PCA，主成分分析是一种有效的降维方法，常用于高维数据的处理，如基因表达谱数据。本教程主要展示PCA在R语言中的实现和可视化。

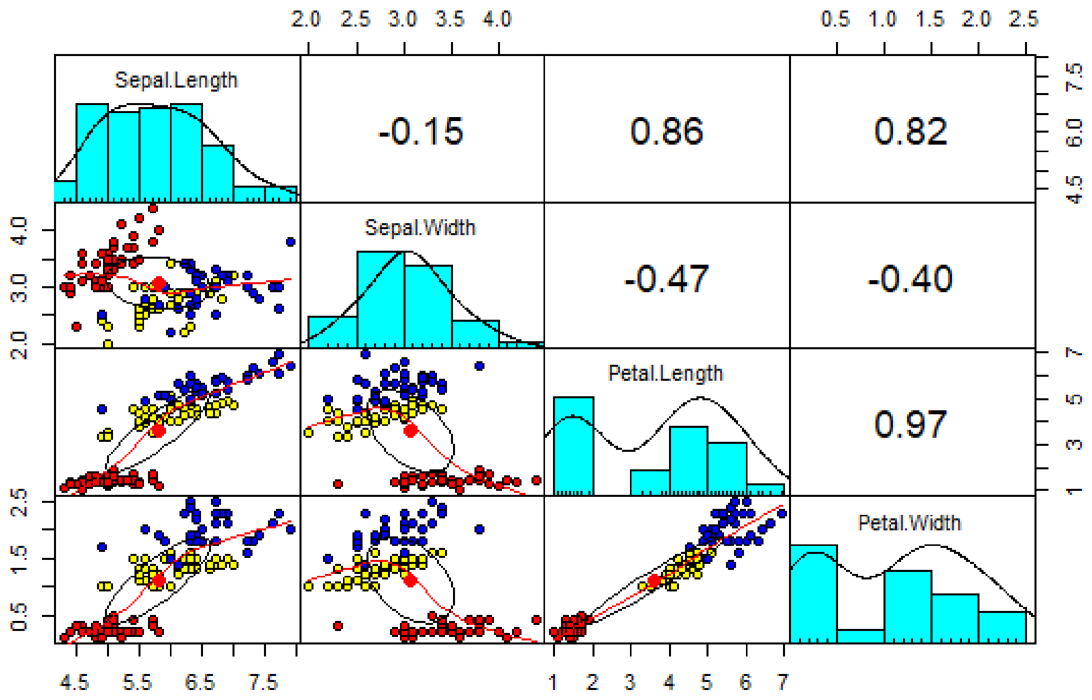
数据准备

所用数据为**iris**数据集。

```
# data
data(iris)
str(iris)

# partition data # 将数据划分成训练集和测试集，一把的分析不用划分
set.seed(111)
ind = sample(2, nrow(iris),replace = T, prob = c(.8,.2))
train = iris[ind == 1,]
test = iris[ind == 2,]

# scatter plot and correlations 绘图展示数据集变量之间的关系
library(psych)
pairs.panels(train[,-5],
             gap = 0,
             bg = c('red','yellow','blue')[train$Species],
             pch = 21)
```



PCA计算

```
# PCA
```

```
pc = prcomp(train[,-5],# 数据最后一列是分类变量，不选择
            center = T,# 数据中心化
            scale. = T) #数据标准化
pc$center
pc$scale
print(pc)

> print(pc)
Standard deviations (1, .., p=4):
[1] 1.7173318 0.9403519 0.3843232 0.1371332

Rotation (n x k) = (4 x 4):
                    PC1         PC2        PC3        PC4
Sepal.Length  0.5147163 -0.39817685  0.7242679  0.2279438
Sepal.Width  -0.2926048 -0.91328503 -0.2557463 -0.1220110
Petal.Length  0.5772530 -0.02932037 -0.1755427 -0.7969342
Petal.Width   0.5623421 -0.08065952 -0.6158040  0.5459403
```

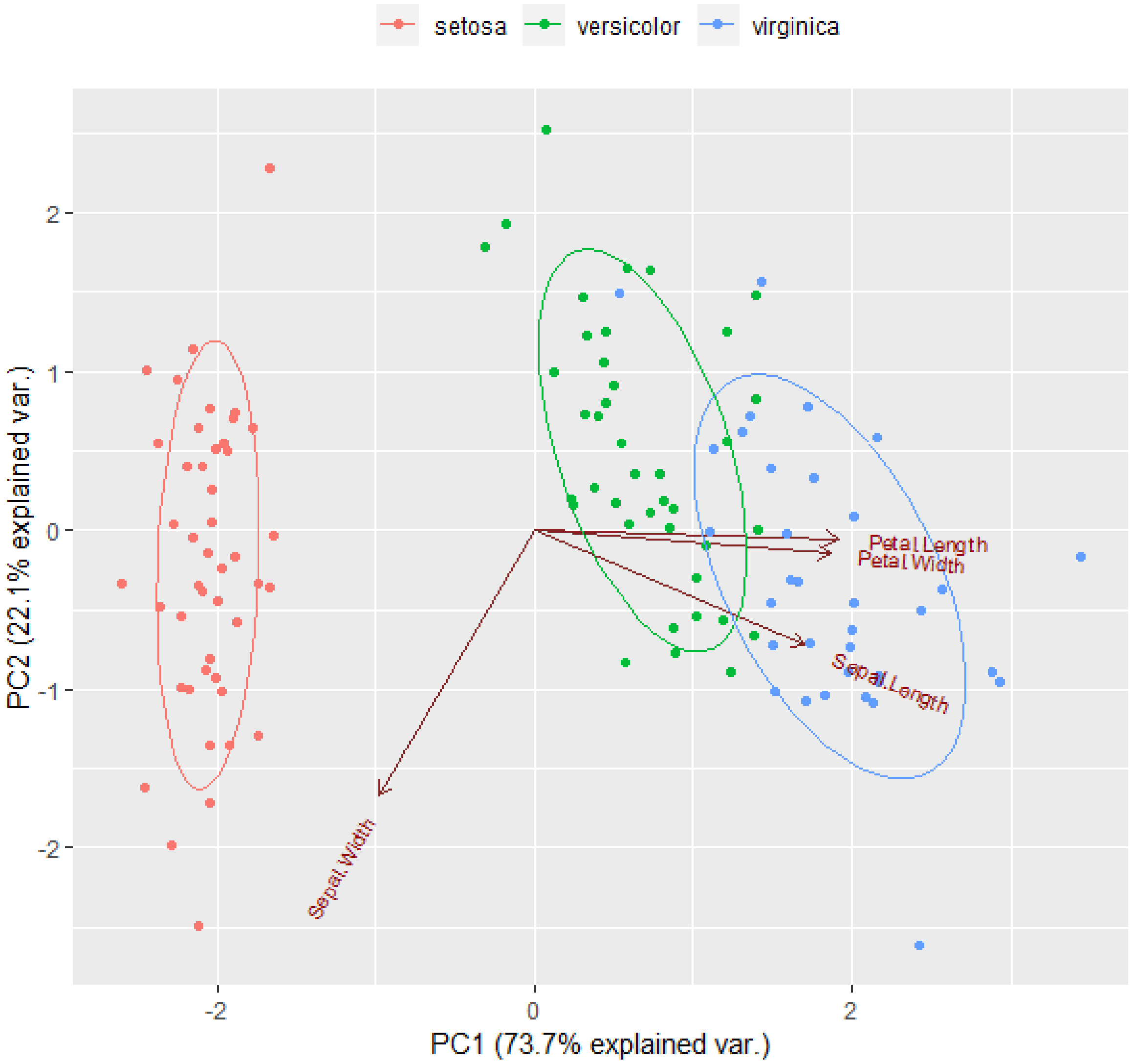
查看**PC**的解释百分百：

```
> summary(pc)
Importance of components:
                          PC1    PC2     PC3    PC4
Standard deviation     1.7173 0.9404 0.38432 0.1371
Proportion of Variance 0.7373 0.2211 0.03693 0.0047
Cumulative Proportion  0.7373 0.9584 0.99530 1.0000
```

直接看**Proportion of Variance**即可。

可视化

```
# bi-plot
library(ggbiplot)
g = ggbiplot(pc,
             obs.scale = 1,
             var.scale = 1,
             groups = train$Species,
             ellipse = T, # 置信区间椭圆
             circle = F,
             ellipse.prob = 0.68) # 椭圆覆盖多少数据点
g = g + scale_color_discrete(name = '') + theme(legend.direction = 'horizontal',
                                                legend.position = 'top')
print(g)
```



模型预测

一般是不会涉及到模型预测的。

```
# prediction with Principal Components
trg = predict(pc, train)
trg = data.frame(trg, train$Species)

tst = predict(pc,test)
tst = data.frame(tst, test$Species)

# multinomial logistic regression with first two PCs
library(nnet)
trg$species = relevel(trg$train.Species, ref = 'setosa')
mymodel = multinom(train.Species~PC1+PC2,data = trg)
summary(mymodel)

# Confusion matrix
p = predict(mymodel, tst)
tab = table(p, test$Species)
tab
```

```
1-sum(diag(tab))/sum(tab)
```

```
> tab

p            setosa versicolor virginica
  setosa          5          0         0
  versicolor      0          9         3
  virginica       0          1        12
> 1-sum(diag(tab))/sum(tab)
[1] 0.1333333
```

致谢： 感谢YouTube博主*Bharatendra Rai*博士的视频!