Exemplo de Análise de Componentes Principais (PCA)

20 de agosto de 2012

Poluentes atmosféricos

Na cidade de Curitiba foi amostrado ar diariamente durante seis meses. Os filtros amostrados em campo foram submetidos às seguintes análises químicas: gravimetria (medição de massa), refletância (determinação de fuligem) e fluorescência (quantificação química de elementos).

A partir dos dados de cada filtro validado após a análise deseja-se verificar a existência de relação entre os dados avaliados e a possibilidade de agrupamento de características de modo a obter a fonte das partículas. Para tanto, será avaliado o método de análise de componentes principais.

Os dados a serem avaliados foram dispostos em planilha com 13 colunas contendo as medições realizadas de Massa, Black carbon, Alumínio, Silicio, Enxofre, Cloro, Potássio, Calcio, Titanio, Magnesio, Ferro, Níquel e Zinco na mesma unidade de concentração. A planilha possui 130 linhas que é o total de amostras válidas no período.

Inserção de dados no R:

```
> dados <- read.table(file.choose(), header=T, dec=",")
> head(dados)
  massa   BC    Al    Si     S    Cl     K    Ca    Ti    Mn    Fe Ni    Zn
1 15.52 5.91 0.025 0.093 0.388 0.000 0.372 0.063 0.005 0.001 0.089  0 0.010
2  8.73 3.39 0.025 0.060 0.107 0.000 0.149 0.010 0.002 0.000 0.051  0 0.006
3 11.19 3.42 0.030 0.047 0.225 0.001 0.159 0.024 0.002 0.004 0.056  0 0.014
4  7.11 1.36 0.000 0.003 0.298 0.000 0.120 0.002 0.001 0.001 0.014  0 0.004
5  8.80 2.82 0.000 0.018 0.404 0.000 0.088 0.006 0.002 0.003 0.040  0 0.011
6  8.53 2.71 0.000 0.000 0.160 0.000 0.115 0.004 0.000 0.001 0.039  0 0.004

>

str(dados)
'data.frame':  130 obs. of  13 variables:
 $ massa: num  15.52 8.73 11.19 7.11 8.8 ...
 $ BC   : num  5.91 3.39 3.42 1.36 2.82 2.71 2.91 2.13 2.86 2.63 ...
 $ Al   : num  0.025 0.025 0.03 0 0 0 0.047 0.015 0.018 0 ...
 $ Si   : num  0.093 0.06 0.047 0.003 0.018 0 0.043 0.007 0.01 0.019 ...
 $ S    : num  0.388 0.107 0.225 0.298 0.404 0.16 0.206 0.088 0.272 0.373 ...
```

```
 $ Cl   : num  0 0 0.001 0 0 0 0.044 0 0 0 ...
 $ K    : num  0.372 0.149 0.159 0.12 0.088 0.115 0.182 0.062 0.038 0.067 ...
 $ Ca   : num  0.063 0.01 0.024 0.002 0.006 0.004 0.012 0.004 0.006 0.005 ...
 $ Ti   : num  0.005 0.002 0.002 0.001 0.002 0 0.001 0.001 0.001 0 ...
 $ Mn   : num  0.001 0 0.004 0.001 0.003 0.001 0.007 0.002 0.004 0.002 ...
 $ Fe   : num  0.089 0.051 0.056 0.014 0.04 0.039 0.085 0.031 0.051 0.035 ...
 $ Ni   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Zn   : num  0.01 0.006 0.014 0.004 0.011 0.004 0.322 0.007 0.012 0.007 ...
```

Resumo dos dados:

```
> summary(dados)
     massa              BC               Al                Si                S          
 Min.   : 1.000   Min.   :0.750   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000  
 1st Qu.: 2.325   1st Qu.:1.300   1st Qu.:0.00000   1st Qu.:0.02300   1st Qu.:0.03322  
 Median : 6.110   Median :1.900   Median :0.00000   Median :0.03650   Median :0.11700  
 Mean   : 7.702   Mean   :2.217   Mean   :0.01730   Mean   :0.04765   Mean   :0.18491  
 3rd Qu.:11.197   3rd Qu.:2.850   3rd Qu.:0.02325   3rd Qu.:0.05593   3rd Qu.:0.28925  
 Max.   :28.840   Max.   :8.200   Max.   :0.30770   Max.   :0.28520   Max.   :0.95170  
       Cl                K                 Ca                Ti                 Mn          
 Min.   :0.00000   Min.   :0.00000   Min.   :0.000000   Min.   :0.000000   Min.   :0.000000  
 1st Qu.:0.00425   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000500   1st Qu.:0.000000  
 Median :0.07275   Median :0.00000   Median :0.000000   Median :0.001000   Median :0.001000  
 Mean   :0.05717   Mean   :0.05985   Mean   :0.009066   Mean   :0.001942   Mean   :0.002089  
 3rd Qu.:0.09240   3rd Qu.:0.08375   3rd Qu.:0.011750   3rd Qu.:0.003000   3rd Qu.:0.002275  
 Max.   :0.27000   Max.   :0.50900   Max.   :0.082000   Max.   :0.013700   Max.   :0.063000  
       Fe                Ni                 Zn          
 Min.   :0.00000   Min.   :0.000000   Min.   :0.000000  
 1st Qu.:0.00645   1st Qu.:0.000000   1st Qu.:0.000000  
 Median :0.02905   Median :0.005400   Median :0.001050  
 Mean   :0.03419   Mean   :0.004933   Mean   :0.009756  
 3rd Qu.:0.05100   3rd Qu.:0.007450   3rd Qu.:0.007000  
 Max.   :0.14300   Max.   :0.061700   Max.   :0.322000  
```

Teste de Bartlett para verificar utilização ou não de ACP:

```
> require (psych)
Carregando pacotes exigidos: psych
> cortest.bartlett(dados)
R was not square, finding R from data
$chisq
[1] 1276.203

$p.value
[1] 5.88739e-216

$df
[1] 78
```

Qui quadrado grande e p.value pequeno, logo existe relação entre as variáveis.

Na planilha os dados de massa estão na escala de 0 a 100 e os dados de concentração de elementos estão na escala de 0 a 1, portanto uma padronização será necessária:

```
> dados <- scale(dados)
> (summary(dados))
     massa                BC                  Al                  Si          
 Min.   :-1.117e+00   Min.   :-1.155e+00   Min.   :-4.635e-01   Min.   :-1.175e+00  
 1st Qu.:-8.959e-01   1st Qu.:-7.220e-01   1st Qu.:-4.635e-01   1st Qu.:-6.080e-01  
 Median :-2.653e-01   Median :-2.498e-01   Median :-4.635e-01   Median :-2.750e-01  
 Mean   :-3.488e-17   Mean   :-8.824e-17   Mean   :-4.475e-18   Mean   :-8.953e-17  
```

```
 3rd Qu.: 5.824e-01   3rd Qu.: 4.978e-01   3rd Qu.: 1.593e-01   3rd Qu.: 2.042e-01
 Max.   : 3.522e+00   Max.   : 4.708e+00   Max.   : 7.779e+00   Max.   : 5.860e+00
       S                    Cl                   K                    Ca
 Min.   :-9.249e-01   Min.   :-1.225e+00   Min.   :-5.715e-01   Min.   :-5.886e-01
 1st Qu.:-7.587e-01   1st Qu.:-1.134e+00   1st Qu.:-5.715e-01   1st Qu.:-5.886e-01
 Median :-3.397e-01   Median : 3.341e-01   Median :-5.715e-01   Median :-5.886e-01
 Mean   : 2.326e-17   Mean   :-5.136e-17   Mean   : 2.342e-17   Mean   :-3.886e-17
 3rd Qu.: 5.219e-01   3rd Qu.: 7.553e-01   3rd Qu.: 2.283e-01   3rd Qu.: 1.743e-01
 Max.   : 3.836e+00   Max.   : 4.562e+00   Max.   : 4.289e+00   Max.   : 4.735e+00
       Ti                   Mn                   Fe                   Ni
 Min.   :-8.770e-01   Min.   :-3.697e-01   Min.   :-1.091e+00   Min.   :-7.649e-01
 1st Qu.:-6.512e-01   1st Qu.:-3.697e-01   1st Qu.:-8.849e-01   1st Qu.:-7.649e-01
 Median :-4.255e-01   Median :-1.927e-01   Median :-1.640e-01   Median : 7.240e-02
 Mean   :-1.955e-17   Mean   :-4.077e-17   Mean   :-1.218e-17   Mean   : 5.279e-17
 3rd Qu.: 4.776e-01   3rd Qu.: 3.287e-02   3rd Qu.: 5.361e-01   3rd Qu.: 3.903e-01
 Max.   : 5.309e+00   Max.   : 1.078e+01   Max.   : 3.470e+00   Max.   : 8.802e+00
       Zn
 Min.   :-2.675e-01
 1st Qu.:-2.675e-01
 Median :-2.387e-01
 Mean   :-2.091e-17
 3rd Qu.:-7.558e-02
 Max.   : 8.562e+00
```

Calculando a matriz de correlação:

```
> dados <- cor(dados)
```

Realizando a análise de componentes:

```
> (pca <- prcomp(dados))
Standard deviations:
[1] 2.874794e+00 3.334388e-01 3.367348e-02 3.850040e-03 2.897064e-07 1.187965e-07 4.160805e-08
[8] 3.535486e-09 2.475120e-10 2.500568e-11 2.831220e-12 4.424326e-14 1.631730e-17

Rotation:
             PC1          PC2         PC3          PC4         PC5          PC6         PC7
massa -0.2902538 -0.129544492 -0.14389657 -0.084911499 -0.01766872  0.236326863  0.09321576
BC    -0.2906995 -0.115171847 -0.14004229 -0.208965378  0.17010538  0.135159021  0.13913318
Al    -0.2552113  0.418611697 -0.11422954  0.331285320 -0.50634275 -0.294612459 -0.38864786
Si    -0.1908526  0.678723096 -0.02558878  0.131168685  0.37390150  0.508220489  0.02922710
S     -0.2918122  0.029370429 -0.07958658 -0.076418013 -0.53767779  0.304070985  0.27494977
Cl     0.2919542  0.093850952  0.04422154  0.063102222  0.30324747  0.016987095 -0.22296207
K     -0.2874571 -0.198092752 -0.08400719 -0.220980659  0.02383573 -0.073455606 -0.08035353
Ca    -0.2901893 -0.142487863 -0.08213161 -0.232318059  0.16330907 -0.166522957 -0.28878541
Ti    -0.2750342  0.280735505 -0.11945272 -0.001145275  0.30569494 -0.660515713  0.35300479
Mn     0.2766262 -0.068518204 -0.93799632  0.154201196  0.02442000  0.038080885  0.03858622
Fe    -0.2922295  0.002886167 -0.08115753 -0.079525330  0.10051510 -0.001478781  0.18896892
Ni     0.2905783  0.128883902  0.10690624  0.017421852 -0.17843138 -0.120875810  0.65205452
Zn    -0.2654359 -0.405847465  0.10846587  0.824096758  0.17402719  0.050788092  0.13380067
             PC8          PC9         PC10         PC11         PC12         PC13
massa -0.41201807 -0.57746087 -0.44263237 -0.054828669 -0.290088284  0.141075318
BC    -0.22440713  0.42840510 -0.35579450 -0.093103670  0.634739196 -0.018645323
Al    -0.26465068  0.17891994 -0.16148080 -0.026093889  0.021281756  0.115517419
Si    -0.04821978 -0.02699001  0.23625634  0.157931581  0.016538851  0.044933719
S      0.56043897 -0.09990264  0.08222012 -0.227556827  0.132515604  0.210903617
Cl     0.26262834  0.10141280 -0.37688470 -0.341428661 -0.072756801  0.644119766
K      0.09605749  0.21031332  0.07728826  0.702202167 -0.179742639  0.474331054
Ca    -0.21178909 -0.15136685  0.60355877 -0.433437265  0.111133267  0.254372062
Ti     0.27794610 -0.26907305 -0.13314610  0.019826604  0.086539309 -0.040821068
Mn     0.02050474  0.01679845  0.10180772  0.020892706 -0.009190603 -0.006653509
Fe     0.01750310  0.52399636 -0.01980957 -0.333595418 -0.658501228 -0.188910791
Ni    -0.43816037  0.12284503  0.19826012 -0.034782586 -0.003333955  0.409870049
Zn     0.02012265  0.02264775  0.09350221 -0.008095173  0.059228670  0.099647967
```

Para conhecer a importância das componentes geradas:

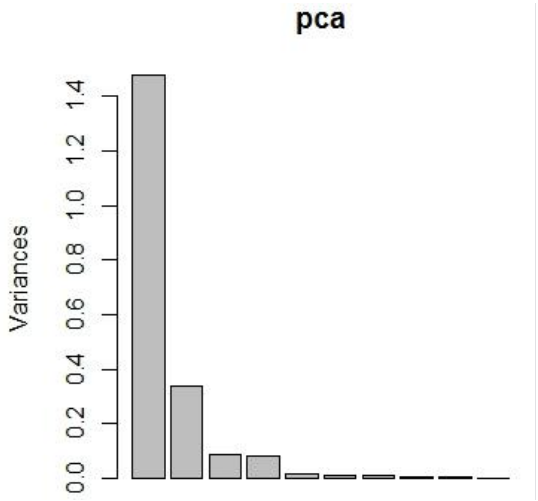
```
> (summary(pca))
```

```
Importance of components:
                          PC1     PC2     PC3     PC4       PC5       PC6       PC7       PC8
Standard deviation     2.8748 0.33344 0.03367 0.00385 2.897e-07 1.188e-07 4.161e-08 3.535e-09
Proportion of Variance 0.9866 0.01327 0.00014 0.00000 0.000e+00 0.000e+00 0.000e+00 0.000e+00
Cumulative Proportion  0.9866 0.99986 1.00000 1.00000 1.000e+00 1.000e+00 1.000e+00 1.000e+00
                           PC9      PC10      PC11      PC12      PC13
Standard deviation     2.475e-10 2.501e-11 2.831e-12 4.424e-14 1.632e-17
Proportion of Variance 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00 1.000e+00 1.000e+00
```

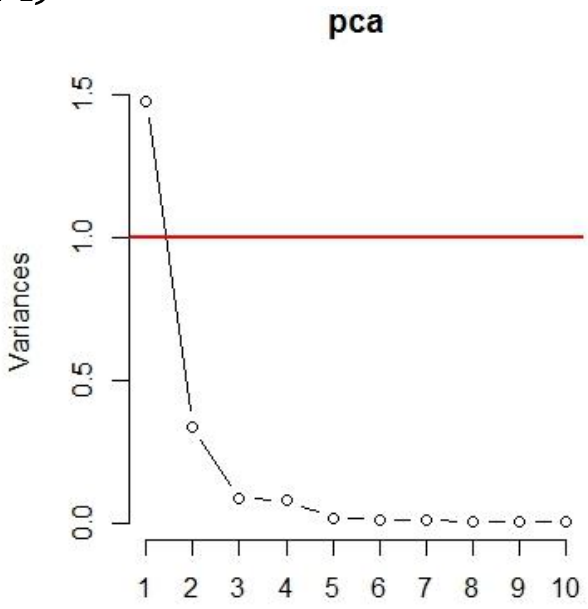
Para conhecer o valor dos autovalores gerados:

```
> (auto$values)
[1] 5.73584815 2.34473205 1.32204829 1.00454338 0.72081720 0.45005906 0.34123389 0.31143799
[9] 0.25775299 0.19070236 0.13347497 0.11599675 0.07135292
```

```
> screeplot(pca)
```




```
> screeplot(pca, type = c("lines"))
> abline(h=1, col=2, lwd=2)
```



Há várias possibilidades de escolha do número de autovalores a serem analisados, seguindo a regra de Kaiser, teremos apenas um autovalor. Como no presente estudo uma componente não será satisfatória, serão selecionadas as primeiras duas componentes a partir do screeplot, pois a partir deste ponto, o incremento das demais é pequeno.

Após realizada a escolha do número de autovalores, é necessário conhecer quais dados contribuem mais para cada uma das componentes principais, tal avaliação é realizada pelo escore da componente.

Matriz de coeficientes das componentes principais:

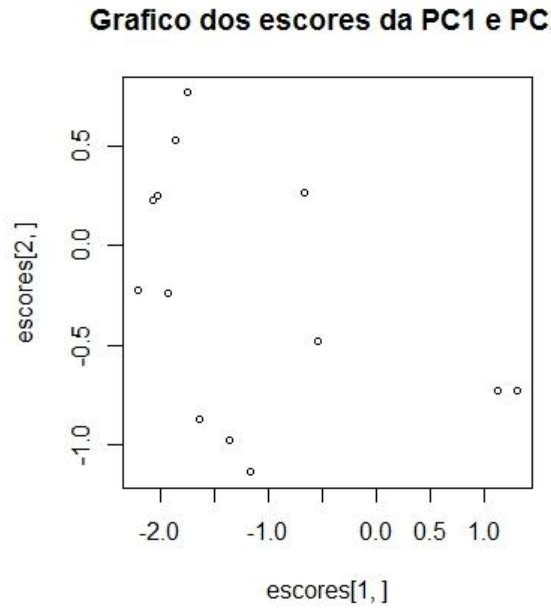
```
> auto <- as.data.frame(eigen(dados))
> m <- round(as.matrix(auto[,-1]),3)
```

13 observations of 13 variables

	vectors.1	vectors.2	vectors.3	vectors.4	vectors.5	vectors.6	vectors.7	vectors.8	vectors.9	vectors.10	vectors.11	vectors.12	vectors.13
1	-0.352	0.109	0.233	0.090	0.115	0.155	-0.278	-0.257	0.456	0.442	0.324	-0.175	0.288
2	-0.359	0.098	0.229	-0.219	-0.040	-0.036	-0.299	-0.276	-0.093	-0.380	0.129	-0.190	-0.622
3	-0.238	-0.417	-0.156	0.244	0.169	0.019	0.387	0.205	0.389	-0.423	0.313	-0.188	-0.038
4	-0.202	-0.487	-0.207	-0.054	-0.088	-0.213	-0.448	-0.002	0.223	0.028	-0.139	0.593	-0.043
5	-0.336	-0.101	-0.160	-0.020	-0.002	0.592	-0.194	0.550	-0.230	0.208	-0.149	-0.155	-0.134
6	0.229	-0.311	0.437	-0.100	-0.202	-0.394	-0.266	0.464	-0.067	0.074	0.209	-0.328	0.081
7	-0.304	0.330	0.155	-0.182	-0.227	-0.016	0.200	0.294	-0.158	-0.112	0.485	0.510	0.175
8	-0.322	0.227	0.075	-0.262	-0.216	-0.299	0.310	0.218	0.426	0.137	-0.519	-0.120	-0.091
9	-0.284	-0.373	-0.072	-0.049	-0.048	-0.229	0.404	-0.260	-0.415	0.511	0.150	-0.063	-0.166
10	-0.095	-0.203	0.693	0.048	0.489	0.136	0.164	0.006	-0.103	-0.063	-0.308	0.266	0.049
11	-0.383	-0.095	-0.057	-0.016	-0.166	-0.084	-0.110	-0.192	-0.316	-0.364	-0.244	-0.243	0.638
12	0.198	-0.313	0.239	-0.213	-0.614	0.501	0.189	-0.245	0.178	-0.052	-0.020	0.038	0.018
13	-0.115	0.115	0.186	0.848	-0.404	-0.069	-0.025	0.014	-0.066	0.024	-0.118	0.077	-0.163

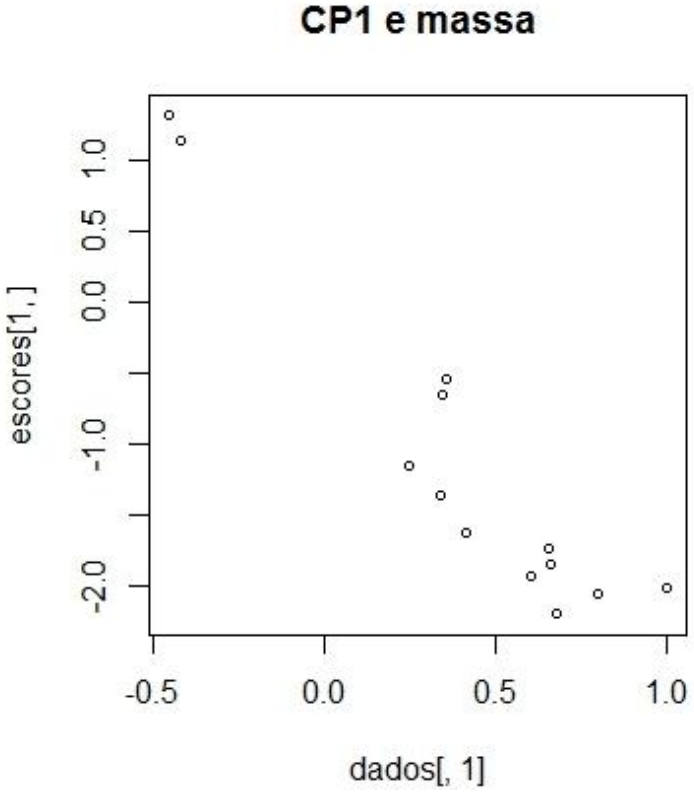
Os escores de cada componente principal podem ser encontrados pela substituição dos valores dos dados padronizados.

```
> escores <- round(t(m) %*% t(dados),2)
> plot(escores[2,] ~ escores[1,], cex = 0.7, main = "Grafico dos escores da PC1 e PC2")
```

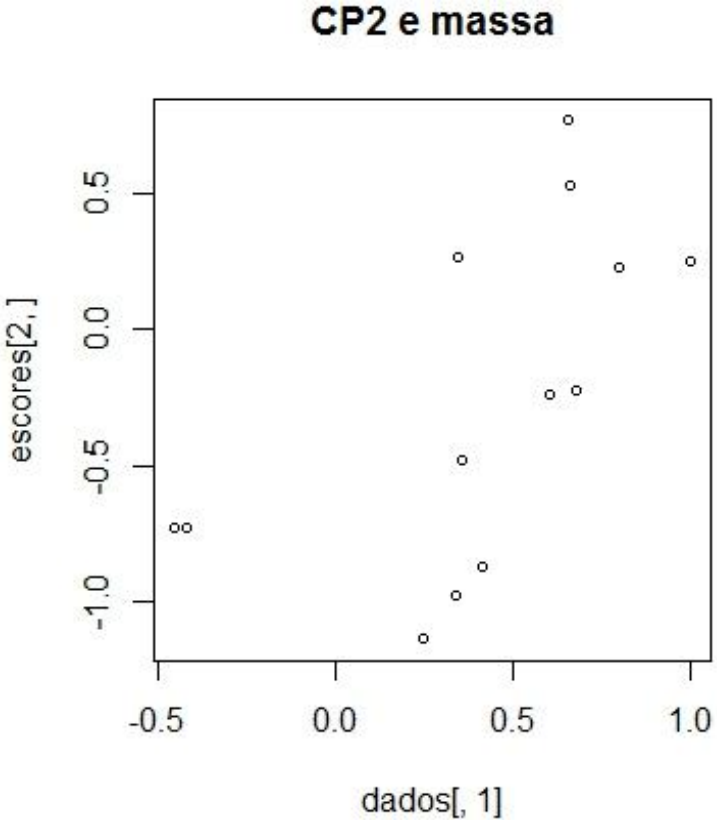


A correlação das componentes principais com os dados pode ser feita da seguinte forma:

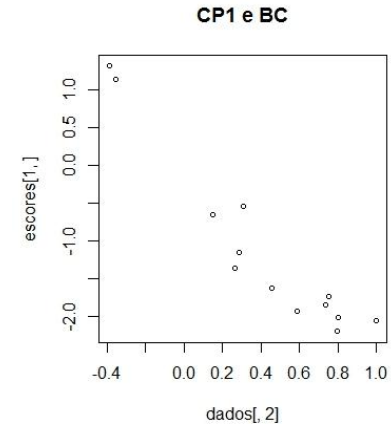
```
> plot(escores[1,] ~ dados[,1], cex = 0.7, main = "CP1 e massa")
```

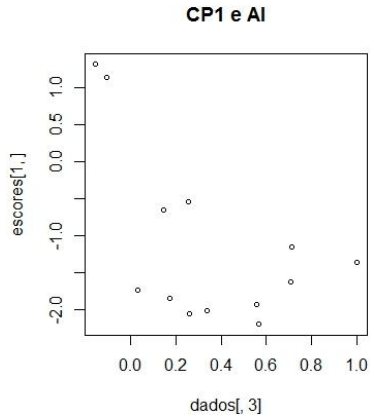


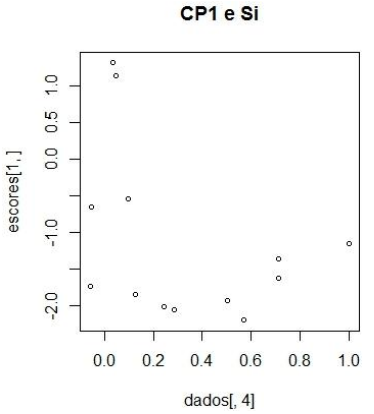

```
> plot(escores[2,] ~ dados[,1], cex = 0.7, main = "CP2 e massa")
```

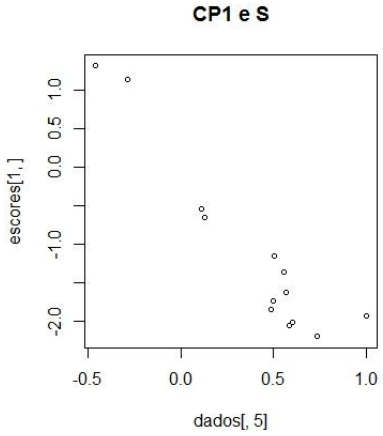


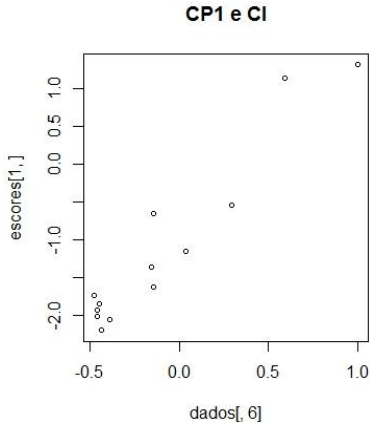
Seguindo o mesmo procedimento obtém-se para a componente 1:

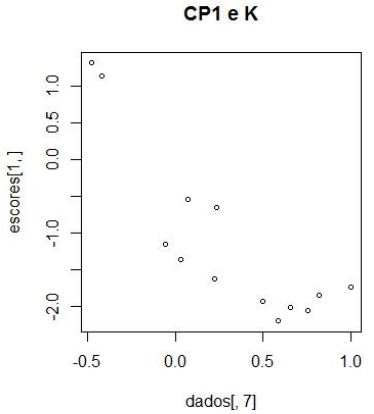


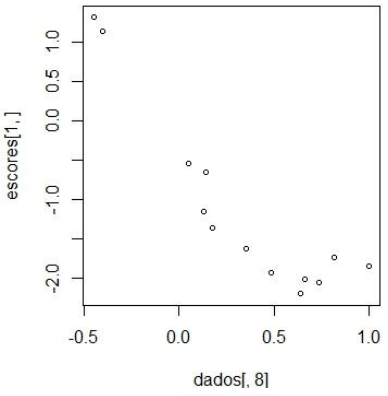


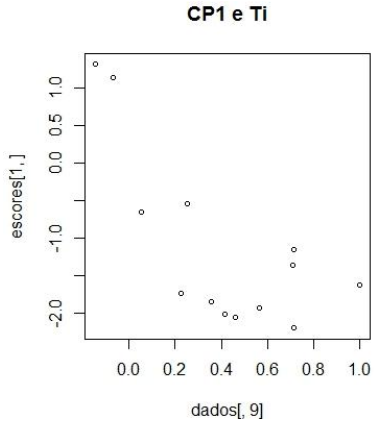


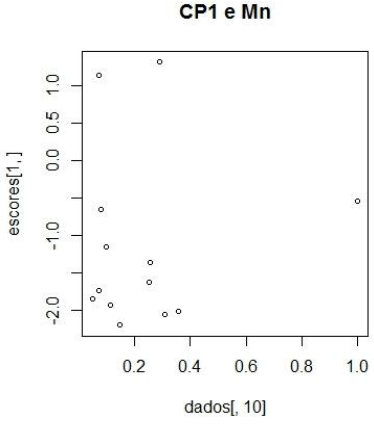


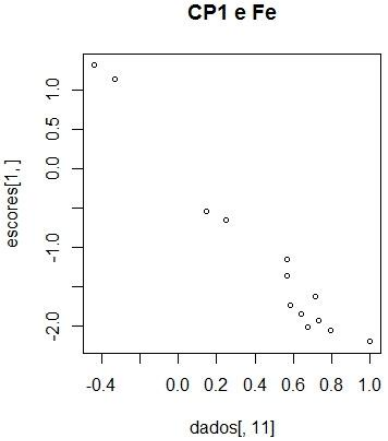


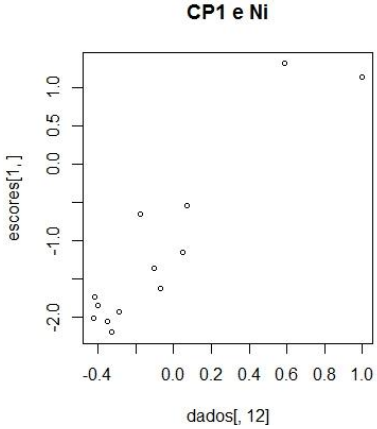


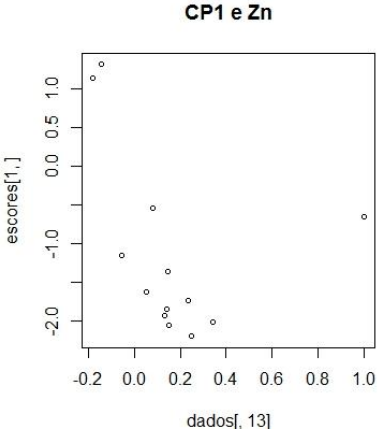




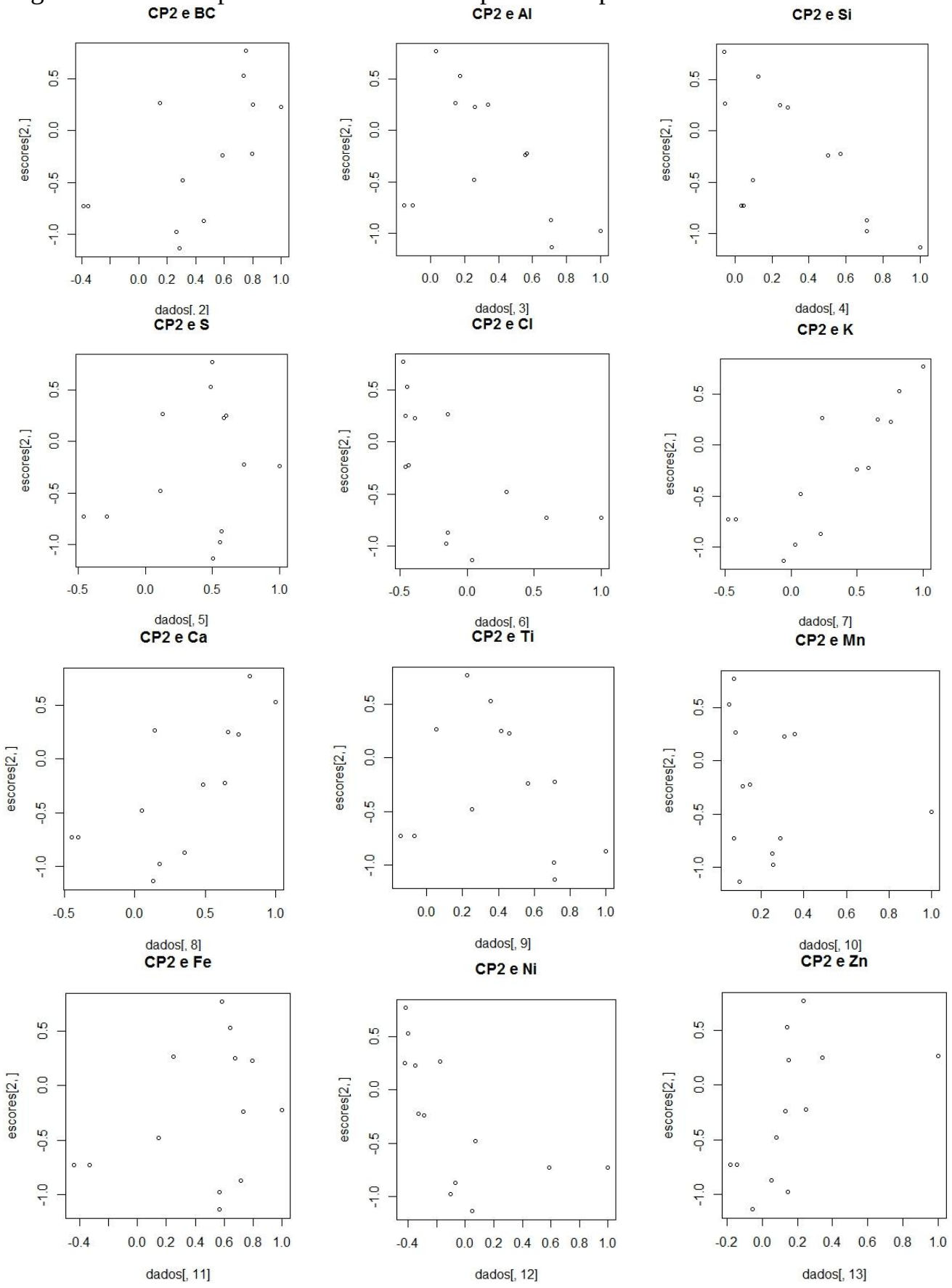








Seguindo o mesmo procedimento obtém-se para a componente 2:



Visualizando a marcação dos pontos no domínio das duas componentes:

```
> biplot(pca, cex=0.6)
```