \lambda_i \qquad e \qquad Cov(Y_i, Y_k) = e_i'\Sigma e_k = 0 (i \neq k)$$

ACP

- Em geral, entre 80 e 90% da variabilidade pode ser explicada por até 3 componentes principais (considerando um grande número de variáveis originais);
- A variabilidade depende do fenômeno em estudo;
- Em alguns casos, pode-se considerar uma variabilidade menor, em geral, em torno de 70%.

ACP

- Há diferenças entre os componentes principais obtidos a partir da matriz de variância e covariância em comparação com a matriz de correlação;
- Recomenda-se a padronização quando as variáveis possuem escalas diferentes (inclusive em magnitude)

ACP

Representação Gráfica

- Utiliza-se um gráfico de dispersão entre as duas componentes principais e o escores de cada componente;

ACP

ACP Definição dos Componentes

- Percentual: depende do fenômeno (scree-plot);
- Interpretação prática do componente;
- Aproximação da matriz de variâncias/covariâncias (original).

ACP

- Quando as variáveis possuem Distribuição Normal (normal multivariada) podem ser realizadas algumas inferências;
- Exemplo: elipses de confiança.

Exemplo: coxinha

```
library(FactoMineR)
```

- Nesse exemplo são considerados quatro atributos: sabor, aroma, massa e recheio.
- Cada avaliador atribuiu uma nota na escala ordinal de 1 a 5. Notas maiores estão relacionadas com melhor qualidade da coxinha.

```
sabor<-c(2.75,3.90,3.12,4.58,3.97,3.01,4.19,3.82)
aroma<-c(4.03,4.12,3.97,4.86,4.34,3.98,4.65,4.12)
massa<-c(2.80,3.40,3.62,4.34,4.28,2.90,4.52,3.62)
recheio<-c(2.62,3.52,3.05,4.82,4.98,2.82,4.77,3.71)
produto<-c(paste( 'C' ,1:8,sep=""))
coxa<-data.frame(sabor,aroma,massa,recheio,produto)
```

- Alguma observação sobre esses valores? sobre as coxinhas?

```
coxa
```

```
  sabor aroma massa recheio produto
1  2.75  4.03  2.80    2.62      C1
2  3.90  4.12  3.40    3.52      C2
3  3.12  3.97  3.62    3.05      C3
4  4.58  4.86  4.34    4.82      C4
5  3.97  4.34  4.28    4.98      C5
6  3.01  3.98  2.90    2.82      C6
7  4.19  4.65  4.52    4.77      C7
8  3.82  4.12  3.62    3.71      C8
```

- Um resumo dos dados:

```
summary(coxa)
```

```
     sabor           aroma           massa           recheio
 Min.   :2.750   Min.   :3.970   Min.   :2.800   Min.   :2.
 1st Qu.:3.092   1st Qu.:4.018   1st Qu.:3.275   1st Qu.:2.
 Median :3.860   Median :4.120   Median :3.620   Median :3.
 Mean   :3.667   Mean   :4.259   Mean   :3.685   Mean   :3.
 3rd Qu.:4.025   3rd Qu.:4.418   3rd Qu.:4.295   3rd Qu.:4.
 Max.   :4.580   Max.   :4.860   Max.   :4.520   Max.   :4.

    produto
 C1      :1
 C2      :1
 C3      :1
 C4      :1
 C5      :1
 C6      :1
```

- Obtendo os componentes principais:

```
prcomp(coxa[,1:4]) # base
```

```
Standard deviations (1, .., p=4):
[1] 1.3178933 0.2547255 0.1670736 0.1499212

Rotation (n x k) = (4 x 4):
                PC1         PC2         PC3          PC4
sabor   0.4557692  0.8160940 -0.1121458 -0.33717690
aroma   0.2234675  0.2149422 -0.2691445  0.91182420
massa   0.4770120 -0.4564207 -0.7177864 -0.22118397
recheio 0.7174930 -0.2819051  0.6322715  0.07724004
```

A função **prcomp** está no base do R. Não é necessário instalar pacotes.

A função **estim_ncp** sugere o número de dimensões:

```
estim_ncp(coxa[,1:4]) # factominer
```

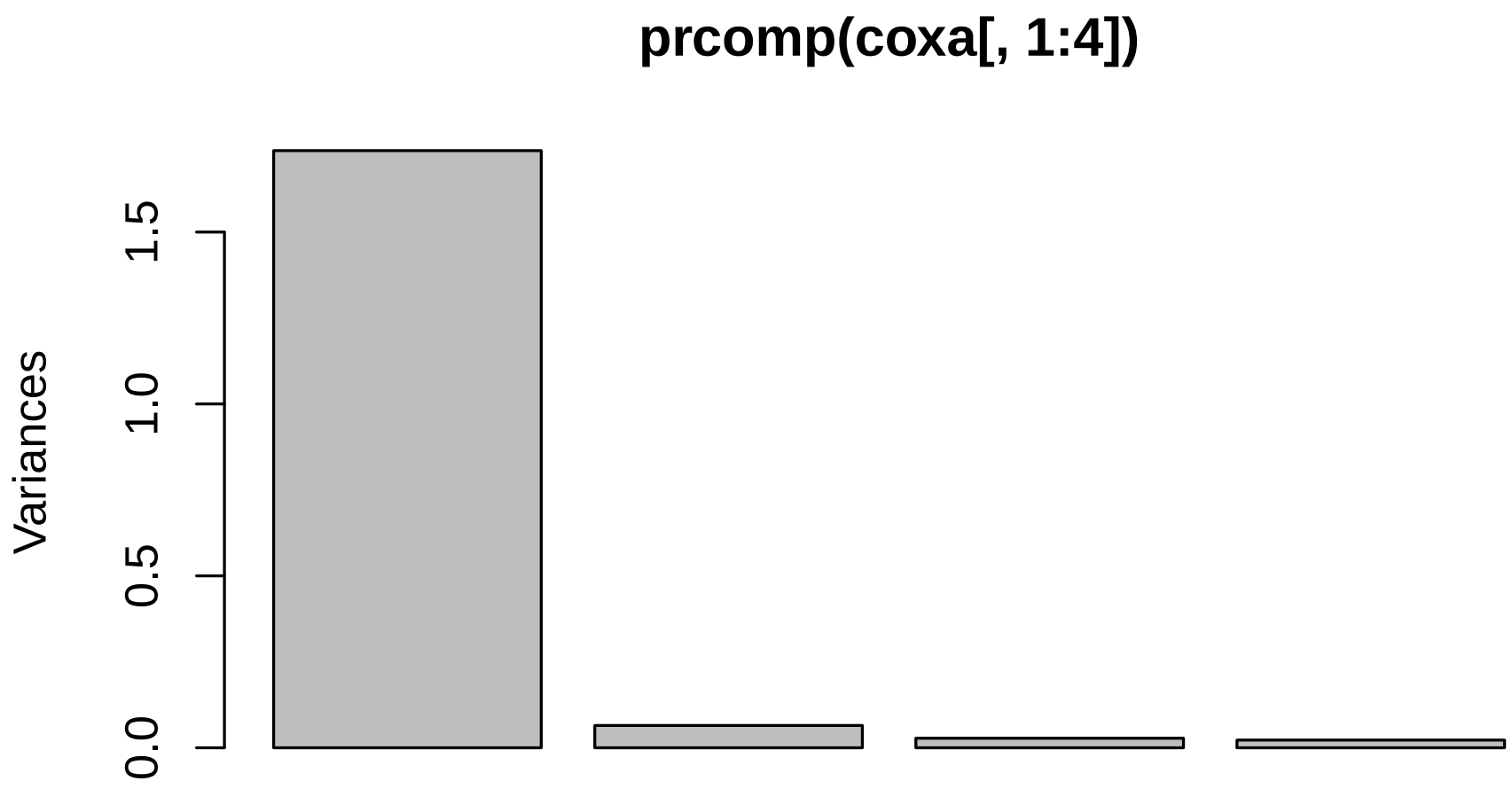
```
$ncp
[1] 1

$criterion
[1] 1.0000000 0.2556116 0.3908284 0.6169136
```

Nesse caso, é sugerida apenas uma dimensão.

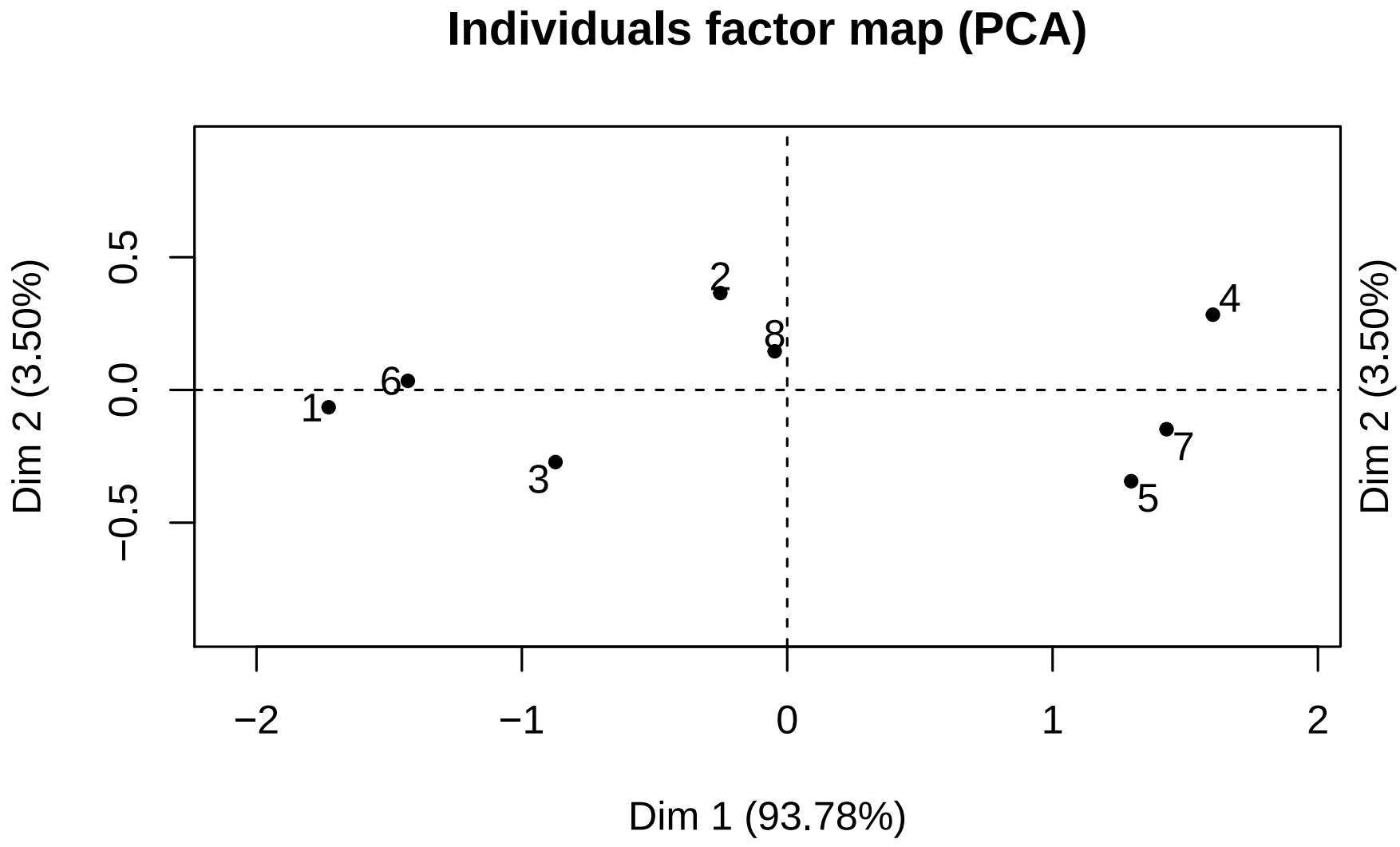
-Graficamente, pode-se obter o mesmo resultado:

```
screeplot(prcomp(coxa[,1:4])) # base
```



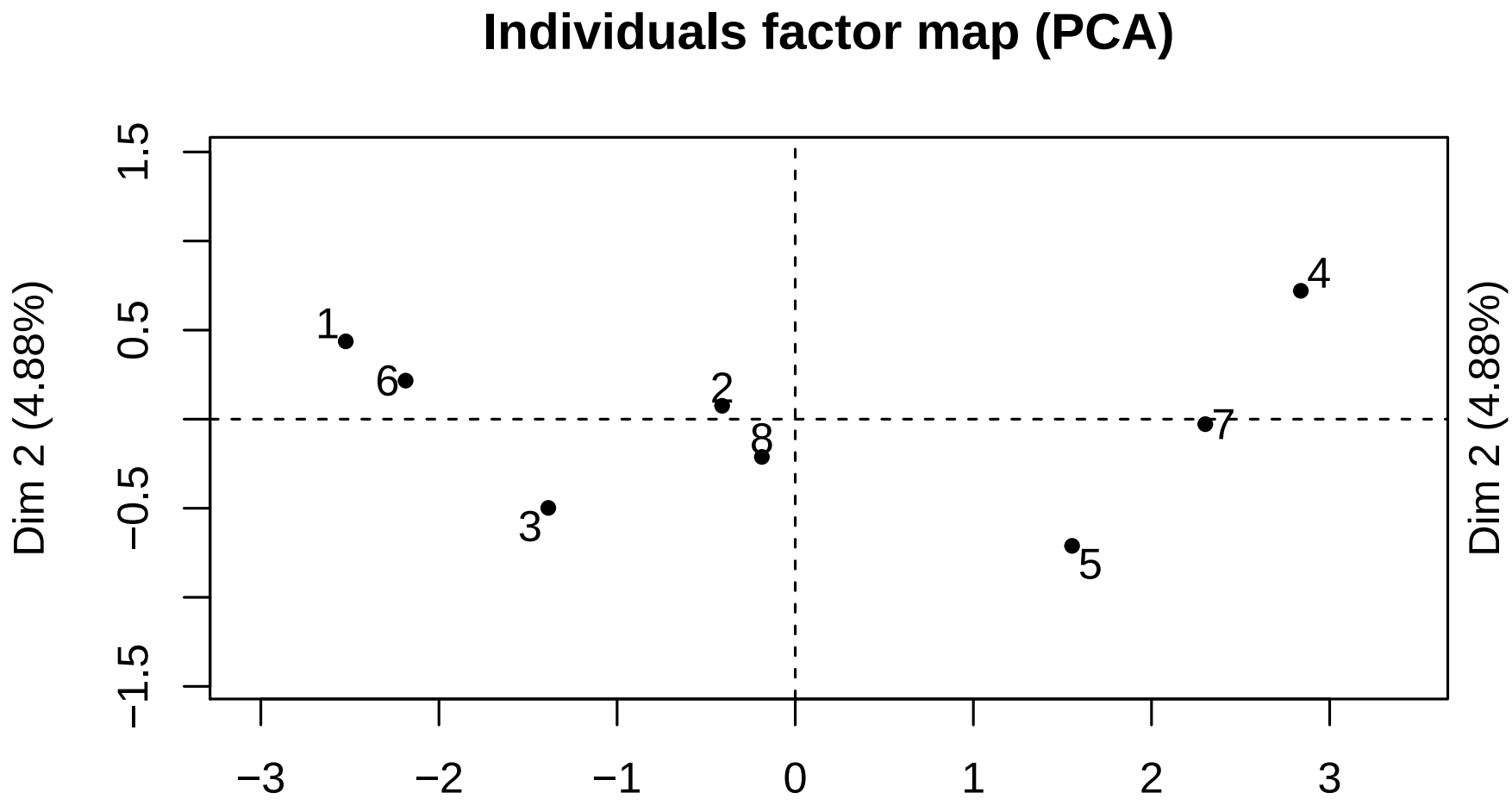
- Obtendo as componentes principais:
- Nesse primeiro caso, utilizamos a opção scale=F. Isso significa que estamos utilizando a escala natural do atributo, sem nenhuma transformação.

```
coxa.pca<-PCA(coxa[,1:4],scale=F)
```



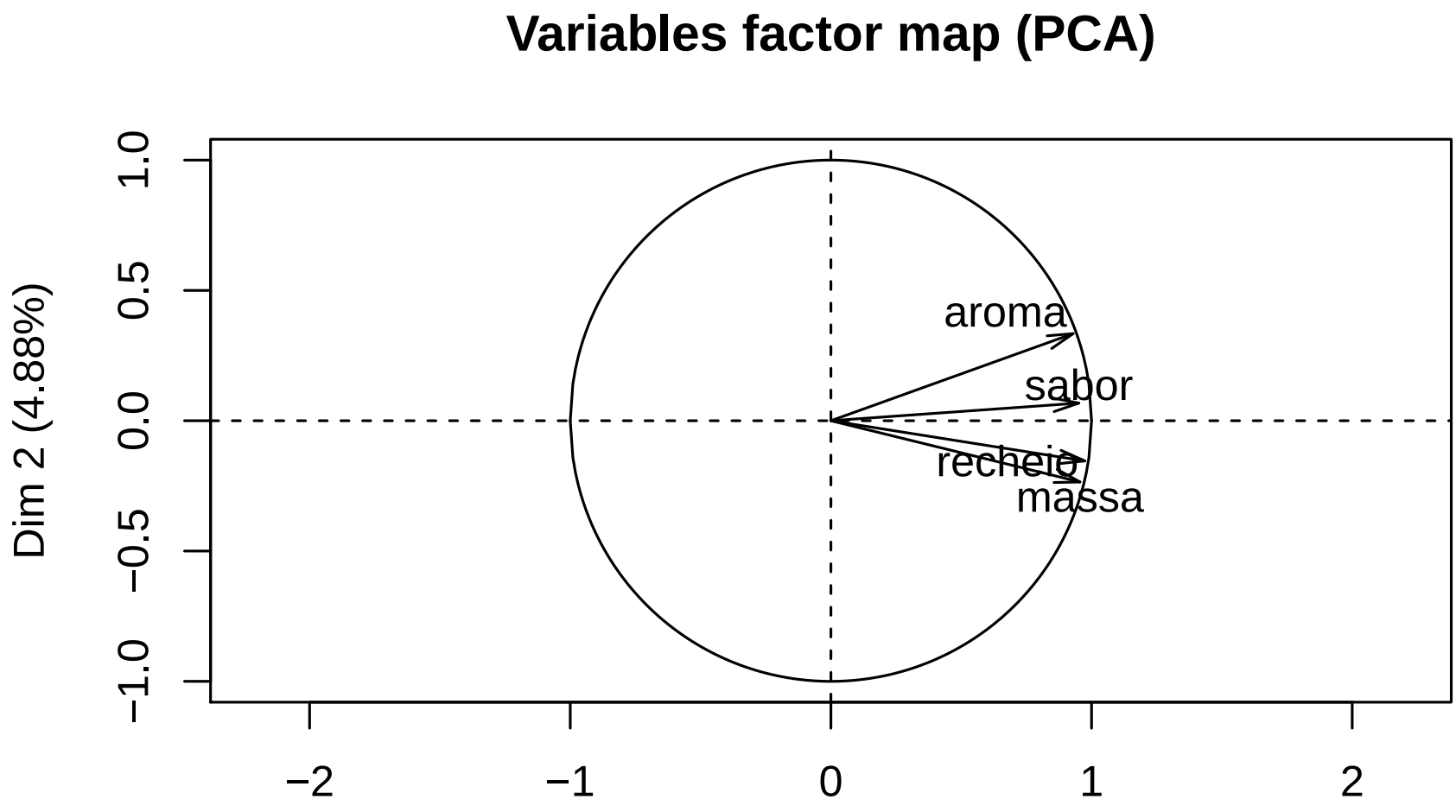
- Nesse segundo caso, utilizamos a opção scale=T. Isso significa que estamos utilizando a escala natural do atributo, ou seja, fazendo uma transformação.

```
coxa.pca<-PCA(coxa[,1:4],scale=T)
```



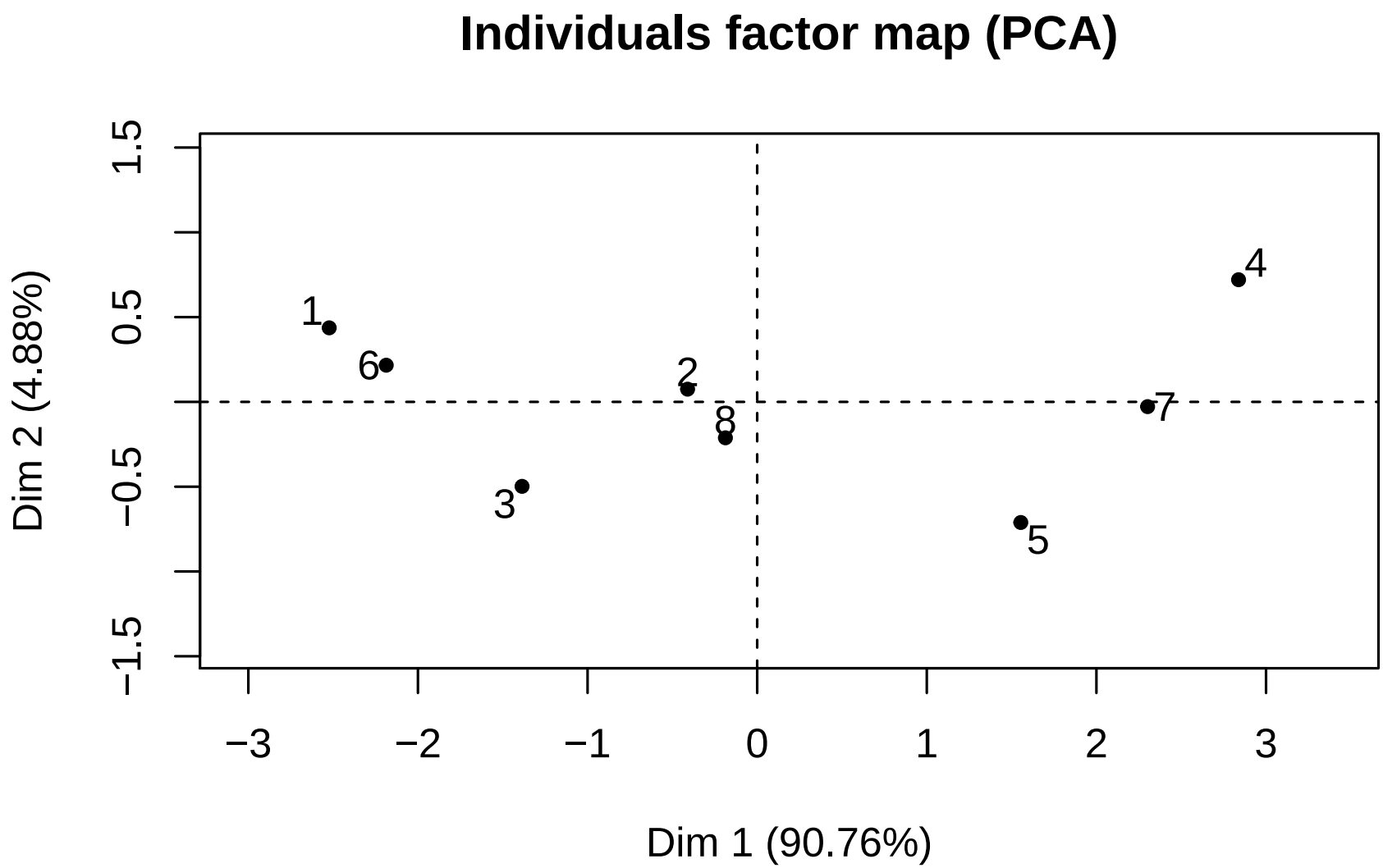
- Graficamente pode-se visualizar o comportamento dos atributos:

```
plot(coxa.pca,choix='var')
```



- E dos tipos de coxinhas:

```
plot(coxa.pca,choix='ind')
```



- O percentual explicado por cada componente pode ser obtido da seguinte forma:

```
coxa.pca$eig
```

```
       eigenvalue percentage of variance cumulative percent
comp 1 3.63027606                 90.756901
comp 2 0.19524697                  4.881174
comp 3 0.13041172                  3.260293
comp 4 0.04406525                  1.101631
```

- Os coeficientes da primeira componente principal podem ser obtidos da seguinte maneira:

```
coxa.pcapr<-prcomp(coxa[,1:4])
names(coxa.pcapr)
```

```
[1] "sdev"     "rotation" "center"   "scale"    "x"
```

```
coxa.pcapr$rotation[,1]
```

```
    sabor     aroma     massa   recheio
0.4557692 0.2234675 0.4770120 0.7174930
```

Onde usar essa informação?

- Para obter os escores para cada tipo de coxinha:

```
apply(coxa.pcapr$rotation[,1]*coxa[1,1:4],1,sum)
```

```
       1
5.369405
```

ou de outra maneira:

```
y1<-0.456*2.75+0.223*4.03+0.477*2.8+0.717*2.62
y1
```

```
[1] 5.36683
```

- Utilizando o critério da componente principal, qual a melhor coxinha?

Exemplo: Café

Arquivo de dados

```
library(readxl)
cafe<-read_excel('cafe.xls')
```

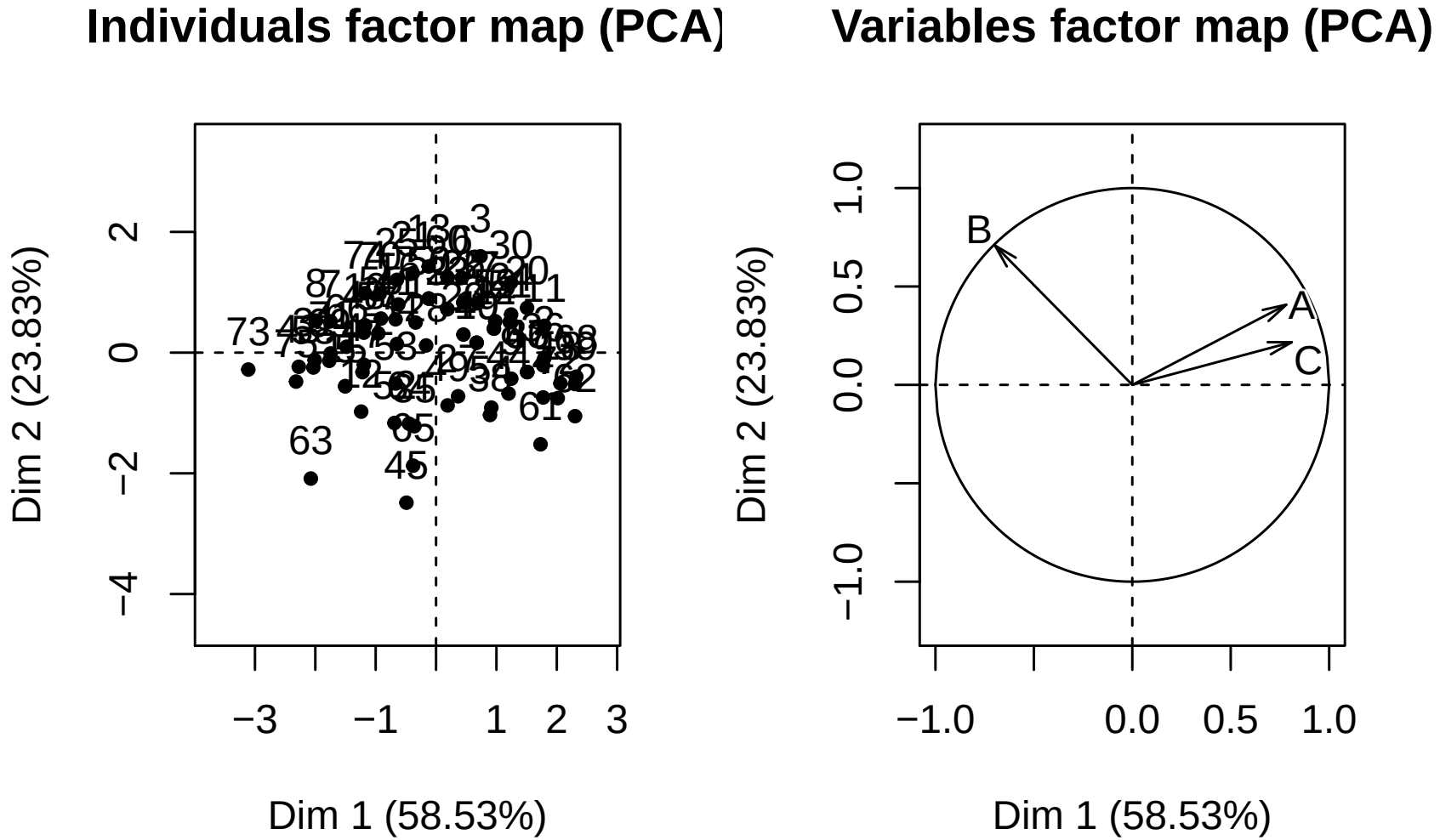
```
head(cafe)
```

```
# A tibble: 6 x 4
  Consumidor     A     B     C
  <chr>      <dbl> <dbl> <dbl>
1 C1             8     7     3
2 C2             9     4     7
3 C3             9     8     7
4 C4             8     3     7
5 C5             3     7     4
6 C6             7     8     8
```

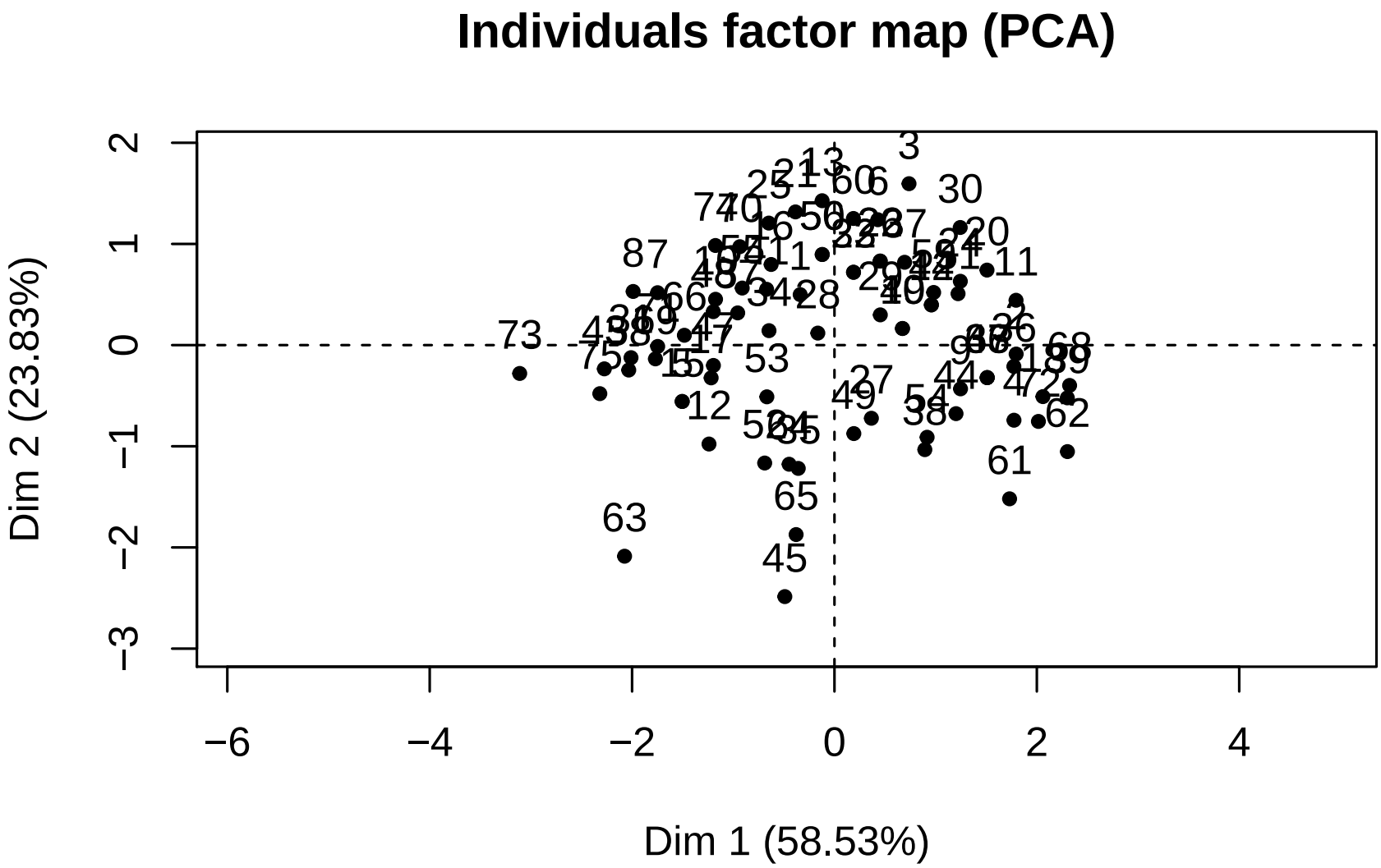
```
summary(cafe)
```

```
  Consumidor                A                B                C
 Length:75          Min.   :1.000   Min.   :2.000   Min.
 Class :character   1st Qu.:4.000   1st Qu.:4.000   1st Qu.
 Mode  :character   Median :6.000   Median :7.000   Median
                    Mean   :5.973   Mean   :6.227   Mean
                    3rd Qu.:8.000   3rd Qu.:8.000   3rd Qu.
                    Max.   :9.000   Max.   :9.000   Max.
```

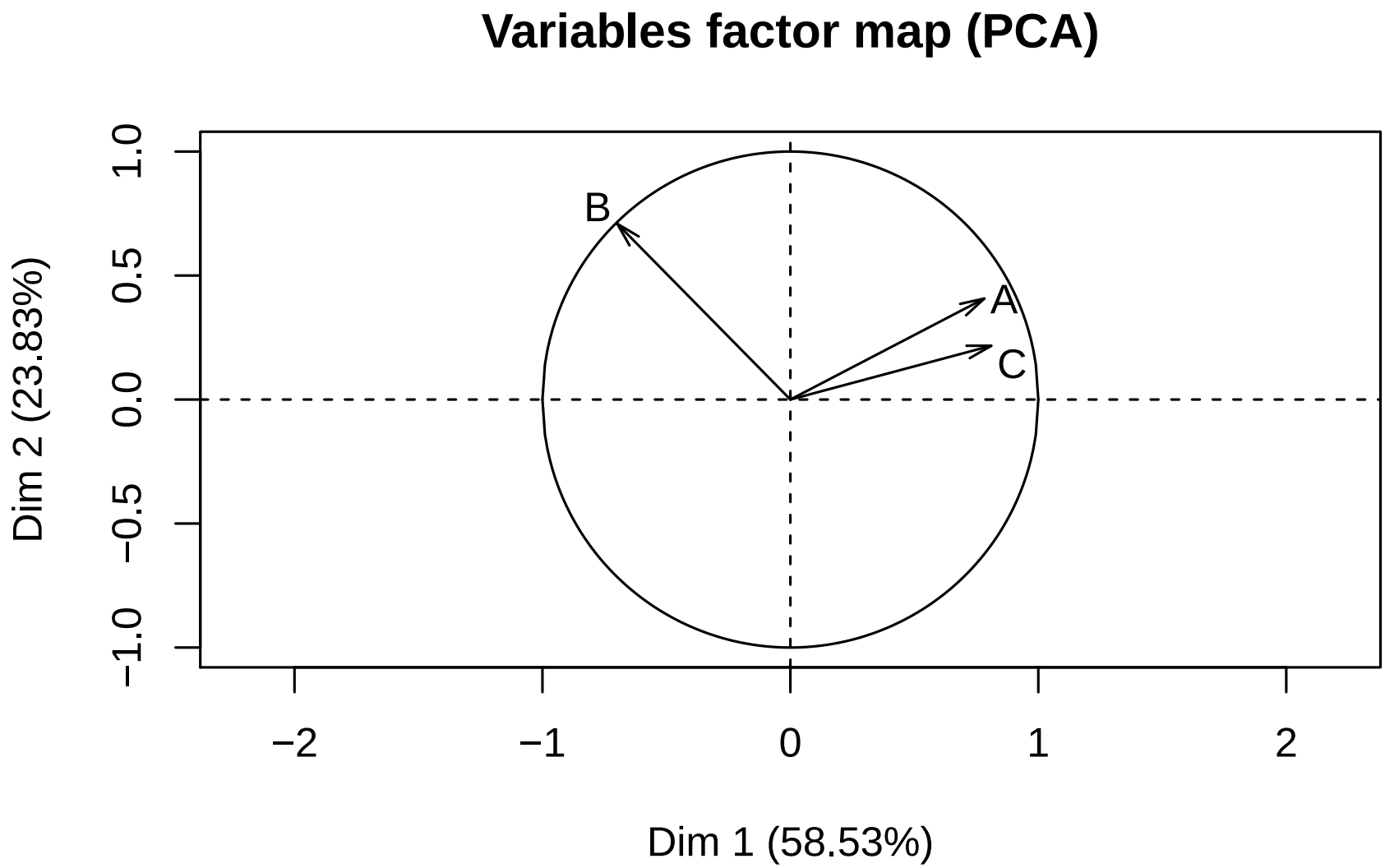
```
library(FactoMineR)
par(mfrow=c(1,2))
cafe.pca<-PCA(cafe[,2:4])
```



```
par(mfrow=c(1,1))
plot(cafe.pca,choix=c('ind')) # linhas
```



```
par(mfrow=c(1,1))
plot(cafe.pca,choix=c('var')) # colunas
```



```
head(round(cafe.pca$ind$contrib[,1:2],2))
```

```
  Dim.1 Dim.2
1  0.09  0.46
2  2.45  0.01
3  0.41  4.75
4  2.39  1.03
5  1.72  0.58
6  0.14  2.87
```

```
round(cafe.pca$var$contrib[,1:2],2)
```

```
  Dim.1 Dim.2
A 34.87 23.22
B 27.75 70.20
C 37.38  6.58
```

```
round(cafe.pca$var$coord[,1:2],2)
```

```
  Dim.1 Dim.2
A  0.78  0.41
B -0.70  0.71
C  0.81  0.22
```

```
round(cafe.pca$eig,2)
```

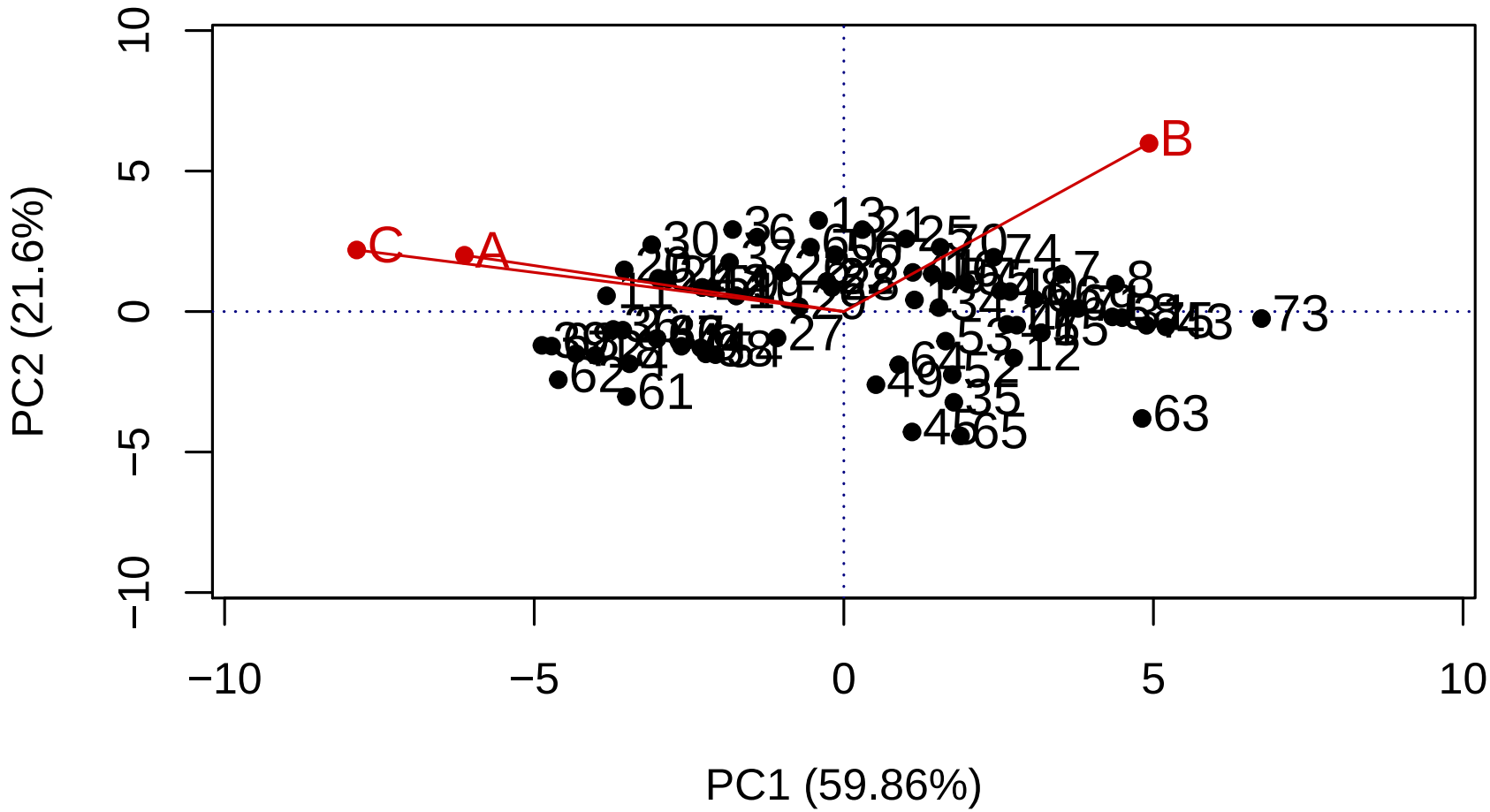
```
       eigenvalue percentage of variance cumulative percent
comp 1       1.76                  58.53
comp 2       0.71                  23.83
comp 3       0.53                  17.64
```

Utilizando o pacote bpca

```
library(bpca)
```

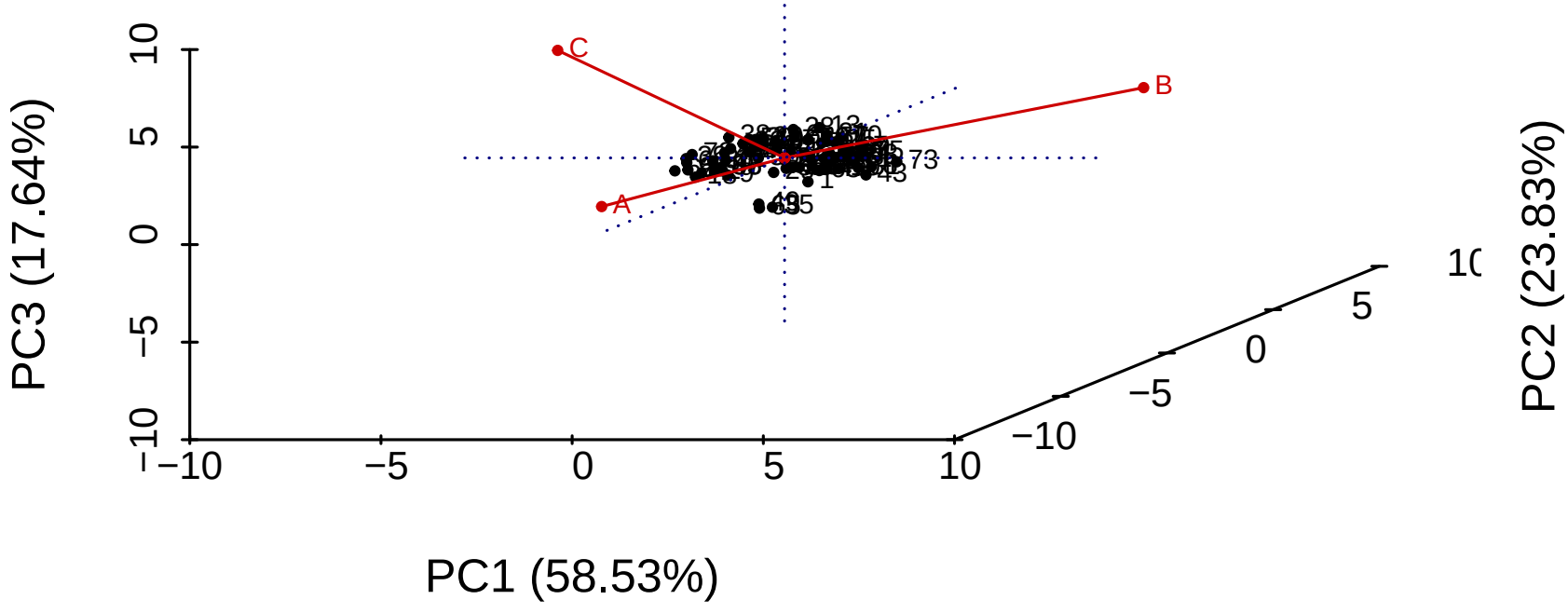
Duas dimensões

```
bp2 <- bpca(cafe[,2:4],d=1:2,scale=F)
plot.bpca.2d(bp2,var.cex=1.2,obj.cex=1.2,var.offset=.2,
    var.factor=.45)
```



Três dimensões

```
bp3 <- bpca(cafe[,2:4],d=1:3)
plot.bpca.3d(bp3)
```



Experimente ratacionar o gráfico

```
plot.bpca.3d(bp3,rgl=T)
```

Exemplo: Suco de Laranja

- Suco de Laranja
- 6 sucos
- 7 variáveis/atributos
- 2 variáveis auxiliares
  - Variáveis 8:15 são quantitativas suplementares e
  - 16 e 17 são qualitativas suplementares

Arquivo de dados

```
orange<-read.table('http://factominer.free.fr/book/orange.c
                   h=T,sep=';',dec='.',row.names=1)
```

Descrição rápida dos dados

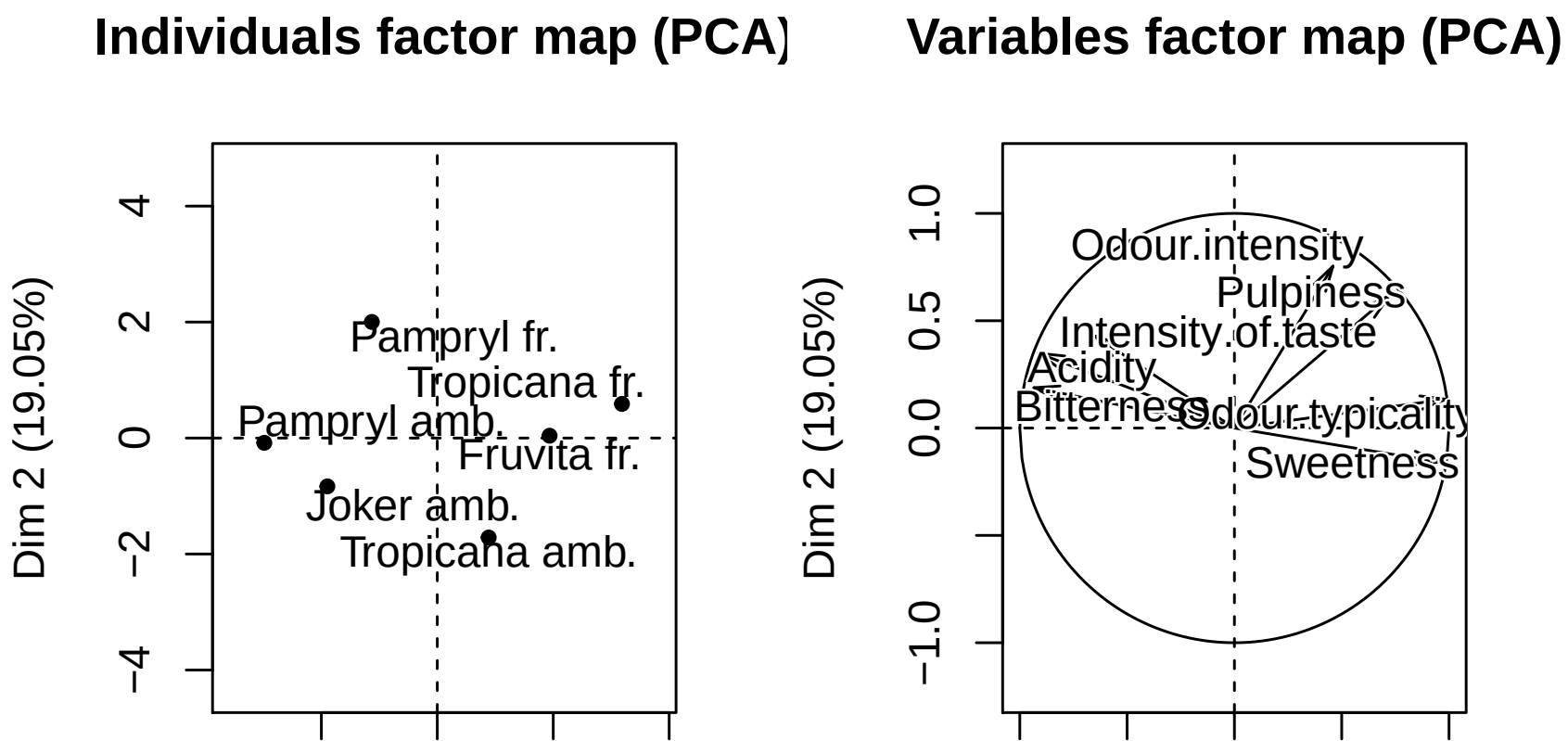
```
summary(orange)
```

```
 Odour.intensity Odour.typicality   Pulpiness      Intensity
 Min.   :2.760   Min.   :2.530    Min.   :1.660   Min.   :3
 1st Qu.:2.775   1st Qu.:2.625    1st Qu.:1.722   1st Qu.:3
 Median :2.825   Median :2.775    Median :2.625   Median :3
 Mean   :2.907   Mean   :2.762    Mean   :2.710   Mean   :3
 3rd Qu.:3.010   3rd Qu.:2.865    3rd Qu.:3.603   3rd Qu.:3
 Max.   :3.200   Max.   :3.020    Max.   :4.000   Max.   :3
    Acidity        Bitterness      Sweetness       Glucose
 Min.   :2.330   Min.   :1.760   Min.   :2.600   Min.   :17
 1st Qu.:2.453   1st Qu.:1.998   1st Qu.:2.825   1st Qu.:22
 Median :2.800   Median :2.320   Median :3.110   Median :24
 Mean   :2.802   Mean   :2.328   Mean   :3.057   Mean   :24
 3rd Qu.:3.125   3rd Qu.:2.612   3rd Qu.:3.335   3rd Qu.:26
 Max.   :3.310   Max.   :2.970   Max.   :3.380   Max.   :32
```

PCA

```
library(FactoMineR)

X11()
par(mfrow=c(1,2))
res.pca <- PCA(orange[,1:7])
```

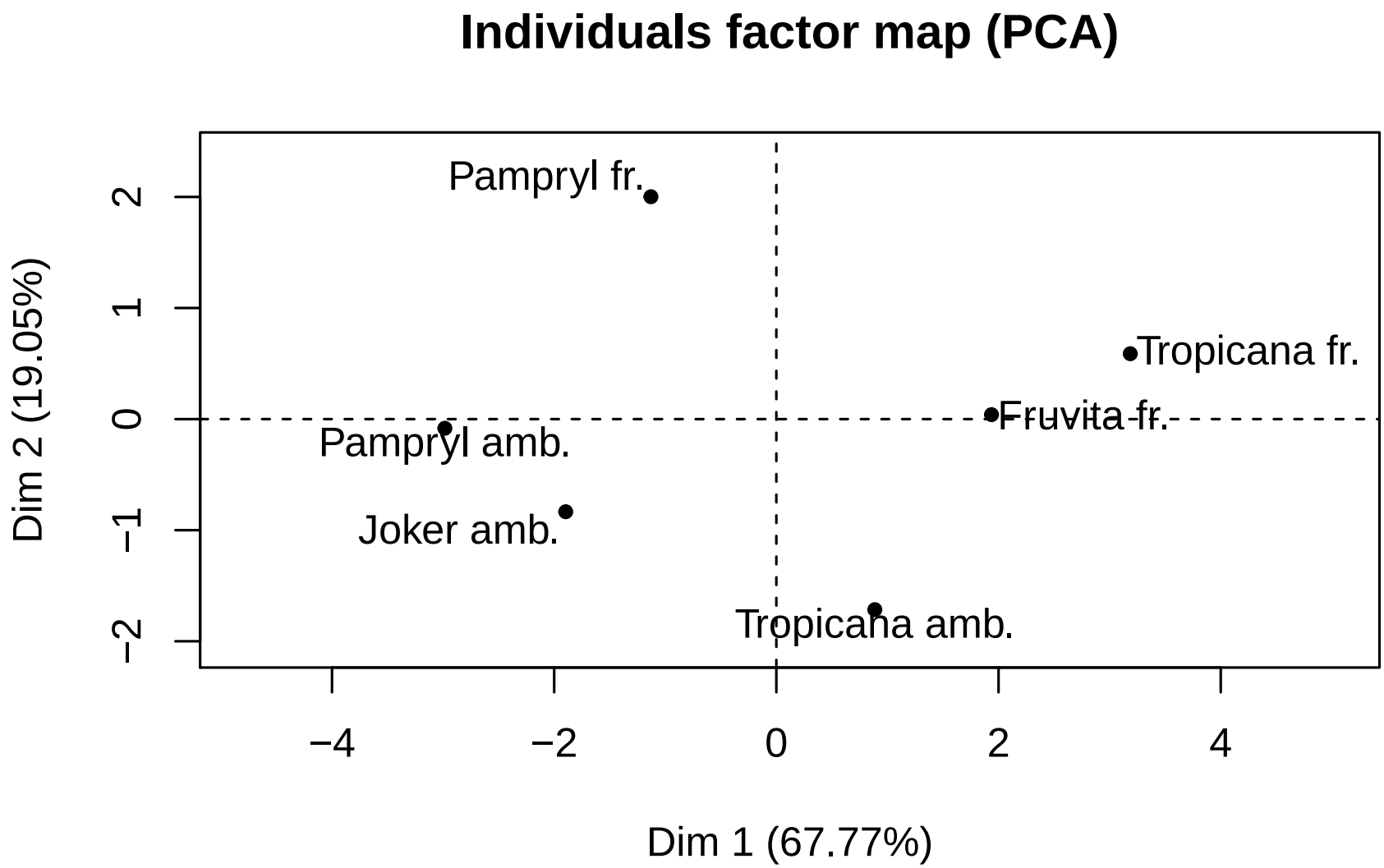


```
res.pca
```

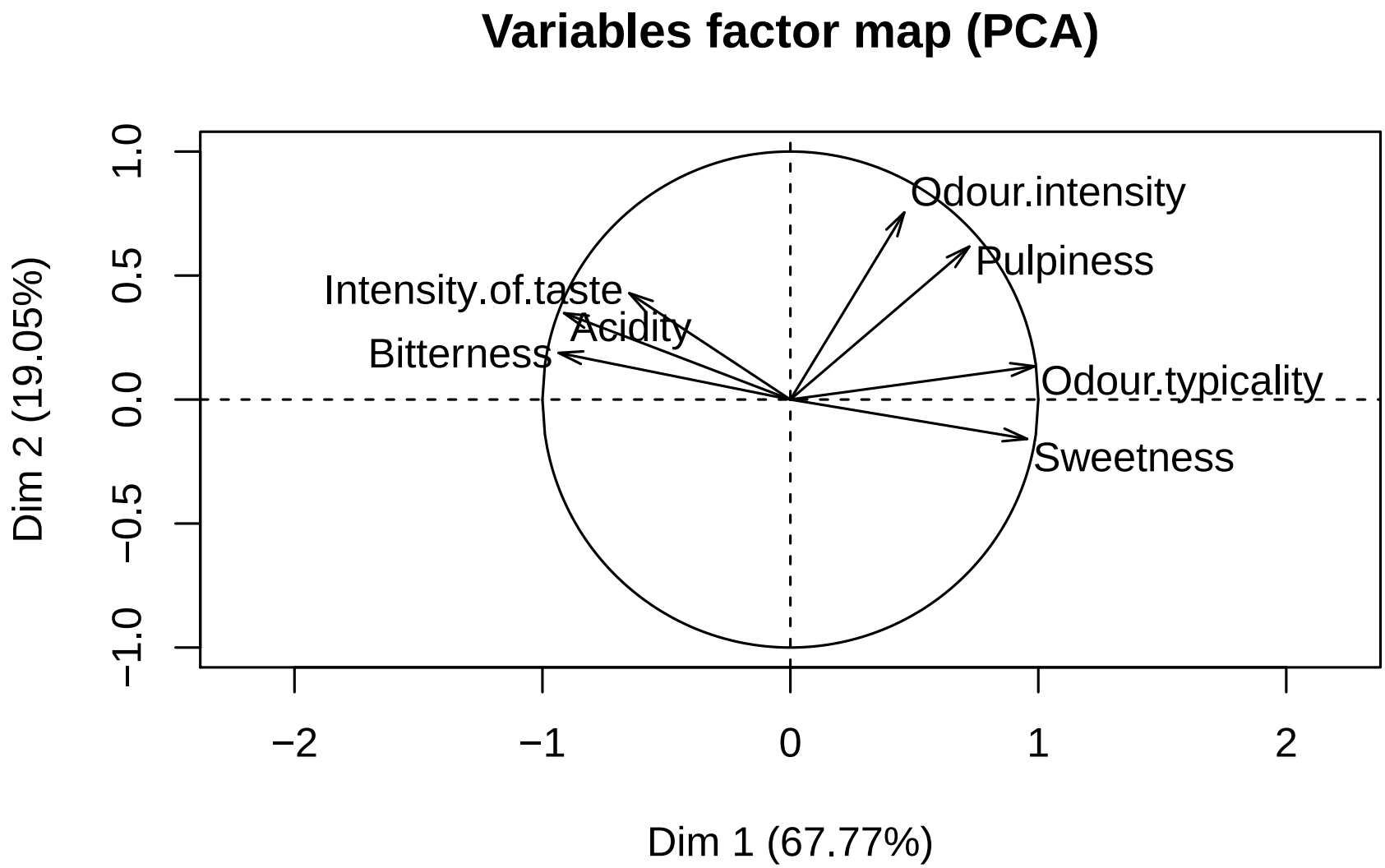
```
**Results for the Principal Component Analysis (PCA)**
The analysis was performed on 6 individuals, described by 7
*The results are available in the following objects:

   name                 description
1  "$eig"               "eigenvalues"
2  "$var"               "results for the variables"
3  "$var$coord"         "coord. for the variables"
4  "$var$cor"           "correlations variables - dimensions"
5  "$var$cos2"          "cos2 for the variables"
6  "$var$contrib"       "contributions of the variables"
7  "$ind"               "results for the individuals"
8  "$ind$coord"         "coord. for the individuals"
9  "$ind$cos2"          "cos2 for the individuals"
10 "$ind$contrib"       "contributions of the individuals"
11 "$call"              "summary statistics"
12 "$call$centre"       "mean of the variables"
```

```
par(mfrow=c(1,1))
plot(res.pca,choix=c('ind'))
```



```
par(mfrow=c(1,1))
plot(res.pca,choix=c('var'))
```



Interpretação

DIM1

- tropicana.fr e Pampryl.amb-> extremos para odour typicality
- tropicana.fr é mais typical e menos bitter
- Pampryl.amb é menos typical e mais bitter

DIM2

- Tropicana.amb tem odor menos intenso
- Pampryl.fr tem odor mais intenso

Correlação entre variáveis e as primeiras duas componentes

```
round(res.pca$var$coord[,1:2],2)
```

```
                    Dim.1 Dim.2
Odour.intensity      0.46  0.75
Odour.typicality     0.99  0.13
Pulpiness            0.72  0.62
Intensity.of.taste -0.65  0.43
Acidity            -0.91  0.35
Bitterness         -0.93  0.19
Sweetness           0.95 -0.16
```

Variância explicada

```
round(res.pca$eig,2)
```

```
       eigenvalue percentage of variance cumulative percent
comp 1       4.74                  67.77
comp 2       1.33                  19.05
comp 3       0.82                  11.71
comp 4       0.08                   1.20
comp 5       0.02                   0.27
```

Contribuição de cada indivíduo ou variável para cada uma das dimensões

```
round(res.pca$ind$contrib[,1:2],2)
```

```
                Dim.1 Dim.2
Pampryl amb.    31.29  0.08
Tropicana amb.   2.76 36.77
Fruvita fr.     13.18  0.02
Joker amb.      12.63  8.69
Tropicana fr.   35.66  4.33
Pampryl fr.      4.48 50.10
```

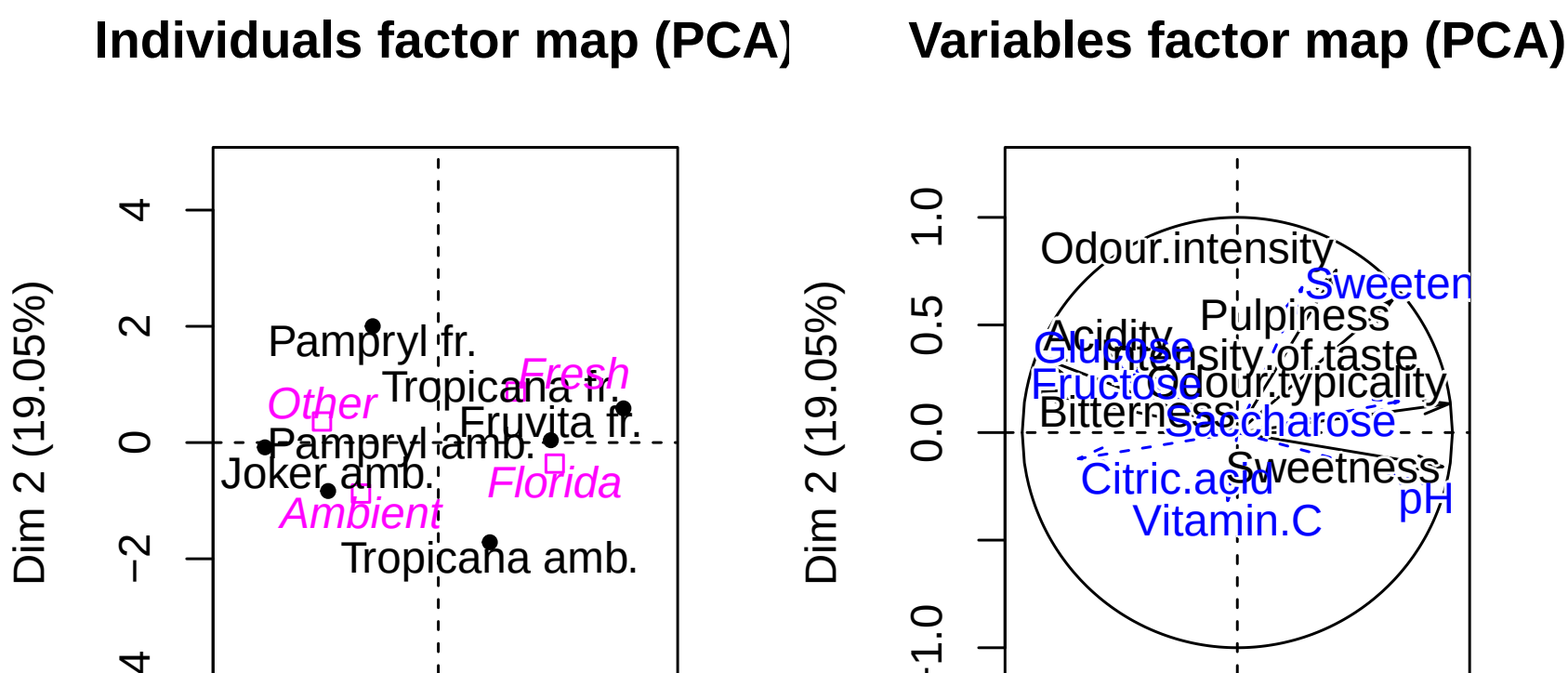
```
round(res.pca$var$contrib[,1:2],2)
```

```
                    Dim.1 Dim.2
Odour.intensity      4.45 42.69
Odour.typicality    20.47  1.35
Pulpiness           10.98 28.52
Intensity.of.taste   8.90 13.80
Acidity             17.56  9.10
Bitterness          18.42  2.65
Sweetness           19.22  1.89
```

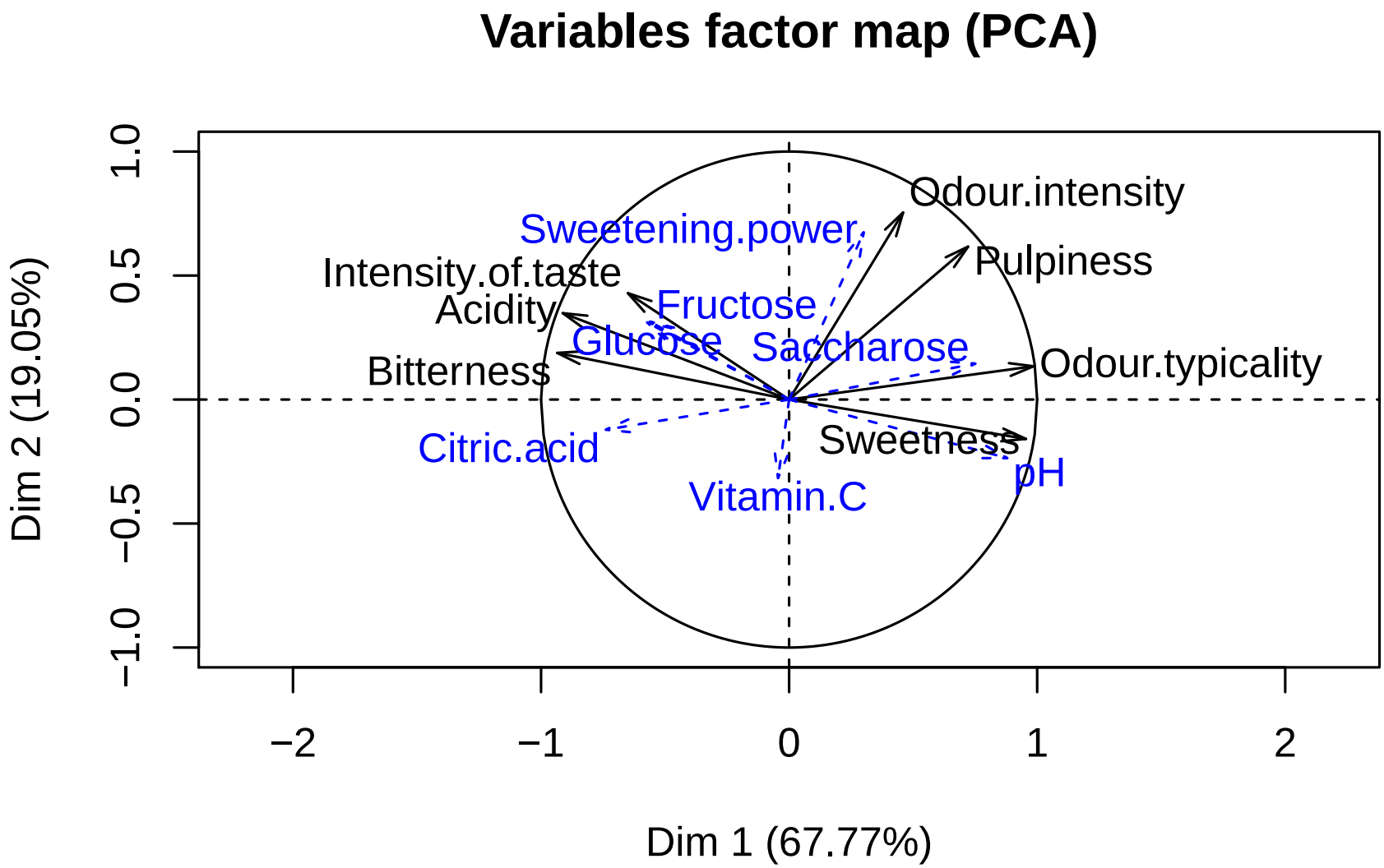
Adicionando informação: variáveis suplementares:

- Variáveis suplementares não contribuem para a construção do PCA auxiliam na interpretação

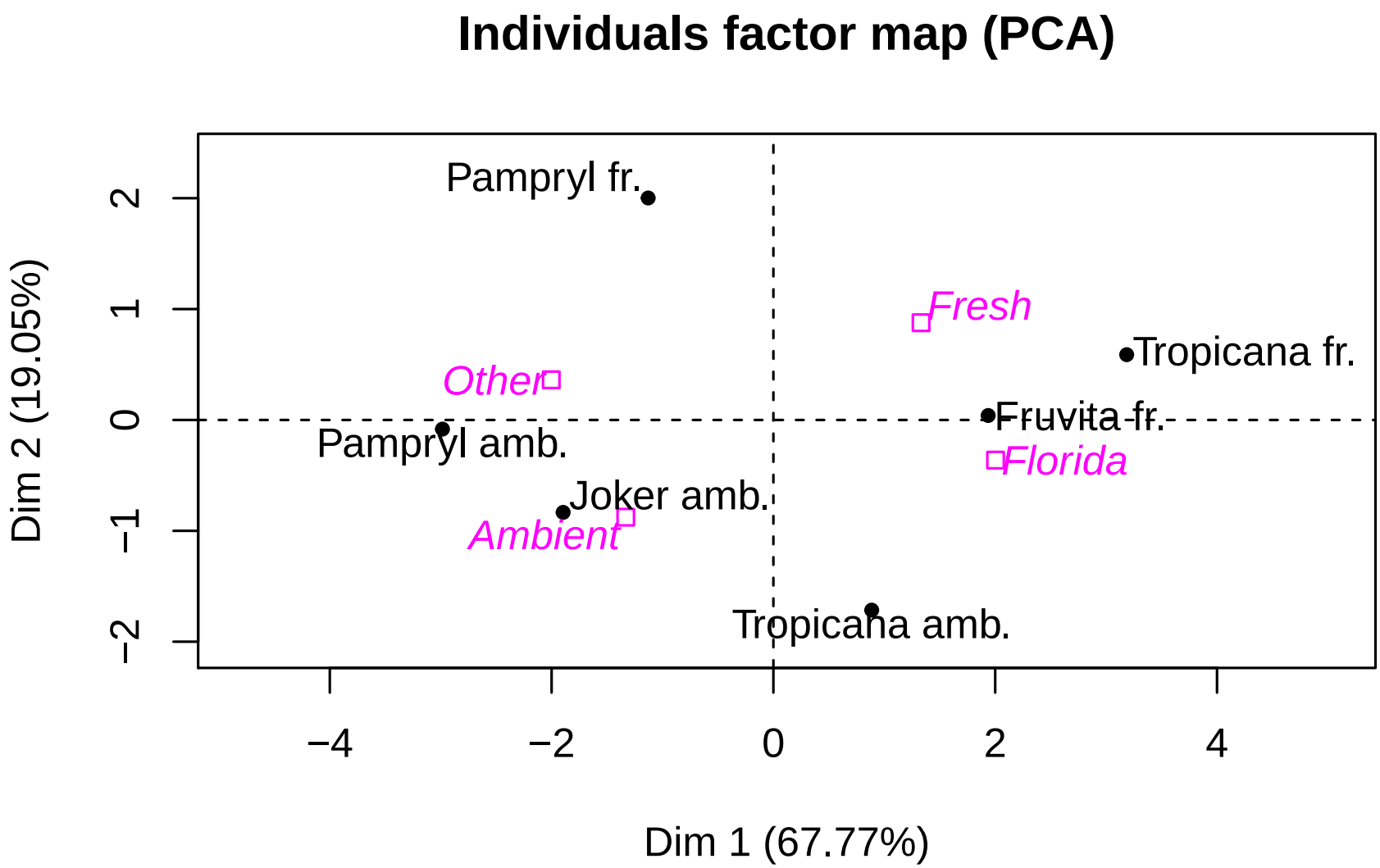
```
par(mfrow=c(1,2))
res.pca <- PCA(orange,quanti.sup=8:14,
                              quali.sup=15:16)
```



```
par(mfrow=c(1,1))
plot(res.pca,choix=('var'))
```



```
par(mfrow=c(1,1))
plot(res.pca,choix=('ind'))
```



Interpretação

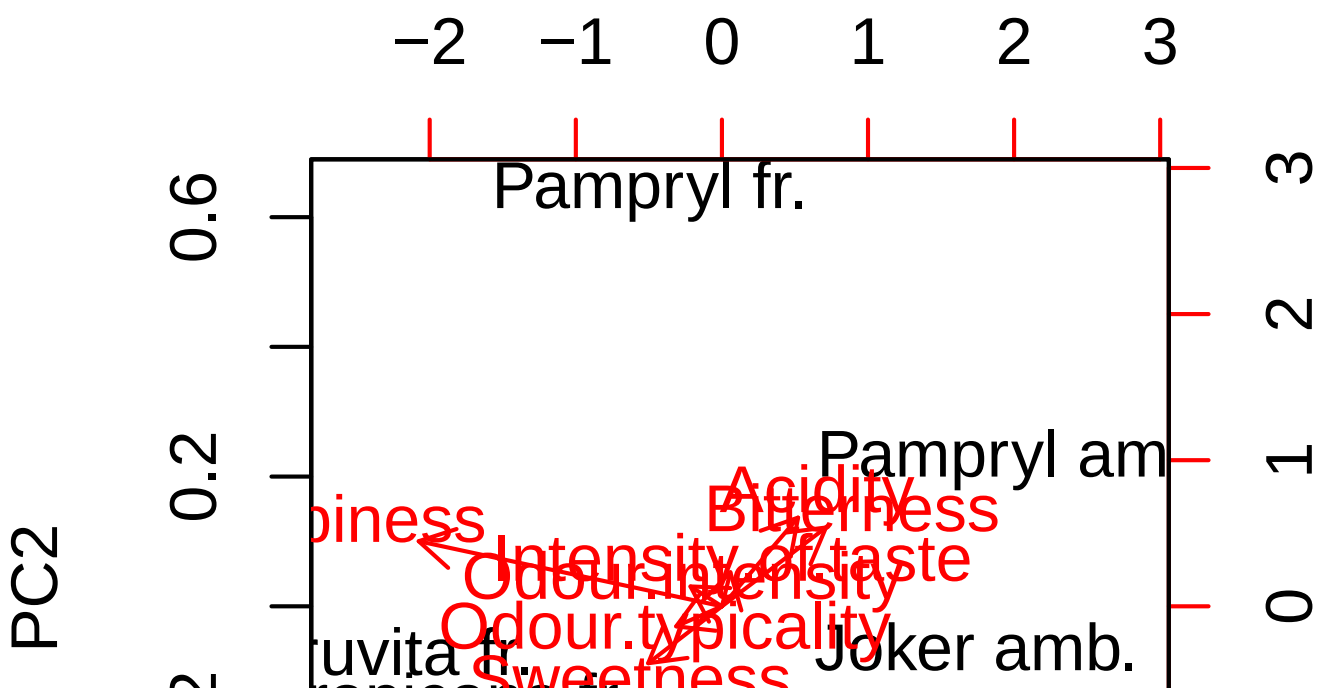
- pH e sacarose são correlacionadas com a primeira dimensão
- Primeira dimensão: doce e ácido
- Em ambiente ácido, sacarose transforma-se em glicose e frutose.
- A variável categórica suplementar é posicionada no centro dos indivíduos aos quais ela pertence

Biplot

- representando produtos/amostras e atributos/variáveis no mesmo gráfico

Função `biplot` (base do R)

```
biplot(prcomp(orange[,1:7]))
```



Os componentes principais

```
prcomp(orange[,1:7])
```

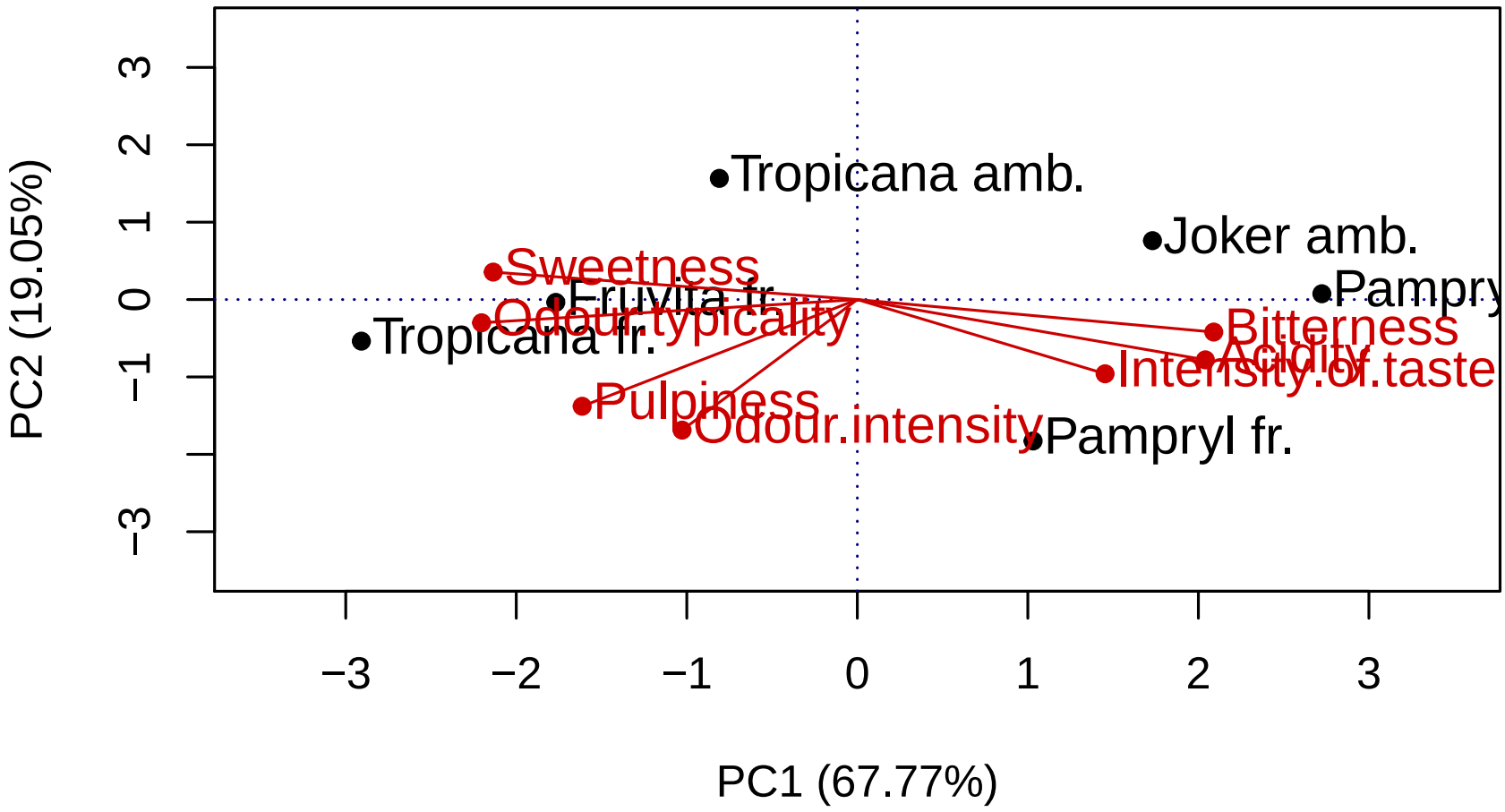
```
Standard deviations (1, .., p=6):
[1] 1.195422e+00 5.354149e-01 1.868676e-01 1.098345e-01 4.6
[6] 2.715608e-16

Rotation (n x k) = (7 x 6):
                          PC1        PC2         PC3
Odour.intensity    -0.09187630  0.1294499  0.67970061  0.25
Odour.typicality   -0.13361141 -0.1305171  0.26045122  0.26
Pulpiness          -0.88585708  0.4244513 -0.03989412 -0.11
Intensity.of.taste  0.02473181  0.2220744 -0.52319792 -0.08
Acidity             0.22252469  0.5786700 -0.23652412  0.68
Bitterness          0.30400795  0.5137590  0.36460721 -0.23
Sweetness          -0.21543965 -0.3713825 -0.07721764  0.55
                          PC5         PC6
```

Utilizando o pacote bpca

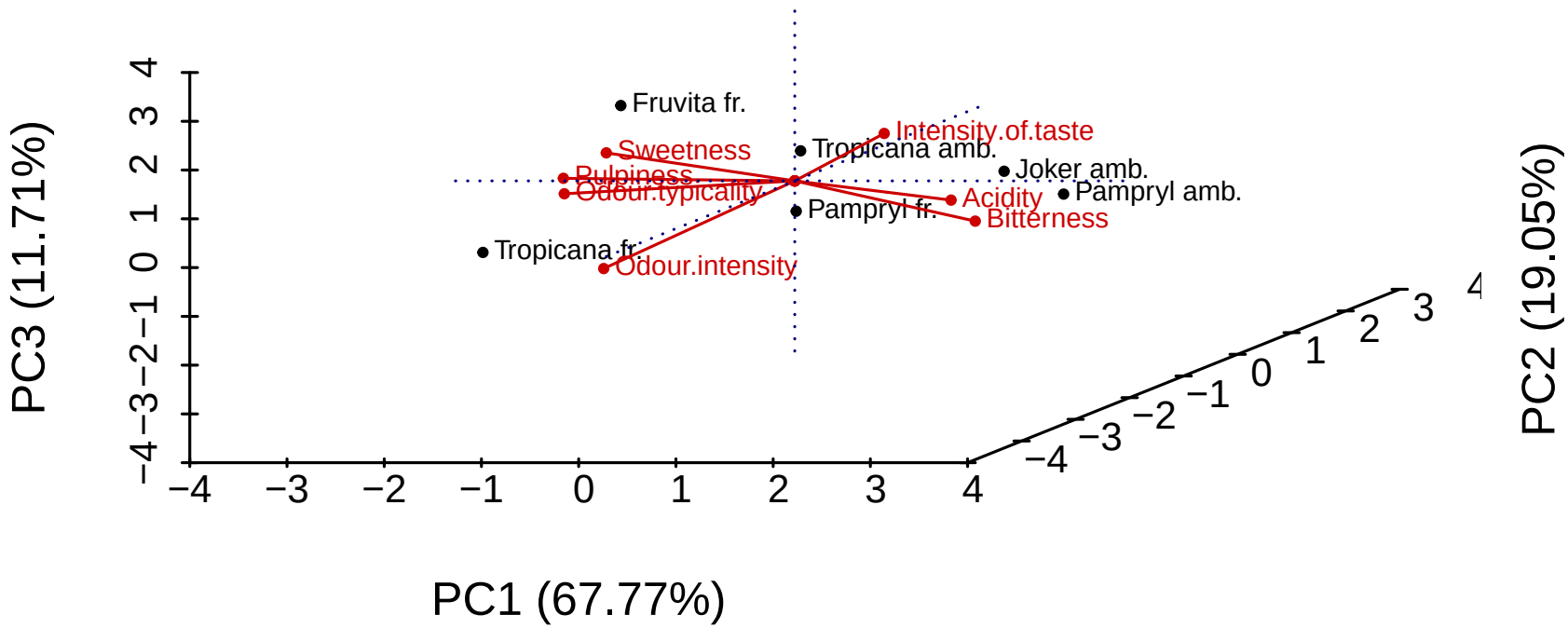
```
library(bpca)
```

```
bp2 <- bpca(orange[,1:7],d=1:2)
plot.bpca.2d(bp2,var.cex=1.2,obj.cex=1.2)
```



Utilizando 3 dimensões

```
bp3 <- bpca(orange[,1:7],d=1:3)
plot.bpca.3d(bp3)
```



Experimente rotacionar o gráfico!

```
plot.bpca.3d(bp3,rgl=T)
```