

Análise de componentes principais

2026

Os dados se referem ao consumo municipal de energia elétrica em MWh nos setores (1) industrial, (2) comércio e serviços, (3) rural, (4) residencial e (5) iluminação, serviços públicos e outros no ano de 2019 em São Paulo. Os dados originiais foram obtidos da página <http://www.imp.seade.gov.br/frontend/#/tabelas>. O município de São Paulo foi excluído por apresentar padrão de consumo bastante diferente dos demais. Outros municípios com dados incompletos também foram excluídos.

A primeira coluna do arquivo contém os nomes dos municípios.

```
## Leitura dos dados
dados <- read.csv2("consumo-ee-SP-2019.csv", header = TRUE, sep = ";",)
n <- nrow(dados)
p <- ncol(dados) - 1
cat("\n n =", n, ", p =", p, "\n")
```

```
##
## n = 630 , p = 5
```

O vetor `setor` contém os nomes dos setores de consumo.

```
setor <- colnames(dados)[-1]
```

Os nomes dos municípios são armazenados em um vetor, a primeira coluna da folha de dados é eliminada e os nomes dos setores são anexados à folha de dados.

```
munic <- dados[, 1]
dados <- dados[, -1]
names(dados) <- setor
munic[1:10]
```

```
## [1] "Adamantina"          "Adolfo"               "Aguai"
## [4] "Aguas da Prata"      "Aguas de Lindoia"     "Aguas de Santa Barbara"
## [7] "Agudos"              "Alambari"             "Alfredo Marcondes"
## [10] "Altinópolis"
```

Apresentamos algumas medidas resumo das sete variáveis.

```
summary(dados)
```

```
##      Industrial      ComServ      Rural      Residencial
## Min.   :    2.0   Min.   :   159   Min.   :    1   Min.   :   583
## 1st Qu.:   385.8   1st Qu.:  1152   1st Qu.: 1290   1st Qu.:   3878
## Median :  3057.5   Median :   3410   Median : 2872   Median :   9790
## Mean   :  65744.0   Mean   :  25635   Mean   :  5463   Mean   :  41829
## 3rd Qu.: 27986.8   3rd Qu.: 14714   3rd Qu.: 5933   3rd Qu.: 31790
## Max.   :3711523.0   Max.   :1151682   Max.   :82055   Max.   :1147449
## IluminServPubOutros
## Min.   :   366
## 1st Qu.:  1400
## Median :   3212
## Mean   : 13048
```

```
## 3rd Qu.: 10154
## Max.    :334779
```

```
## Desvios padrão
```

```
apply(dados, MARGIN = 2, FUN = sd)
```

```
##          Industrial          ComServ          Rural          Residencial
##          222933.762          84816.263          8375.389          103389.266
## IluminServPubOutros
##          32635.285
```

Para reduzir a magnitude dos dados optamos pelo consumo em GWh.

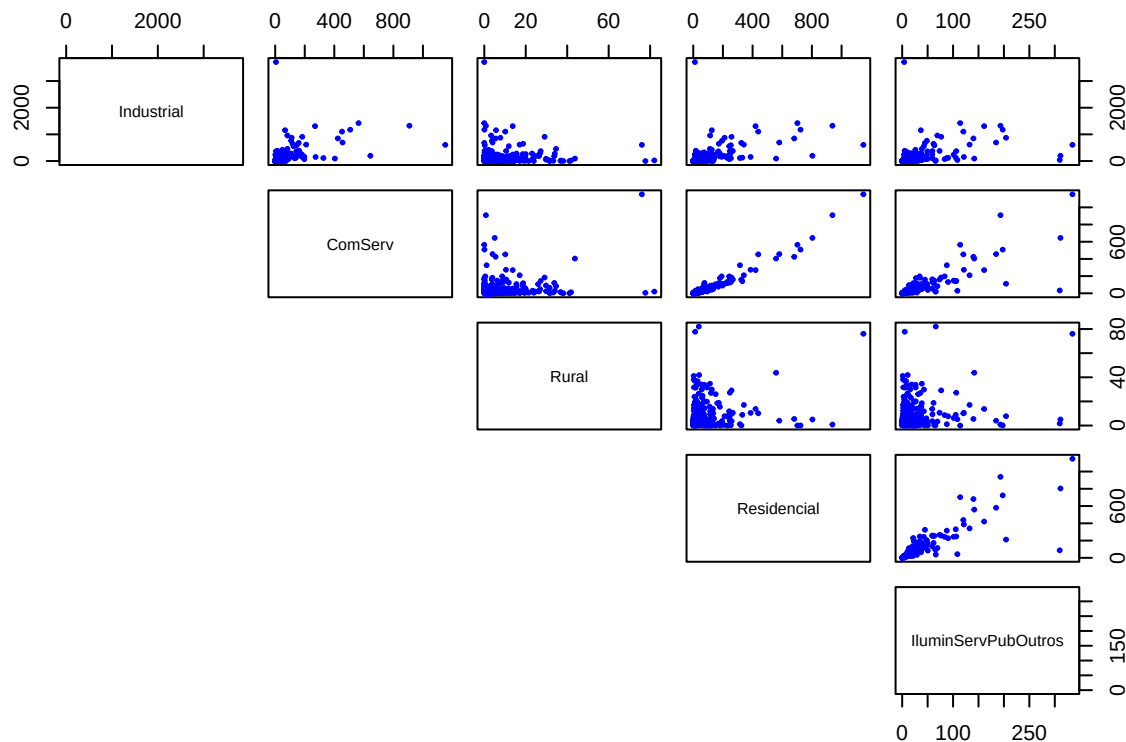
```
dados <- dados / 1000
```

```
apply(dados, MARGIN = 2, FUN = sd)
```

```
##          Industrial          ComServ          Rural          Residencial
##          222.933762          84.816263          8.375389          103.389266
## IluminServPubOutros
##          32.635285
```

Uma matriz de gráficos de dispersão permite representar graficamente os dados. Notamos um padrão de associação positiva entre alguns pares de variáveis.

```
pairs(dados, lower.panel = NULL, pch = 20, col = "blue", cex = 0.6)
```



As correlações amostrais variam de 0,10 a 0,97.

```
matcor <- cor(dados)
```

```
print(matcor, digits = 3)
```

```
##          Industrial ComServ Rural Residencial IluminServPubOutros
## Industrial          1.000   0.514 0.102          0.551          0.475
## ComServ            0.514   1.000 0.272          0.971          0.837
```

```
## Rural          0.102  0.272 1.000      0.262      0.280
## Residencial    0.551  0.971 0.262      1.000      0.871
## IluminServPubOutros 0.475  0.837 0.280      0.871      1.000
```

```
range(matcor[matcor < 1.00])
```

```
## [1] 0.1023928 0.9712954
```

As variáveis são consumos de energia elétrica e são medidas na mesma unidade (GWh). A análise de componentes principais será baseada na matriz de covariâncias amostral.

```
mcov <- cov(dados) # S
acp <- eigen(mcov) # cálculo de "lambda" e "e"
E <- acp$vectors   # e1, ..., ep
colnames(E) <- paste("CP", 1:p, sep = "")
lambda <- acp$values
```

Em seguida apresentamos as estimativas das variâncias dos CP e as percentagens acumuladas da variância total explicada.

```
cat("\n Variâncias dos CP = ", lambda)
```

```
##
```

```
## Variâncias dos CP = 56699.35 11454.78 311.961 188.1935 63.52943
```

```
cat("\n Média das variâncias =", mean(lambda), "\n")
```

```
##
```

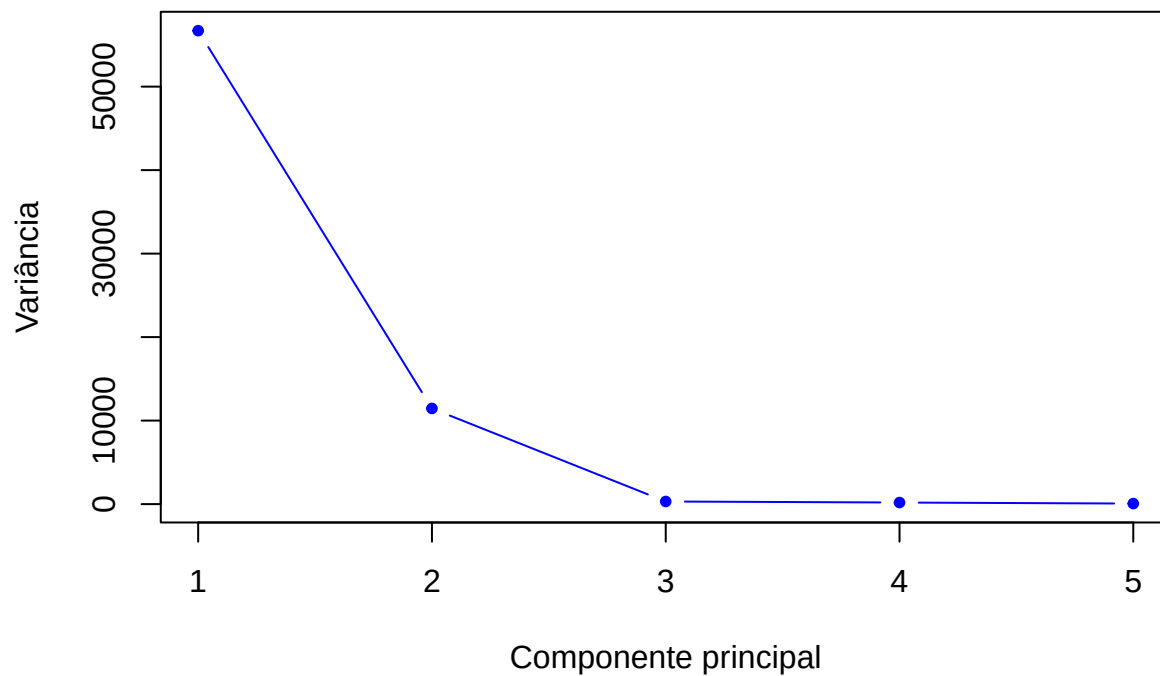
```
## Média das variâncias = 13743.56
```

```
var <- sum(lambda) # variância total
propa <- cumsum(lambda) / var
cbind(1:p, propa * 100)
```

```
##      [,1]      [,2]
## [1,]    1  82.51041
## [2,]    2  99.17971
## [3,]    3  99.63369
## [4,]    4  99.90755
## [5,]    5 100.00000
```

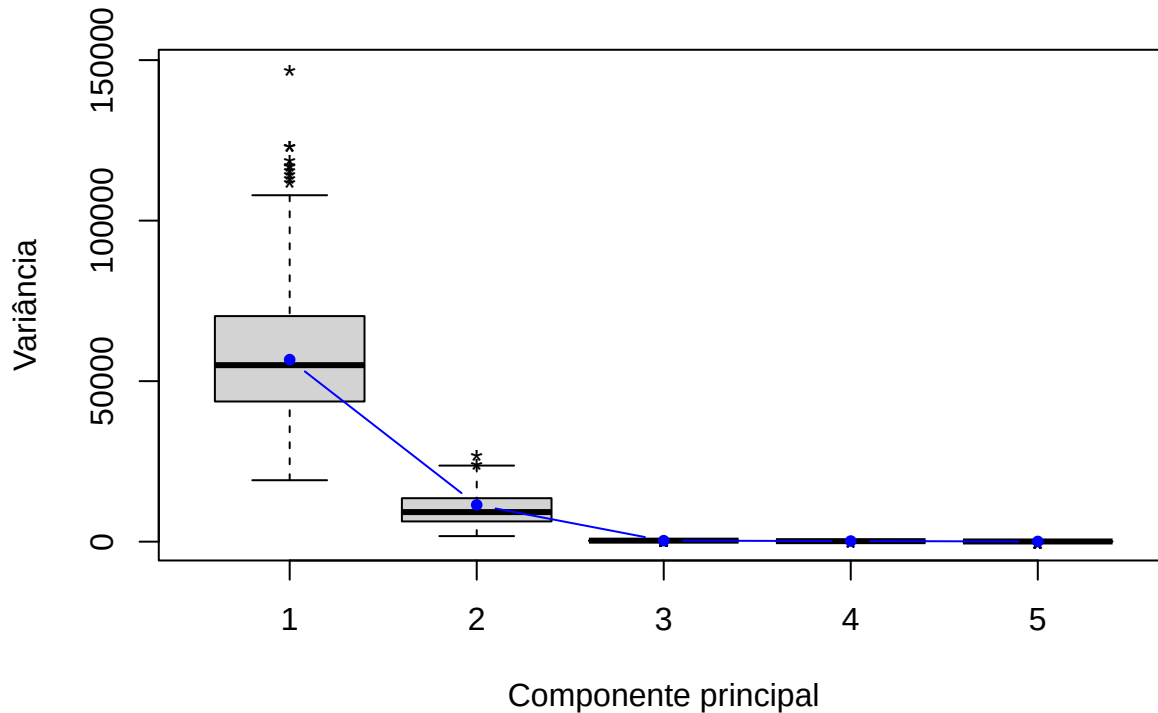
O primeiro componente principal é o único com variância maior do que a média das variâncias. O primeiro componente principal responde por cerca de 83% da variância total dos dados, ao passo que se tomarmos os dois primeiros componentes atingimos cerca de 99% da variância total. O gráfico abaixo (chamado de *scree plot*) indica que o número de componentes a reter pode ser 2.

```
plot(1:p, lambda, type = "b", xlab = "Componente principal", ylab = "Variância",
     pch = 20, col = "blue")
```



Utilizando reamostragem pelo método *bootstrap* com $B = 500$ amostras, obtivemos a figura abaixo, que traz informação sobre a variabilidade das estimativas das variâncias dos componentes principais.

```
B <- 500
lambdae <- matrix(NA, B, p)
for (b in 1:B) {
  iboot <- sample(n, n, replace = TRUE)
  acpe <- eigen(cov(dados[iboot, ]))
  lambdae[b,] <- acpe$values
}
boxplot(lambdae, pch = "*", xlab = "Componente principal", ylab = "Variância")
points(lambda, pch = 20, col = "blue", type = "b")
```



Entre um ou dois CP, continuamos o exemplo com dois CP ($k = 2$).

```
k <- 2
```

As estimativas dos coeficientes dos componentes principais são os autovetores da matriz de covariâncias amostral. Os k primeiros são mostrados abaixo.

```
print(as.data.frame(cbind(setor, round(E[, 1:k], 5))))
```

```
##          setor    CP1    CP2
## 1   Industrial 0.91942 0.39302
## 2    ComServ 0.23657 -0.58065
## 3      Rural 0.00524 -0.01826
## 4  Residencial 0.30295 -0.68615
## 5 IluminServPubOutros 0.08302 -0.19299
```

Como os elementos de $\hat{e}_1$ têm o mesmo sinal, podemos interpretar o primeiro componente principal como sendo um índice de consumo municipal de energia elétrica em GWh no ano de 2019 (“tamanho” do consumo). Como $\hat{e}_{21} > 0$ e $\hat{e}_{22}, \dots, \hat{e}_{25} < 0$, o segundo componente principal pode ser interpretado como a diferença entre o setor industrial e os demais setores, ou seja, a segunda principal fonte de variabilidade no consumo de energia elétrica deve-se à diferença entre consumo industrial e os demais setores de consumo.

Nota 1. Se $\hat{e}_j$ é um autovetor normalizado, $-\hat{e}_j$ também o é. Neste exemplo, se todos os elementos de $\hat{e}_1$ tivessem sinal negativo (depende da implementação computacional), poderíamos usar $-\hat{e}_1$ e manter a interpretação do CP_1 .

Exercício 1. Se $\hat{e}_j$ e $\hat{e}_m$, $j \neq m$, são dois autovetores, é possível que os elementos de cada par $(e_{j\ell}, e_{m\ell})$, $\ell = 1, \dots, p$, tenham o mesmo sinal?

As estimativas do coeficiente de correlação entre os dois primeiros componentes principais e as variáveis são dadas abaixo. O primeiro componente principal tem correlação forte com o consumo industrial, ao passo que o segundo componente está mais associado com os consumos residencial e de comércio e serviços.

```
corYX <- sqrt(lambda[1:k]) * t(E[, 1:k]) / sqrt(diag(mcov)[1:k])
colnames(corYX) <- setor
```

```
print(corYX, digits = 3)
```

```
##      Industrial ComServ   Rural Residencial IluminServPubOutros
## CP1      0.982    0.253  0.0056      0.324              0.0887
## CP2      0.496   -0.733 -0.0230     -0.866             -0.2435
```

Para cada observação i , os escores dos componentes principais são calculados como

$$\widehat{\mathbf{Y}}_i = (\widehat{Y}_{i1}, \dots, \widehat{Y}_{ip})^\top, \quad \widehat{Y}_{ij} = \widehat{\mathbf{e}}_j^\top \mathbf{X}_i = \sum_{l=1}^p \widehat{e}_{jl} X_{il}, \quad j = 1, \dots, p, \quad i = 1, \dots, n. \quad (1)$$

De outra forma, tomando $\mathbf{E} = [\widehat{\mathbf{e}}_1 \dots \widehat{\mathbf{e}}_p]$, obtemos

$$\widehat{\mathbf{Y}}_i = \mathbf{E}^\top \mathbf{X}_i, \quad i = 1, \dots, n. \quad (2)$$

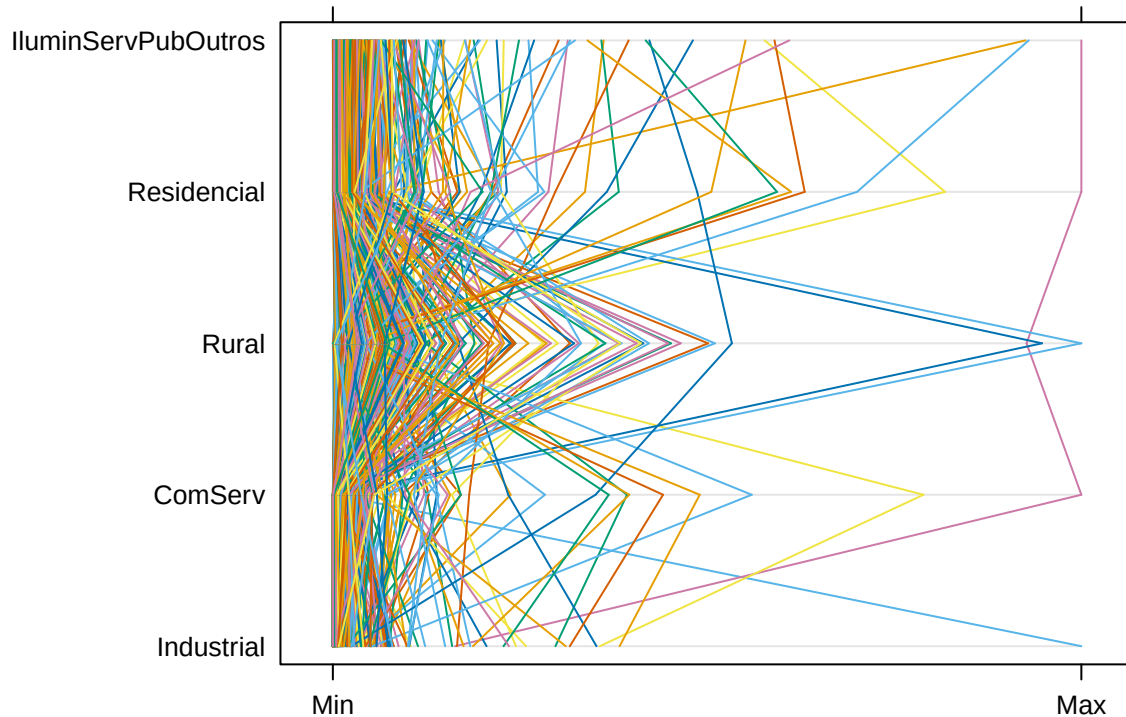
Nota 2. Se utilizarmos variáveis padronizadas, devemos usar $\mathbf{Z}_i$ e Z_{il} no lugar de $\mathbf{X}_i$ e X_{il} em (1) e (2). Além disso, os autovetores são obtidos da matriz de correlações amostral.

```
escores <- matrix(NA, n, p)
for (i in 1:n) {
  escores[i, ] <- t(E) %*% as.numeric(dados[i, ])
}
colnames(escores) <- colnames(E)
```

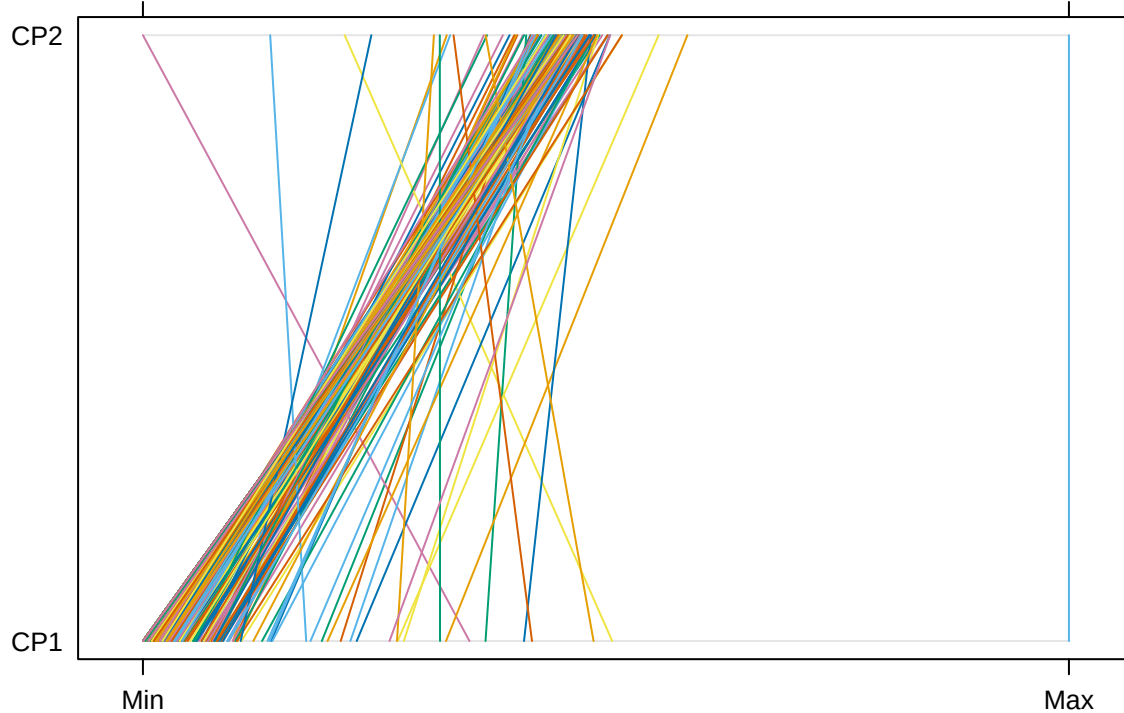
Exercício 2. Qual o resultado do comando `cor(escores)`? Surpresa?

Os gráficos de coordenadas paralelas dos dados de consumo e dos escores dos dois primeiros CP são mostrados abaixo. O gráfico com os CP permite uma distinção mais nítida entre os municípios.

```
library(lattice)
parallelplot(dados)
```



```
parallelplot(escores[, 1:k])
```



O município mais à direita no gráfico acima é Alumínio (número 12). O município mais à esquerda nos escores de CP_2 é Campinas (número 105).

```
(imax <- which.max(escores[, 1]))
```

```
## [1] 12
```

```
munic[imax]
```

```
## [1] "Aluminio"
```

```
(imin <- which.min(escores[, 2]))
```

```
## [1] 105
```

```
munic[imin]
```

```
## [1] "Campinas"
```

Tendo em vista que o primeiro componente principal pode ser visto como um indicador de consumo de energia elétrica, utilizamos os escores deste componente para ordenar os municípios.

```
escores1 <- escores[, 1]
names(escores1) <- munic
ordem <- order(escores1, decreasing = TRUE)
plot(escores1[ordem][1:20], type = "h", axes = FALSE, lwd = 2, col = "blue",
     xlab = "", ylab = "índice de consumo (GWh)")
axis(2)
axis(1, 1:20, munic[ordem][1:20], las = 2, cex.axis = 0.4)
box()
```

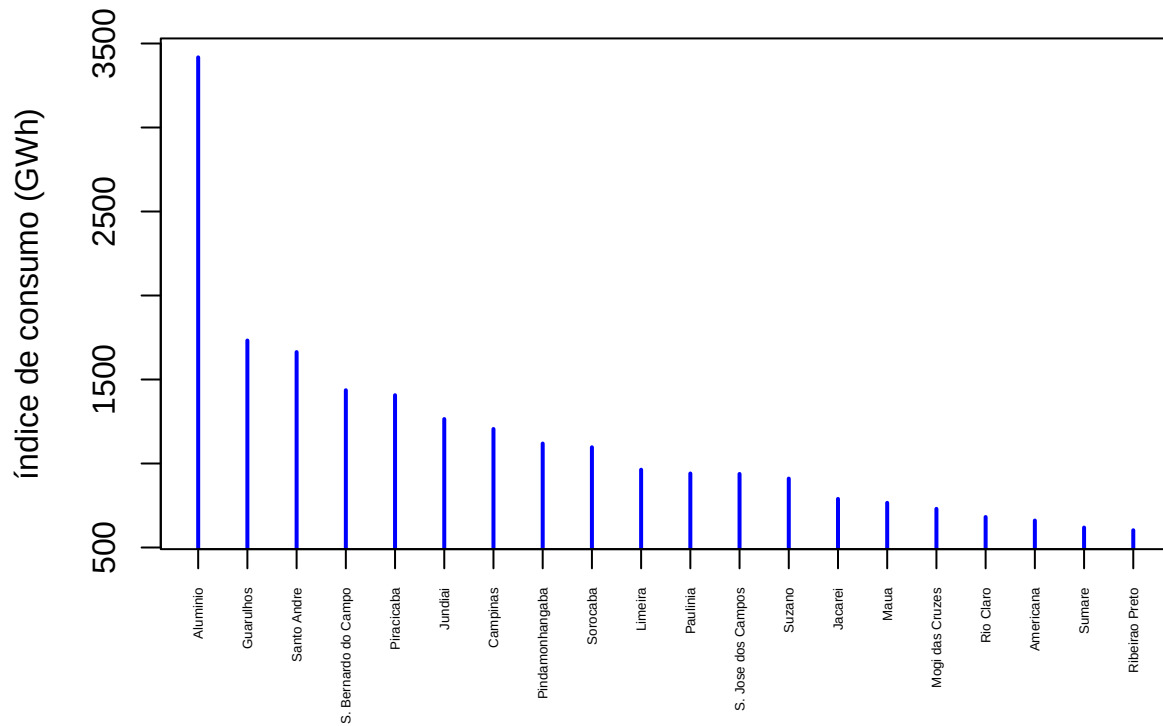
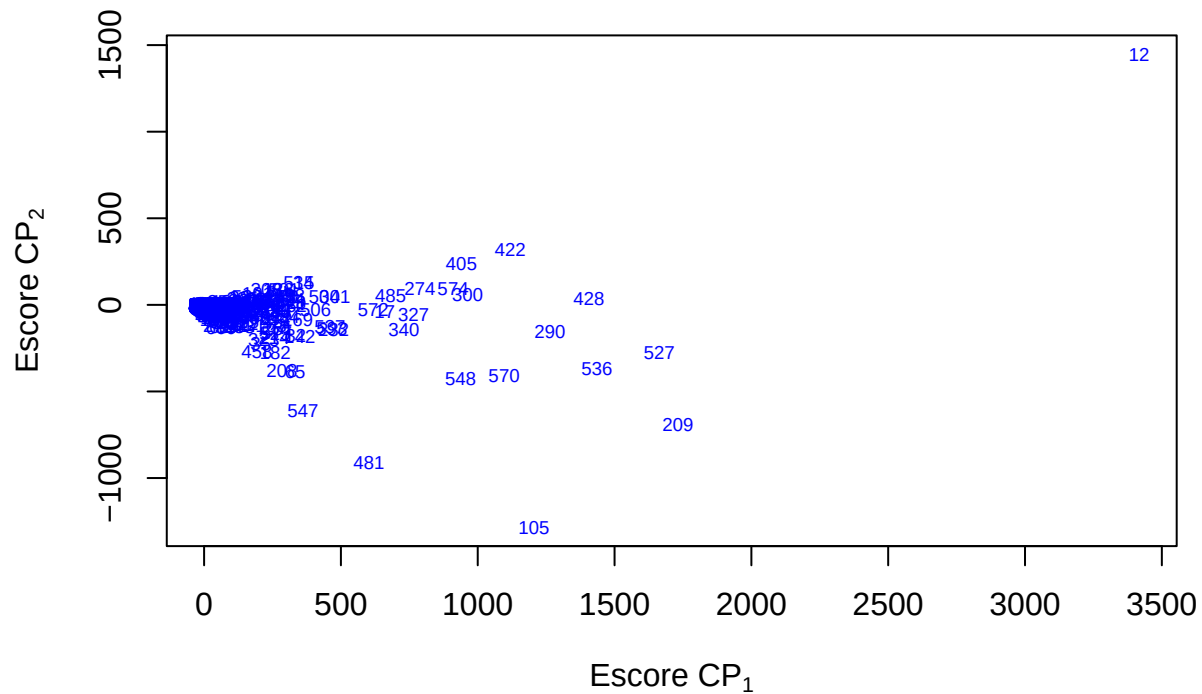


Gráfico de dispersão de pares de componentes principais podem ser úteis na identificação de observações aberrantes (*outliers*).

```
plot(escores[, 1], escores[, 2], col = "blue", pch = "",
     xlab = expression(paste("Escore ", CP[1])),
     ylab = expression(paste("Escore ", CP[2])),
     text(escores[, 1], escores[, 2], 1:n, col = "blue", cex = 0.6))
```



Vemos no gráfico acima que o município 12 (Alumínio) tem contribuição destacada para os dois primeiros CP. Este município corresponde ao maior consumo industrial, que corresponde ao elemento com maior valor no

```
which.max(dados[, "Industrial"])
```

No gráfico abaixo (dois últimos componentes principais) o município 12 não se destaca tanto (vimos acima que sua contribuição é forte para os dois primeiros CP). Os municípios 105, 222, 246 e 317 têm mais destaque no gráfico abaixo. O passo seguinte seria analisar o comportamento das cinco variáveis nestes municípios e também dos coeficientes de CP_4 e CP_5 ($\hat{e}_4$ e $\hat{e}_5$). O que muda se estes municípios forem excluídos?

Exercício 3. Refaça o exemplo usando a matriz de correlações amostral.