Análise de Componentes Principais

Francisco Rubens

## Análise descritiva

Fazendo primeiro a análise descritiva dos dados.

data(iris)  
head(iris, 3)

Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
1 5.1 3.5 1.4 0.2 setosa  
2 4.9 3.0 1.4 0.2 setosa  
3 4.7 3.2 1.3 0.2 setosa

diris=iris[,1:4]  
(summary(diris))

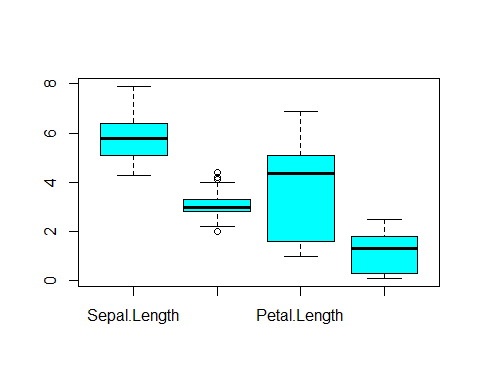
Sepal.Length Sepal.Width Petal.Length Petal.Width   
 Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100   
 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300   
 Median :5.800 Median :3.000 Median :4.350 Median :1.300   
 Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199   
 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800   
 Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500

Acima tem-se a média, os quartis e os mínimos e máximos das variáveis.  
Nota-se que o comprimento da pétala possui uma distribuição bem assimétrica em torno da média, pois a mediana é 4,35 e a média é 3,76.

(varian=var(diris))

Sepal.Length Sepal.Width Petal.Length Petal.Width  
Sepal.Length 0.6856935 -0.0424340 1.2743154 0.5162707  
Sepal.Width -0.0424340 0.1899794 -0.3296564 -0.1216394  
Petal.Length 1.2743154 -0.3296564 3.1162779 1.2956094  
Petal.Width 0.5162707 -0.1216394 1.2956094 0.5810063

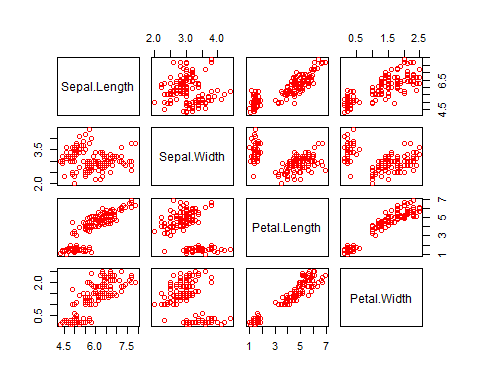
boxplot(diris,col="cyan")



(correl=cor(diris))

Sepal.Length Sepal.Width Petal.Length Petal.Width  
Sepal.Length 1.0000000 -0.1175698 0.8717538 0.8179411  
Sepal.Width -0.1175698 1.0000000 -0.4284401 -0.3661259  
Petal.Length 0.8717538 -0.4284401 1.0000000 0.9628654  
Petal.Width 0.8179411 -0.3661259 0.9628654 1.0000000

plot(diris,col="red")



Podemos ver pelos gráficos de correlação e a matriz de correlação uma correlação forte, principalmente, entre o comprimento da pétala com a largura da pétala (0,96), mas também vemos correlação forte entre comprimento da sépala com comprimento da pétala (0,87) e entre comprimento da sépala e largura da pétala (0,81).

## Análise de componentes principais

Começando a análise de componentes principais serão achados os autovalores e autovetores da matriz de covariância e da matriz de correlação.

(autov <- eigen(varian))

eigen() decomposition  
$values  
[1] 4.22824171 0.24267075 0.07820950 0.02383509  
  
$vectors  
 [,1] [,2] [,3] [,4]  
[1,] 0.36138659 -0.65658877 -0.58202985 0.3154872  
[2,] -0.08452251 -0.73016143 0.59791083 -0.3197231  
[3,] 0.85667061 0.17337266 0.07623608 -0.4798390  
[4,] 0.35828920 0.07548102 0.54583143 0.7536574

(autoc <- eigen(correl))

eigen() decomposition  
$values  
[1] 2.91849782 0.91403047 0.14675688 0.02071484  
  
$vectors  
 [,1] [,2] [,3] [,4]  
[1,] 0.5210659 -0.37741762 0.7195664 0.2612863  
[2,] -0.2693474 -0.92329566 -0.2443818 -0.1235096  
[3,] 0.5804131 -0.02449161 -0.1421264 -0.8014492  
[4,] 0.5648565 -0.06694199 -0.6342727 0.5235971

Os auvalores são as variâncias dos componentes principais. Para a matriz de covariância são:  
var() = 4,22824171  
var() = 0,24267075  
var() = 0,07820950  
var() = 0,02383509

Para a matriz de correlação:  
var() = 2,91849782  
var() = 0,91403047  
var() = 0,14675688  
var() = 0,02071484

Para se ter noção da proporção da variância das componentes:

autov$values[1:4]/sum(autov$values)

[1] 0.924618723 0.053066483 0.017102610 0.005212184

autoc$values[1:4]/sum(autoc$values)

[1] 0.729624454 0.228507618 0.036689219 0.005178709

Nota-se que para se ter ao menos 99% da proporção da variância das componentes, não é necessária a componente principal 4 para os dois métodos.  
O método usando a matriz de correlação é o método quando padroniza as variáveis, pois padronizando as variáveis a matriz de covariância é igual a matriz de correlação.

padiris=scale(diris)  
(varianp=var(padiris))

Sepal.Length Sepal.Width Petal.Length Petal.Width  
Sepal.Length 1.0000000 -0.1175698 0.8717538 0.8179411  
Sepal.Width -0.1175698 1.0000000 -0.4284401 -0.3661259  
Petal.Length 0.8717538 -0.4284401 1.0000000 0.9628654  
Petal.Width 0.8179411 -0.3661259 0.9628654 1.0000000

correl-varianp

Sepal.Length Sepal.Width Petal.Length Petal.Width  
Sepal.Length -2.220446e-16 0 0.000000e+00 1.110223e-16  
Sepal.Width 0.000000e+00 0 0.000000e+00 0.000000e+00  
Petal.Length 0.000000e+00 0 1.110223e-16 0.000000e+00  
Petal.Width 1.110223e-16 0 0.000000e+00 1.110223e-16

É possível usar a função prcomp para simplificar o processo de análise de componentes principais.

(cp=prcomp(diris)) # sem padronização (usa matriz de covariância)

Standard deviations (1, .., p=4):  
[1] 2.0562689 0.4926162 0.2796596 0.1543862  
  
Rotation (n x k) = (4 x 4):  
 PC1 PC2 PC3 PC4  
Sepal.Length 0.36138659 -0.65658877 0.58202985 0.3154872  
Sepal.Width -0.08452251 -0.73016143 -0.59791083 -0.3197231  
Petal.Length 0.85667061 0.17337266 -0.07623608 -0.4798390  
Petal.Width 0.35828920 0.07548102 -0.54583143 0.7536574

(cp1=prcomp(diris,scale=TRUE)) # com padronização (usa matriz de correlação)

Standard deviations (1, .., p=4):  
[1] 1.7083611 0.9560494 0.3830886 0.1439265  
  
Rotation (n x k) = (4 x 4):  
 PC1 PC2 PC3 PC4  
Sepal.Length 0.5210659 -0.37741762 0.7195664 0.2612863  
Sepal.Width -0.2693474 -0.92329566 -0.2443818 -0.1235096  
Petal.Length 0.5804131 -0.02449161 -0.1421264 -0.8014492  
Petal.Width 0.5648565 -0.06694199 -0.6342727 0.5235971

summary(cp)

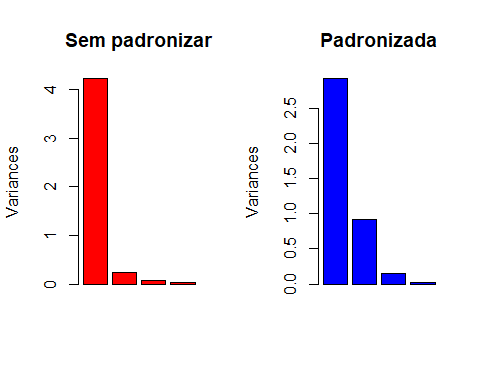
Importance of components:  
 PC1 PC2 PC3 PC4  
Standard deviation 2.0563 0.49262 0.2797 0.15439  
Proportion of Variance 0.9246 0.05307 0.0171 0.00521  
Cumulative Proportion 0.9246 0.97769 0.9948 1.00000

summary(cp1)

Importance of components:  
 PC1 PC2 PC3 PC4  
Standard deviation 1.7084 0.9560 0.38309 0.14393  
Proportion of Variance 0.7296 0.2285 0.03669 0.00518  
Cumulative Proportion 0.7296 0.9581 0.99482 1.00000

Nota-se que as proporções de variâncias são as mesmas achadas pelos autovalores, porém esse método retorna também a proporção acumulada. Com isso podemos ver a diferença entre padronizar ou não. E para esses dados podemos ver que, sem padronizar, aproximadamente 98% da variação são explicadas pelas componentes principais 1 e 2, e padronizando, aproximadamente 96% são explicadas pelas CP1 e CP2. Mas a principal diferença padronizando ou não está na CP1, que padronizando cai de 92,5% para 73%. O que faria diferença se fosse adotar o uso de componentes principais que expliquem ao menos 90% da variabilidade.

par(mfrow=c(1,2))  
plot(cp,col="red", main="Sem padronizar")  
plot(cp1,col="blue", main="Padronizada")



Pelos gráficos acima é possível perceber o impacto na proporção das variâncias quando se padroniza as variáveis. A CP1 continua tendo um grande impacto na variância total, mas a CP2 já tem um aumento expressivo quando se padroniza.