主成分分析结果可视化

庄闪闪

目录

1 简介

主成分分析法，也被称为主分量分析法，是很常用的一种数据降维方法。采用一个线性变换将数据变换到一个新的坐标系统，使得任何数据点投影到第一个坐标（成为第一主成分）的方差最大，在第二个坐标（第二主成分）的方差为第二大，以此类推。因此，主成分分析可以减少数据的维数，并保持对方差贡献最大的特征，相当于保留低阶主成分，忽略高阶主成分。

关于主成分的理论介绍和 R 语言代码实现可见前段时间赵西西写的推文：

但是后面留了一个小尾巴，如果想对主成分结果进行可视化，那得怎么实现？有没有简便的方法呢？

正好这几天有读者问起，那今天就来说说这个问题吧。

2 方法一

使用 ggbiplot 包中的 ggbiplot() 函数，该函数使用 ggplot2 对主成分进行可视化。函数内部参数如下

```
ggbiplot(pcobj, choices = 1:2, scale = 1, pc.biplot =
TRUE, obs.scale = 1 - scale, var.scale = scale, groups =
NULL, ellipse = FALSE, ellipse.prob = 0.68, labels =
NULL, labels.size = 3, alpha = 1, var.axes = TRUE, circle
= FALSE, circle.prob = 0.69, varname.size = 3,
varname.adjust = 1.5, varname.abbrev = FALSE, ...)
```

内部参数过多，就不做详细解释。如果对内部参数有兴趣可以通过帮助文档进行查询（?ggbiplot）。

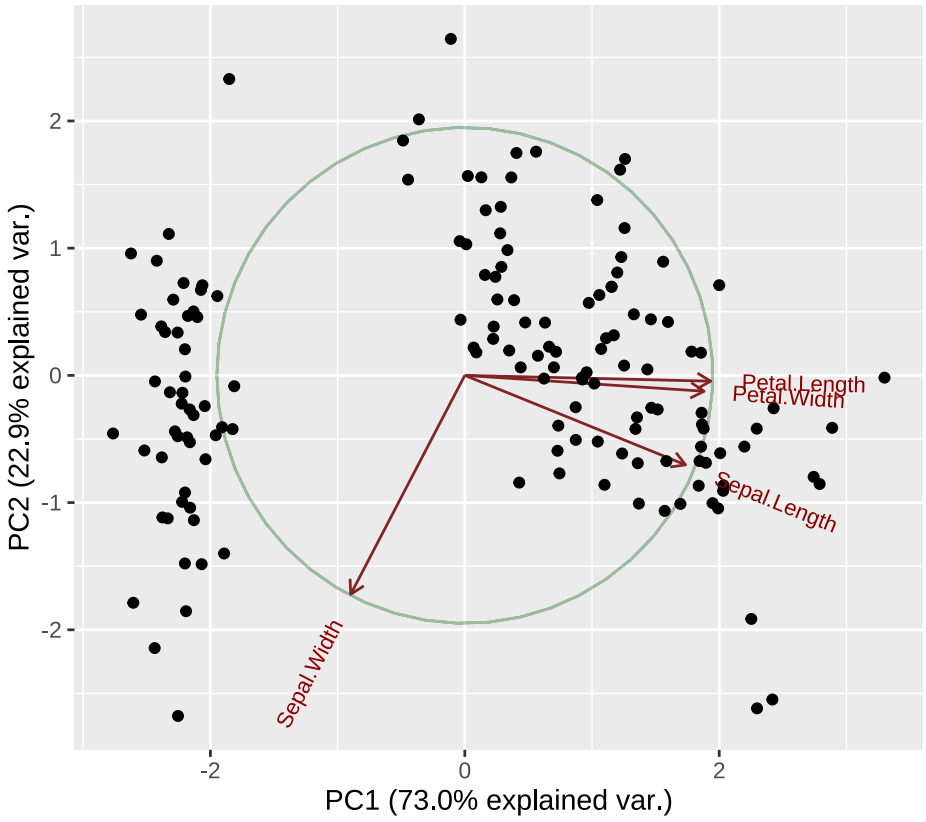
这里使用鸢尾花数据，给出一个简单的例子。大家可以将自己的数据进行导入（如何导入？可见推文：xxx），替换鸢尾花数据。

注意：检查自己数据集的数据结构是否和鸢尾花数据结构一致

这个包在 github 中，官方说可以使用以下参数进行下载（但是小编下载不了，只能通过强暴的方法进行，具体可见推文:。该压缩包已经处理成 tar.gz 放到公众号内了，如有需要，后台回复 [ggbiplot] 即可获得）。

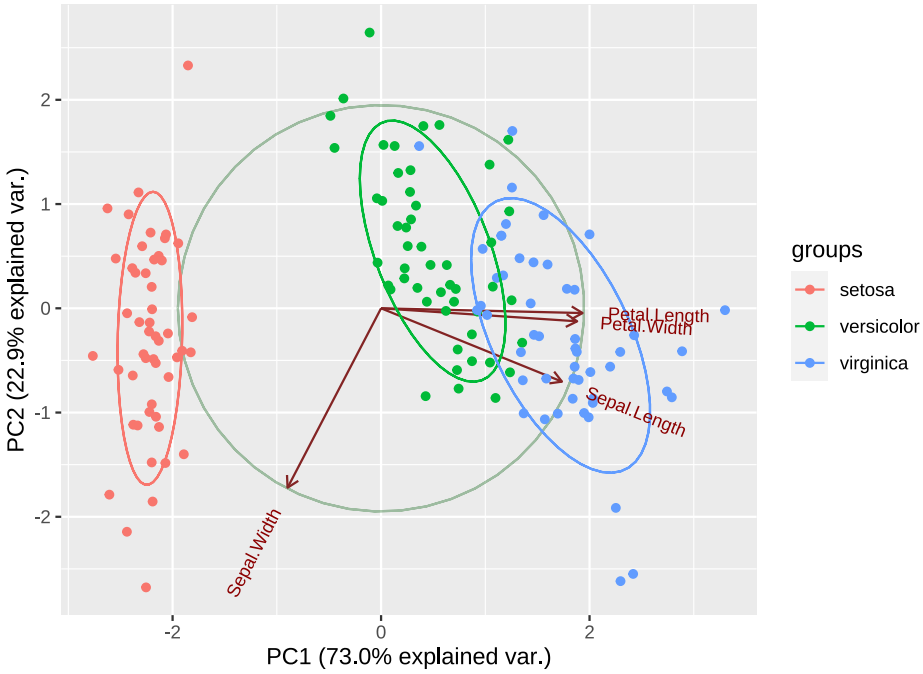
之后使用 `prcomp()` 进行主成分分析，然后将结果保存到 res.pca 变量中。之后使用 `ggbiplot()` 进行可视化。其中观测的尺度因子为 1（`obs.scale = 1`），变量的尺度因子为 1（`var.scale = 1`），每组绘制一个椭圆（`ellipse = TRUE`）并添加相关系数的圆。

```
# install_github("vqv/ggbiplot")
library(ggbiplot)
res.pca <- prcomp(iris[, -5],  scale = TRUE)
ggbiplot(res.pca, obs.scale = 1, var.scale = 1, ellipse = TRUE, circle = TRUE)
```



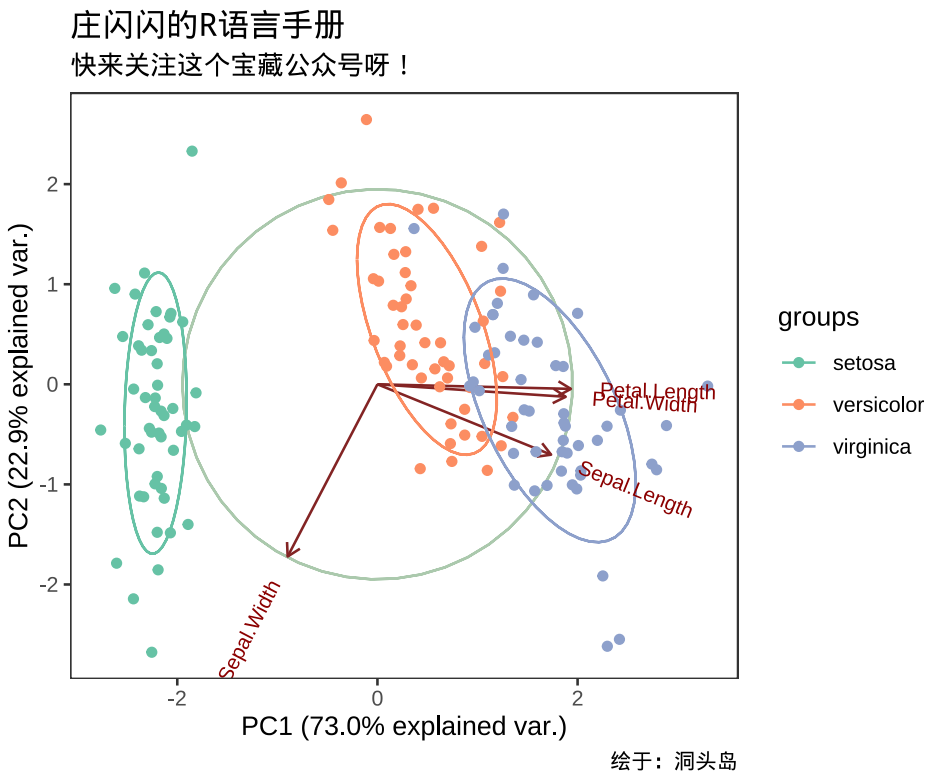
如果想给不同组别添加分别显示不同颜色，则可以使用参数 groups，然后设定为原始数据对应的组别向量（如果原始数据没有，可以自行构造一个向量。）

```
# 添加组别颜色
ggbiplot(res.pca, obs.scale = 1, var.scale = 1, ellipse = TRUE,groups = iris$Species, c
```



当然你可以在此基础上加入 ggplot 内部的参数，比如更改主题，更改颜色，添加标题等一系列操作。

```
# 更改主题
ggbiplot(res.pca, obs.scale = 1, var.scale = 1, ellipse = TRUE,groups = iris$Species, c
  theme_bw() +
  theme(panel.grid = element_blank()) +
  scale_color_brewer(palette = "Set2") +
  labs(title = " 庄闪闪的 R 语言手册",subtitle = " 快来关注这个宝藏公众号呀! ",caption =" 绘
```



小编最近有幸上了两节线上的 R 语言数据可视化公益课，把 R 语言 base 包以及 ggplot 语法系统的过了一遍，如果有需要补补可视化基础的朋友，可移步我的 b 站 [账号名：庄闪闪]，视频回放已等候多时了。

3 方法二

使用 `FactoMineR` 包的 `PCA()` 函数或者使用基础包的 `prcomp()` 函数进行数据降维处理，然后使用 `factoextra` 包的 `fviz_pca_ind()` 函数对结果进行可视化。

这里还是以鸢尾花的数据作为例子，沿用方法一的主成分分析结果 res.pca。

这个包内部有四个主要绘制主成分结果的函数。

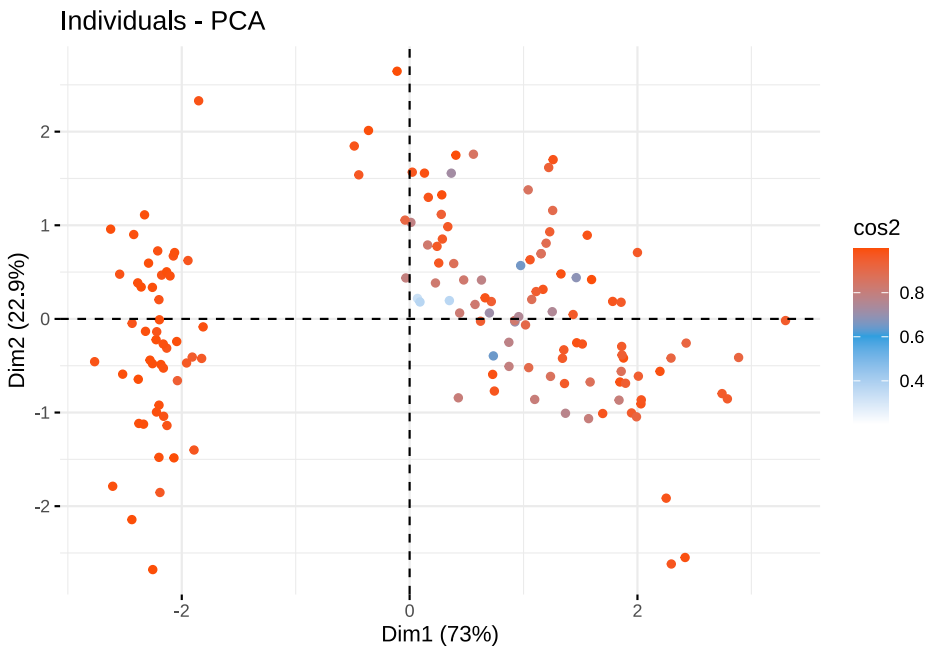
- fviz_pca_ind(): 各个样本图

- fviz_pca_var(): 变量图
- fviz_pca_biplot(): 各个样本和变量的联合图
- fviz_pca(): fviz_pca_biplot() 的别名

内部参数不做过多介绍，有兴趣的读者请看帮助文档。这里只对下面的代码中出现的参数进行解释。

使用散点图进行绘制(`geom = "point"`),颜色使用“cos2”(`col.ind="cos2"`),使用 3 阶梯度颜色 (`gradient.cols = c("white", "#2E9FDF", "#FC4E07" )`)。

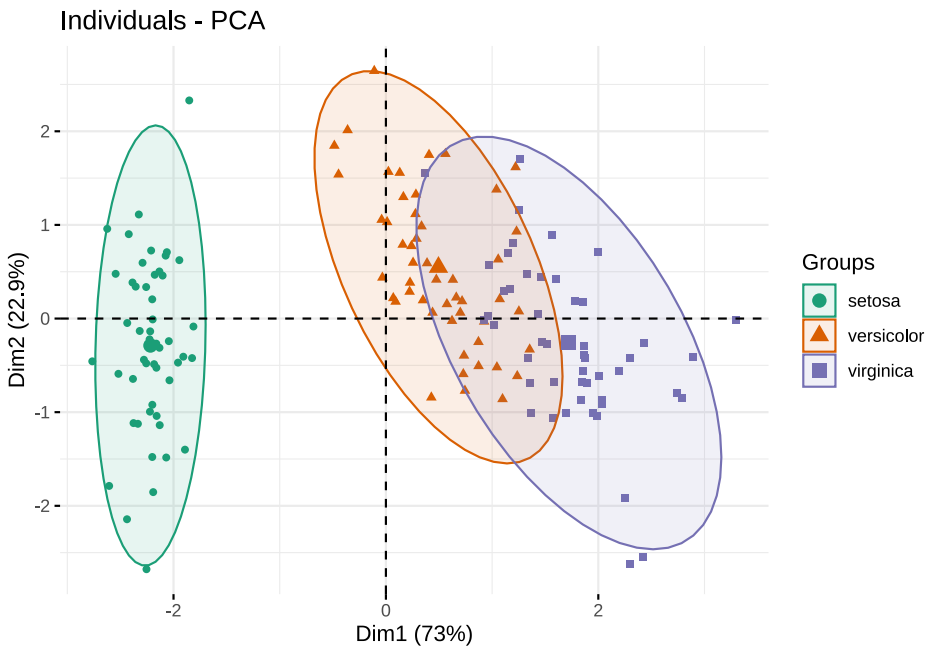
```
library(ggplot2)
library(factoextra) # 可以直接通过 install.packages() 进行下载
# 各个样本图
fviz_pca_ind(res.pca, col.ind="cos2", geom = "point",
             gradient.cols = c("white", "#2E9FDF", "#FC4E07" ))
```



如果想展示分组变量信息，可以通过 habillage 参数设定，和第一种方法类似，这里还加入了一些细节：各组添加椭圆（`addEllipses=TRUE`），图的版

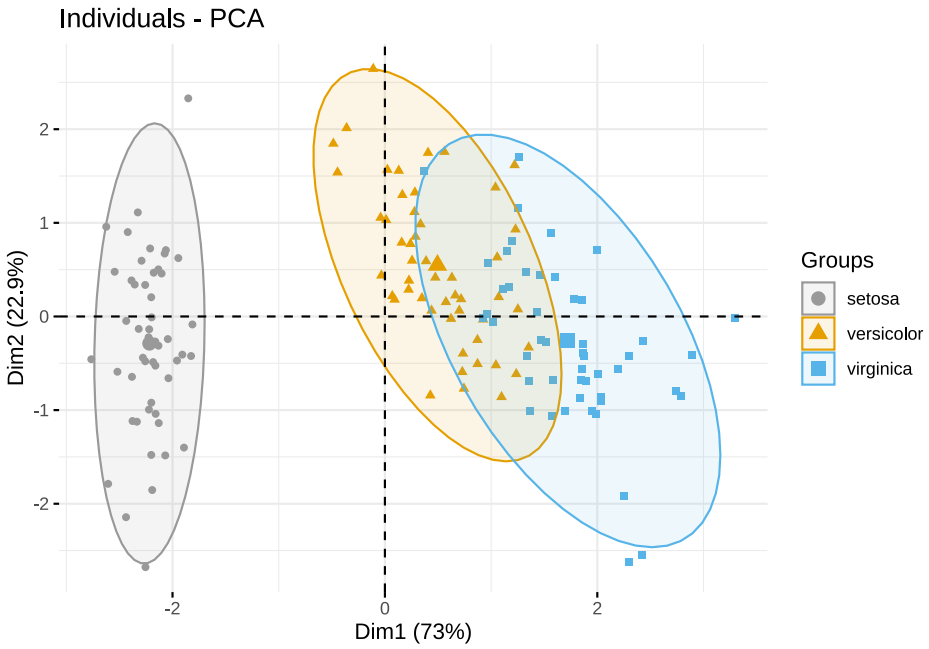
式使用 “Dark2”（palette = "Dark2"）。

```
fviz_pca_ind(res.pca, label="none", habillage=iris$Species,
             addEllipses=TRUE, ellipse.level=0.95, palette = "Dark2")
```



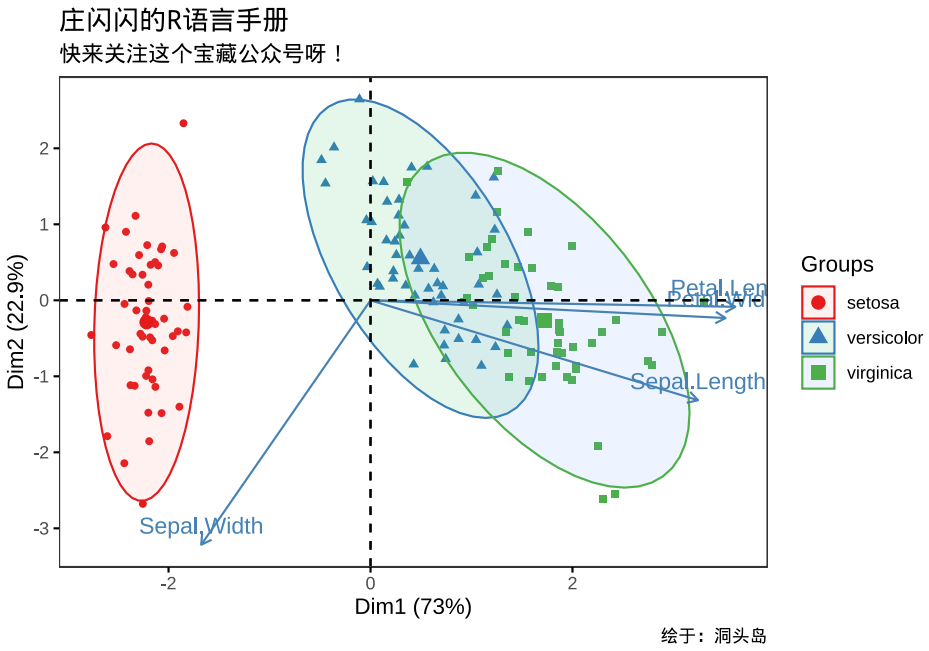
当然可以使用 palette = c("#999999", "#E69F00", "#56B4E9")，根据论文全文配色，进行手动调整。

```
fviz_pca_ind(res.pca, label="none", habillage=iris$Species,
             addEllipses=TRUE, ellipse.level=0.95,
             palette = c("#999999", "#E69F00", "#56B4E9"))
```



如果想绘制个体和变量的双图，可以使用 `fviz_pca_biplot()`，内部其他参数构造相同，然后可以添加各种其他 ggplot 的函数。

```
# 个体和变量的双图
# 只保留变量的标签 # 按组改变颜色，添加省略号
fviz_pca_biplot(res.pca, label = "var", habillage=iris$Species,
                addEllipses=TRUE, ellipse.level=0.95,
                ggtheme = theme_bw()) +
  theme(panel.grid = element_blank()) +
  scale_color_brewer(palette = "Set1") +
  labs(title = " 庄闪闪的 R 语言手册",subtitle = " 快来关注这个宝藏公众号呀! ",caption =" 给
```


庄闪闪的R语言手册
快来关注这个宝藏公众号呀！
Dim2 (22.9%)
Dim1 (73%)
Sepal.Width
Sepal.Length
Groups
setosa
versicolor
virginica
绘于：洞头岛