# PCA 和降维分析

## HUST AIBIO course series for undergraduates

Wei-Hua Chen (CC BY-NC 4.0)

13 April, 2021

# Table of Contents

# Section 1: PCA

# 什么是 PCA？

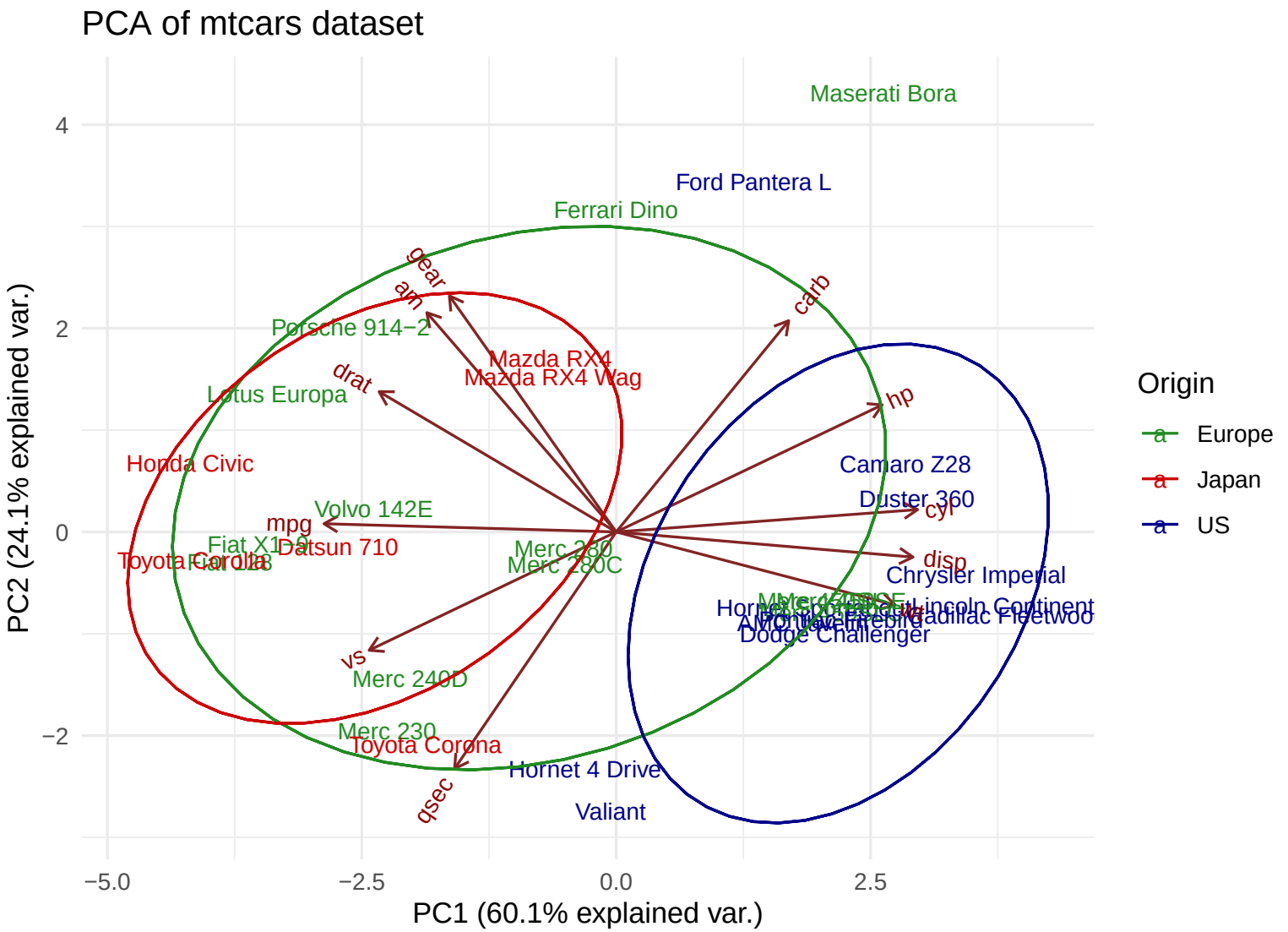
Principle component analysis (PCA) 是一种降维分析方法，主要用于简化、揭示**样本之间的关系**（距离），**并显示变量之间的关系及对样本分组的贡献度**。

PCA 通常用所谓的 biplot（双标图）来表示；Biplot is a generalization of the simple two-variable scatterplot; biplot allows information on both **samples** and **variables of a data matrix** to be displayed graphically.

这里用 `mtcars` 给出一个例子；先看一下数据：

| | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mazda RX4 | 21.0 | 6 | 160 | 110 | 3.90 | 2.620 | 16.46 | 0 | 1 | 4 | 4 |
| Mazda RX4 Wag | 21.0 | 6 | 160 | 110 | 3.90 | 2.875 | 17.02 | 0 | 1 | 4 | 4 |
| Datsun 710 | 22.8 | 4 | 108 | 93 | 3.85 | 2.320 | 18.61 | 1 | 1 | 4 | 1 |
| Hornet 4 Drive | 21.4 | 6 | 258 | 110 | 3.08 | 3.215 | 19.44 | 1 | 0 | 3 | 1 |
| Hornet Sportabout | 18.7 | 8 | 360 | 175 | 3.15 | 3.440 | 17.02 | 0 | 0 | 3 | 2 |
| Valiant | 18.1 | 6 | 225 | 105 | 2.76 | 3.460 | 20.22 | 1 | 0 | 3 | 1 |

# A PCA plot


PCA of mtcars dataset
PC2 (24.1% explained var.)
PC1 (60.1% explained var.)
4
2
0
−2
−5.0
−2.5
0.0
2.5
Maserati Bora
Ford Pantera L
Ferrari Dino
gear
am
carb
Porsche 914−2
Mazda RX4
Mazda RX4 Wag
drat
Lotus Europa
hp
Honda Civic
Camaro Z28
Duster 360
cyl
Volvo 142E
mpg
Fiat X1−9
Datsun 710
Toyota Corolla
Merc 280
Merc 280C
disp
Chrysler Imperial
Lincoln Continent
Cadillac Fleetwoo
Dodge Challenger
vs
Merc 240D
Merc 230
Toyota Corona
Hornet 4 Drive
qsec
Valiant
Origin
Europe
Japan
US

# 降维

`mtcars` 共有 11 维数据：

| | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mazda RX4 | 21.0 | 6 | 160 | 110 | 3.90 | 2.620 | 16.46 | 0 | 1 | 4 | 4 |
| Mazda RX4 Wag | 21.0 | 6 | 160 | 110 | 3.90 | 2.875 | 17.02 | 0 | 1 | 4 | 4 |
| Datsun 710 | 22.8 | 4 | 108 | 93 | 3.85 | 2.320 | 18.61 | 1 | 1 | 4 | 1 |

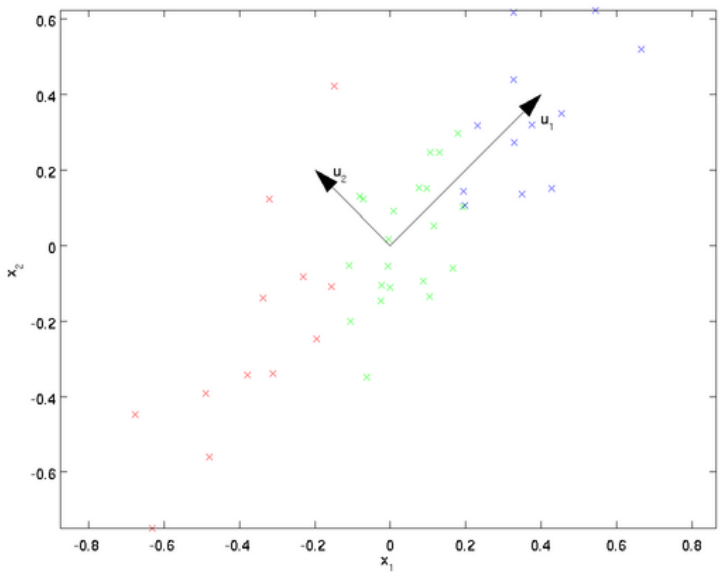
但通常情况下，人只能处理 2~3 维的数据。因此，理想情况是将**任何多维数据**转化为 **2~3 维**，以供人眼识别。

但操作中这很难实现。解决方法如下：

1. 创造新的、数量与**变量数一样的维度**（components）；
2. 使第一维尽可能多的捕捉整体数据的**变化（variation）**，**实现最大可分性**；并使第二 ... N 维依次捕捉独立于之前维度的独立变化；或者说，按维度的重要性（对样本 variation 的解释度）进行排序，最重要的排第一，依次排列；
3. 所有维度对样本解释度的总和为 100%；
4. 可以用最重要的 2~3 个维度做图，虽然这样会损失部分信息；损失程度与前几维的重要性有关。

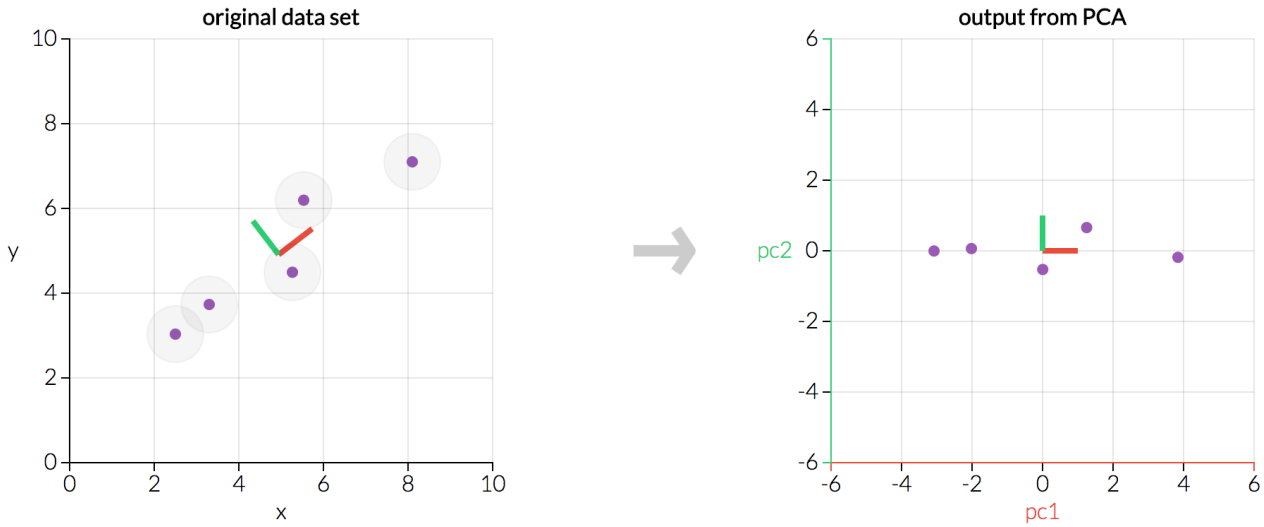
# 如何实现最大可分性？

假设我们有一个二维数据，可以用 x1 和 x2 表示：



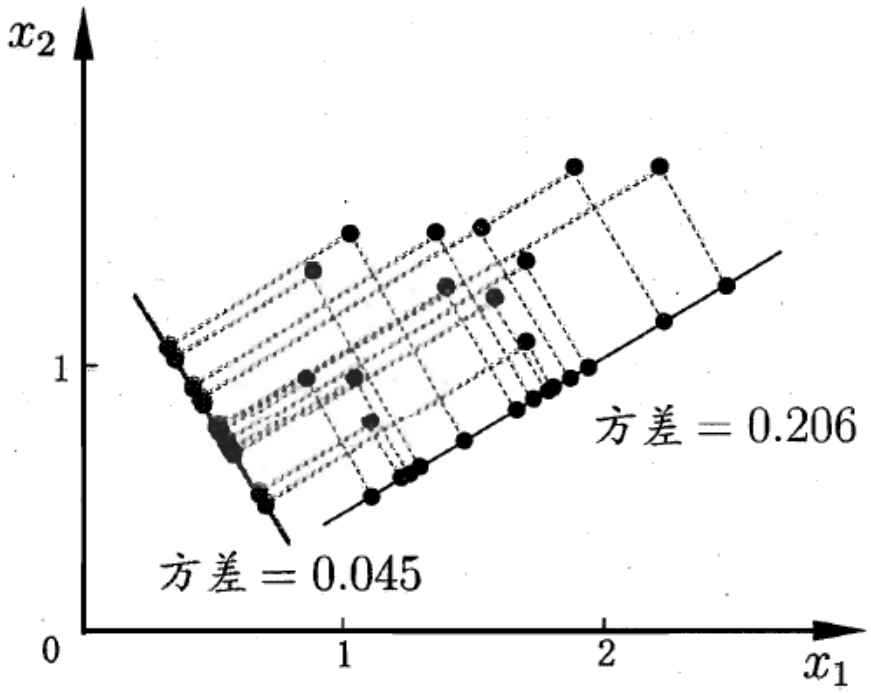

从直观上也可以看出，我们可以产生两个新的维度，以更好的表示当前数据；而且，u1 比 u2 好，因此前者可以作为第一维。这就是我们所说的**最大可分性**。

# 最大可分性


original data set
y
x
0
2
4
6
8
10
output from PCA
pc2
pc1
-6
-4
-2
0
2
4
6

# 降维

两个维度的重要性可以用数据点与维度的距离判定:




注: 插图来自知乎.

# 多维数据的降维？

一般来说，欲获得原始数据新的表示空间，最简单的是对原始数据进行线性变换（基变换）：

$$Y = PX$$

其中 $Y$ 是样本在新空间的表达，$P$ 是基向量，$X$ 是原始样本。我们可知选择不同的基可以对一组数据给出不同的表示，同时当基的数量少于原始样本本身的维数则可达到降维的效果，矩阵表示如下：

$$\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_R \end{pmatrix} \begin{pmatrix} a_1 & a_2 & \cdots & a_M \end{pmatrix} = \begin{pmatrix} p_1a_1 & p_1a_2 & \cdots & p_1a_M \\ p_2a_1 & p_2a_2 & \cdots & p_2a_M \\ \vdots & \vdots & \ddots & \vdots \\ p_Ra_1 & p_Ra_2 & \cdots & p_Ra_M \end{pmatrix}$$

注：详情见知乎页面.

# PCA 算法的优缺点

**优点：**

- 仅仅需要以方差衡量信息量，不受数据集以外的因素影响。
- 各主成分之间正交，可消除原始数据成分间的相互影响的因素。
- 计算方法简单，主要运算是特征值分解，易于实现。

**缺点：**

- 主成分各个特征维度的含义具有一定的模糊性，不如原始样本特征的解释性强。
- 方差小的非主成分也可能含有对样本差异的重要信息，因降维丢弃可能对后续数据处理有影响。

# PCA 计算方法与结果解读

| | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mazda RX4 | 21.0 | 6 | 160 | 110 | 3.90 | 2.620 | 16.46 | 0 | 1 | 4 | 4 |
| Mazda RX4 Wag | 21.0 | 6 | 160 | 110 | 3.90 | 2.875 | 17.02 | 0 | 1 | 4 | 4 |
| Datsun 710 | 22.8 | 4 | 108 | 93 | 3.85 | 2.320 | 18.61 | 1 | 1 | 4 | 1 |

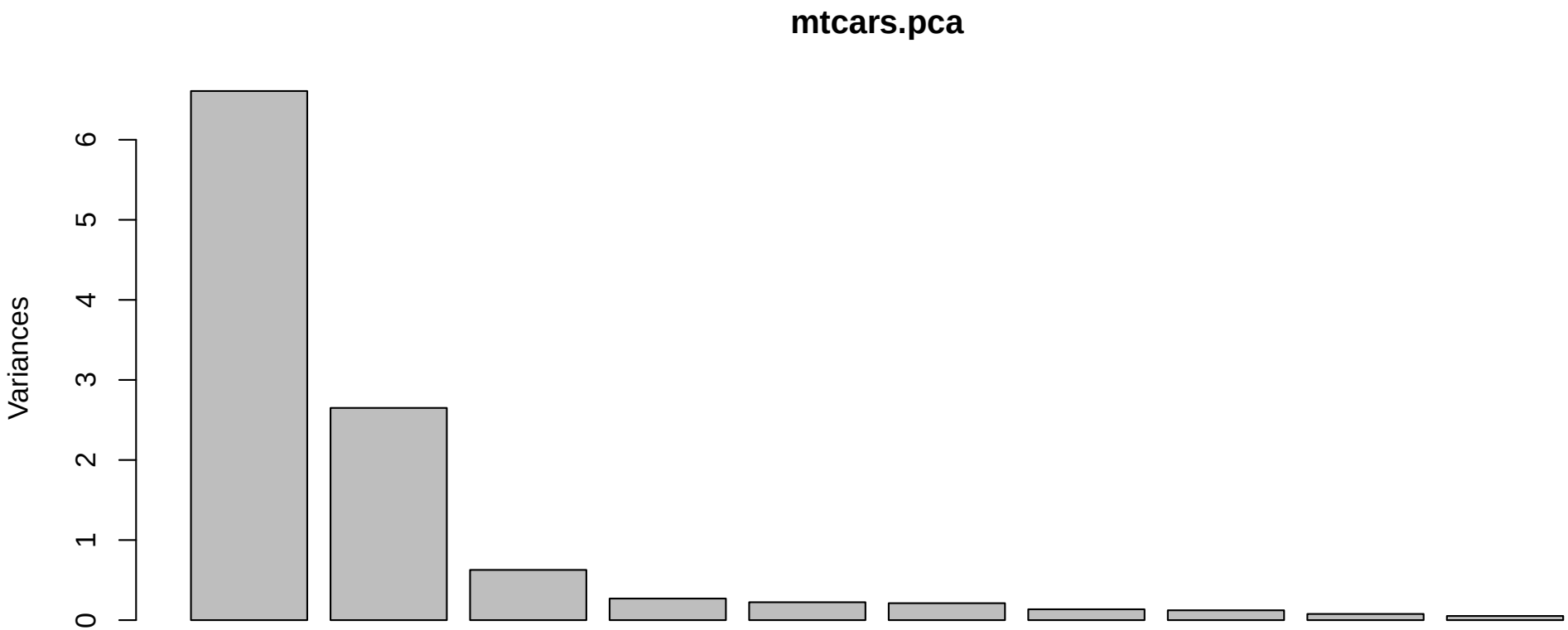
```
mtcars.pca <- prcomp(mtcars, scale = T);
names(mtcars.pca);
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

```
summary(mtcars.pca);
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6    PC7
## Standard deviation     2.5707 1.6280 0.79196 0.51923 0.47271 0.46000 0.3678
## Proportion of Variance 0.6008 0.2409 0.05702 0.02451 0.02031 0.01924 0.0123
## Cumulative Proportion  0.6008 0.8417 0.89873 0.92324 0.94356 0.96279 0.9751
##                           PC8    PC9    PC10   PC11
## Standard deviation     0.35057 0.2776 0.22811 0.1485
## Proportion of Variance 0.01117 0.0070 0.00473 0.0020
## Cumulative Proportion  0.98626 0.9933 0.99800 1.0000
```

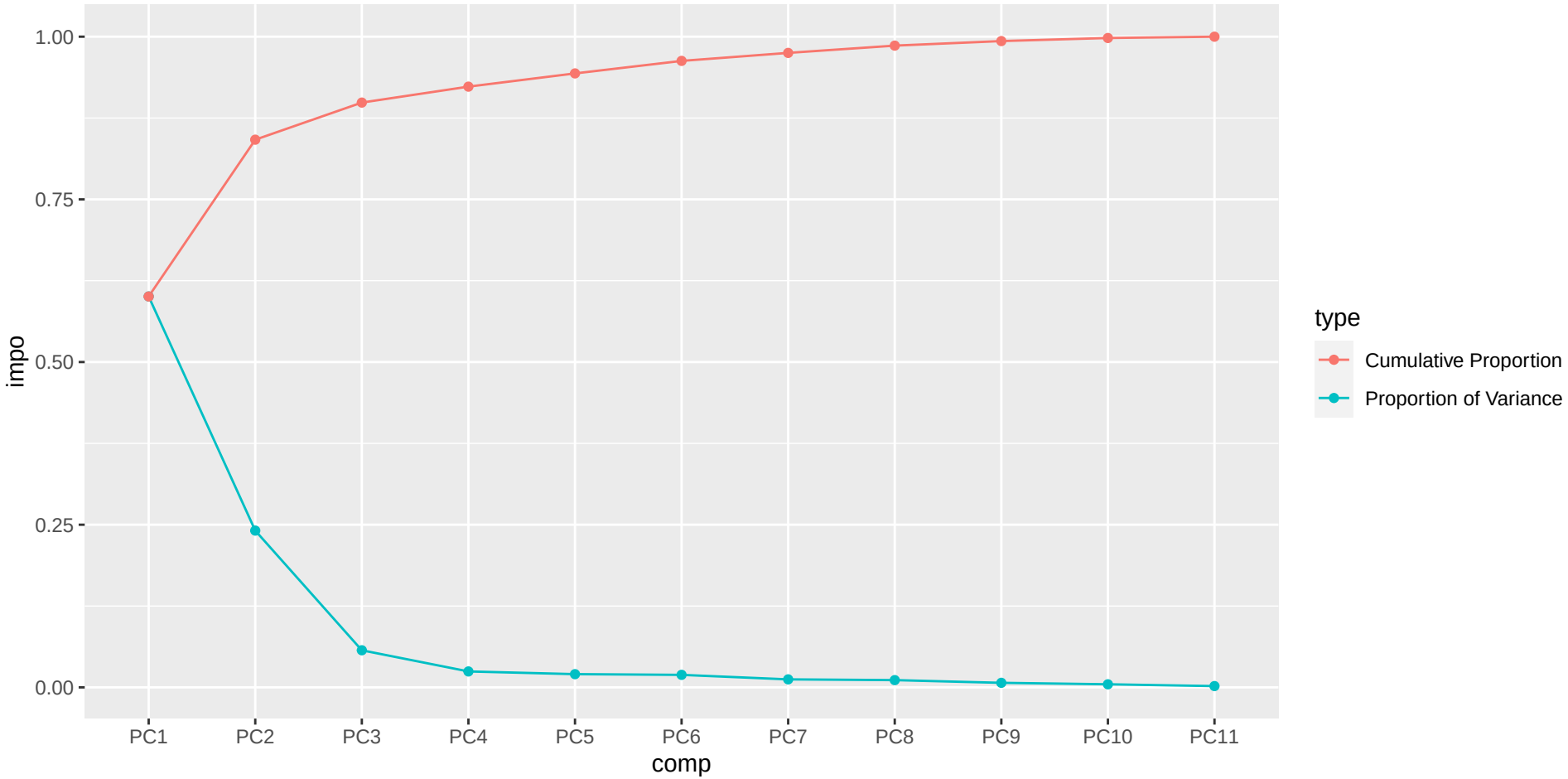
# Variance explained by components



# Plot both variance and cumulative variance

```
## -- let's first get all data ready --
impor <- summary( mtcars.pca )$importance[2:3,]; ## get importance
impor2 <- data.frame( impor, type = rownames(impor) );

## -- transform
require(tidyverse);
dat <- impor2 %>% gather( comp, impo, -type );
dat$comp <- factor(dat$comp, levels = colnames(impor));
plot3 <-
  ggplot( dat, aes( comp, impo, group = type, color = type ) ) +
  geom_line() +
  geom_point();
```

# Plot both variance and cumulative variance, cont.

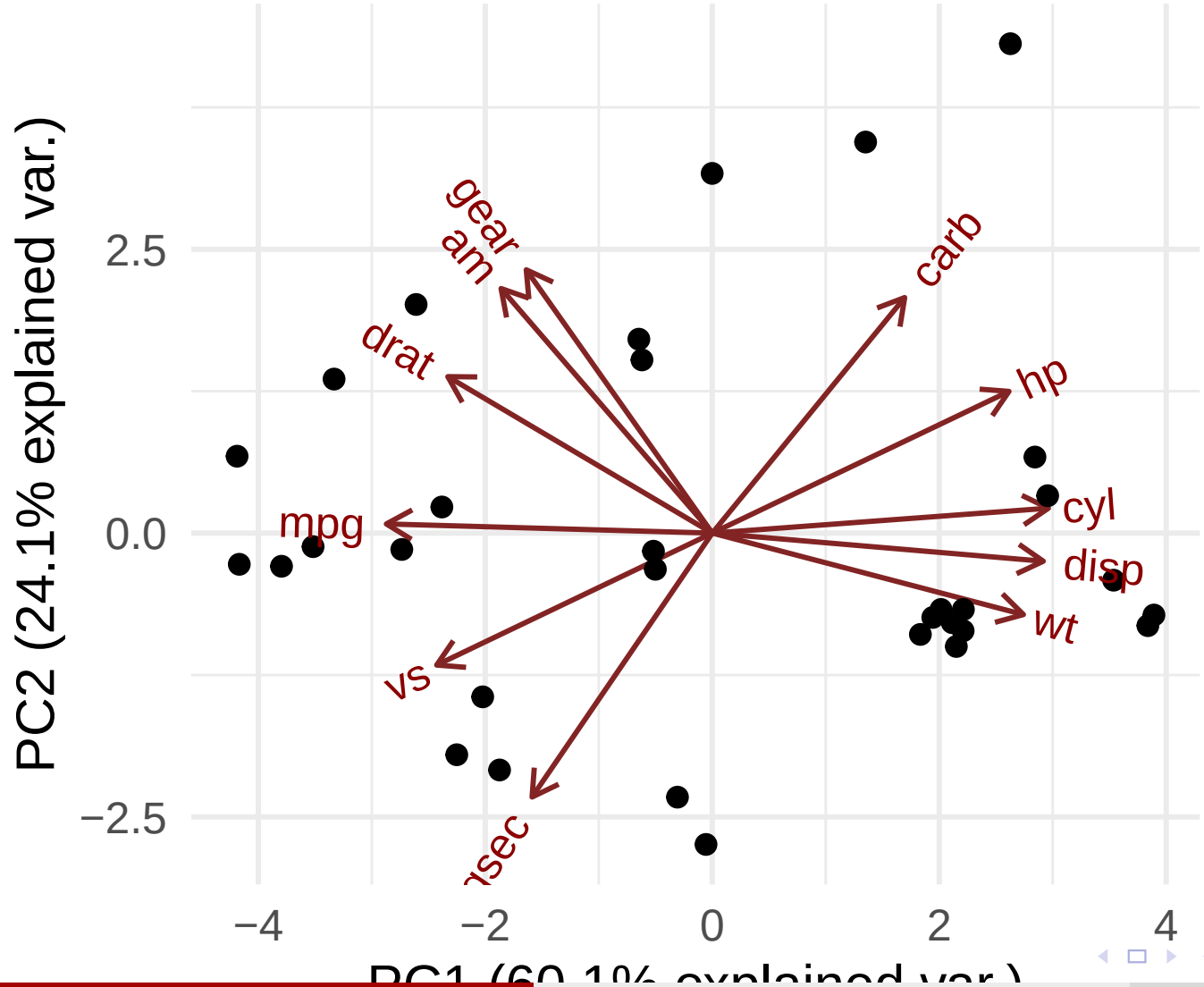


# Plot the first two components

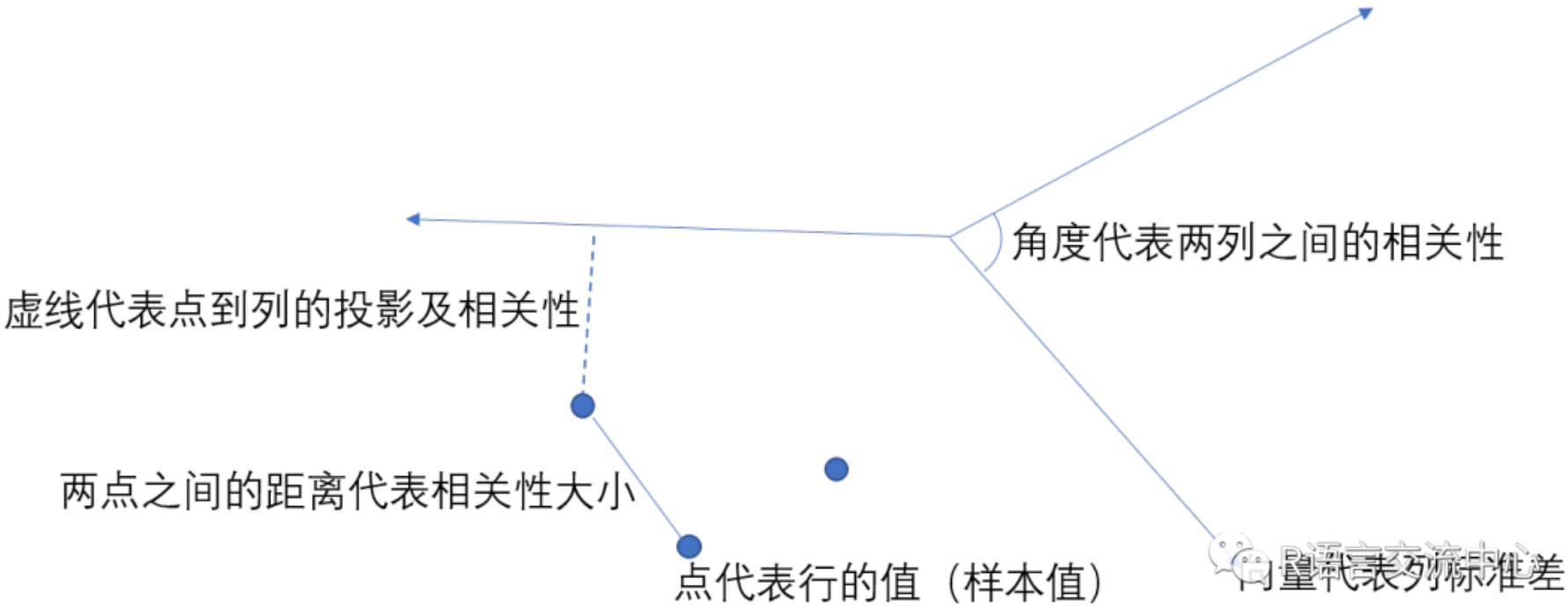
```
require(ggbiplot);
plot1 <-
  ggbiplot(mtcars.pca, obs.scale = 1, var.scale = 1) + theme_minimal();
```

# Plot the first two components, cont.

```
plot1;
```



# 双标图中各个线段、点所代表的意义




注：图片来自腾讯云.

# XY 数据从何而来？

```
knitr::kable(mtcars.pca$x[1:10, 1:4])
```

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| Mazda RX4 | -0.6468627 | 1.7081142 | -0.5917309 | 0.1137022 |
| Mazda RX4 Wag | -0.6194831 | 1.5256219 | -0.3763013 | 0.1991212 |
| Datsun 710 | -2.7356243 | -0.1441501 | -0.2374391 | -0.2452155 |
| Hornet 4 Drive | -0.3068606 | -2.3258038 | -0.1336213 | -0.5038004 |
| Hornet Sportabout | 1.9433927 | -0.7425211 | -1.1165366 | 0.0744620 |
| Valiant | -0.0552534 | -2.7421229 | 0.1612456 | -0.9751674 |
| Duster 360 | 2.9553851 | 0.3296133 | -0.3570461 | -0.0515292 |
| Merc 240D | -2.0229593 | -1.4421056 | 0.9290295 | -0.1421291 |
| Merc 230 | -2.2513840 | -1.9522879 | 1.7689364 | 0.2872110 |
| Merc 280 | -0.5180912 | -0.1594610 | 1.4692603 | 0.0662634 |

# Components and their contributing variables

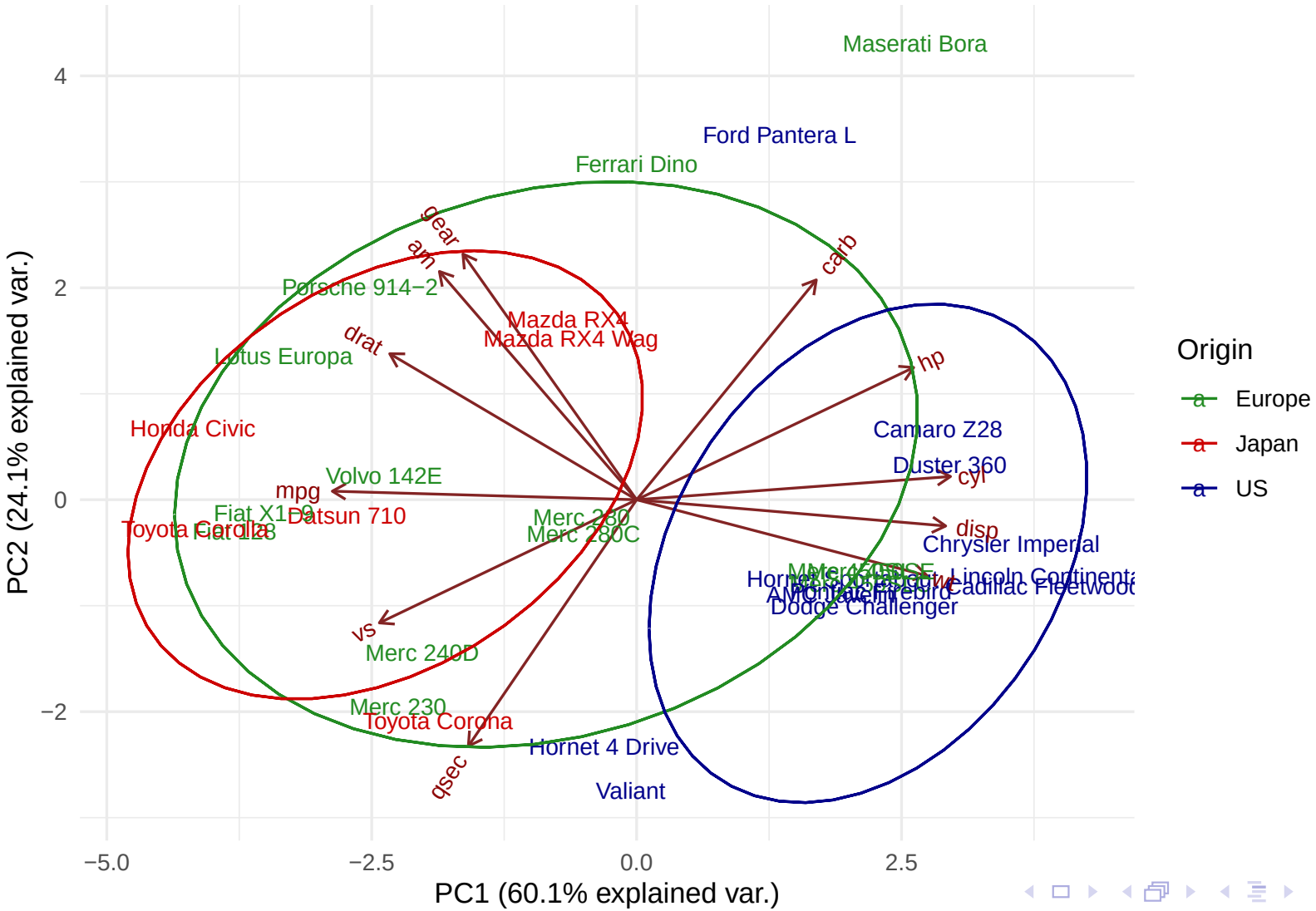
| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| mpg | -0.3625305 | 0.0161244 | -0.2257442 | -0.0225403 | 0.1028447 |
| cyl | 0.3739160 | 0.0437437 | -0.1753112 | -0.0025918 | 0.0584838 |
| disp | 0.3681852 | -0.0493241 | -0.0614841 | 0.2566079 | 0.3939953 |
| hp | 0.3300569 | 0.2487840 | 0.1400148 | -0.0676762 | 0.5400474 |
| drat | -0.2941514 | 0.2746941 | 0.1611888 | 0.8548287 | 0.0773273 |
| wt | 0.3461033 | -0.1430383 | 0.3418185 | 0.2458993 | -0.0750291 |
| qsec | -0.2004563 | -0.4633748 | 0.4031690 | 0.0680765 | -0.1646659 |
| vs | -0.3065113 | -0.2316470 | 0.4288152 | -0.2148486 | 0.5995396 |
| am | -0.2349429 | 0.4294177 | -0.2057666 | -0.0304629 | 0.0897813 |
| gear | -0.2069162 | 0.4623486 | 0.2897799 | -0.2646905 | 0.0483296 |
| carb | 0.2140177 | 0.4135711 | 0.5285446 | -0.1267892 | -0.3613187 |

# PCA 常用分析 1：组间差别

```
mtcars.country <- c(rep("Japan", 3), rep("US",4), rep("Europe", 7),rep("US",3),
                    "Europe", rep("Japan", 3), rep("US",4), rep("Europe", 3),
                    "US", rep("Europe", 3))

## 只做图不显示;
plot2 <-
  ggbiplot(mtcars.pca, ellipse=TRUE, obs.scale = 1, var.scale = 1, labels=rownames(mtcars),
         groups=mtcars.country) +
  scale_colour_manual(name="Origin",
                      values= c("forest green", "red3", "dark blue"))+
  theme_minimal();
```

# PCR 常用分析 1：组间差别，cont.

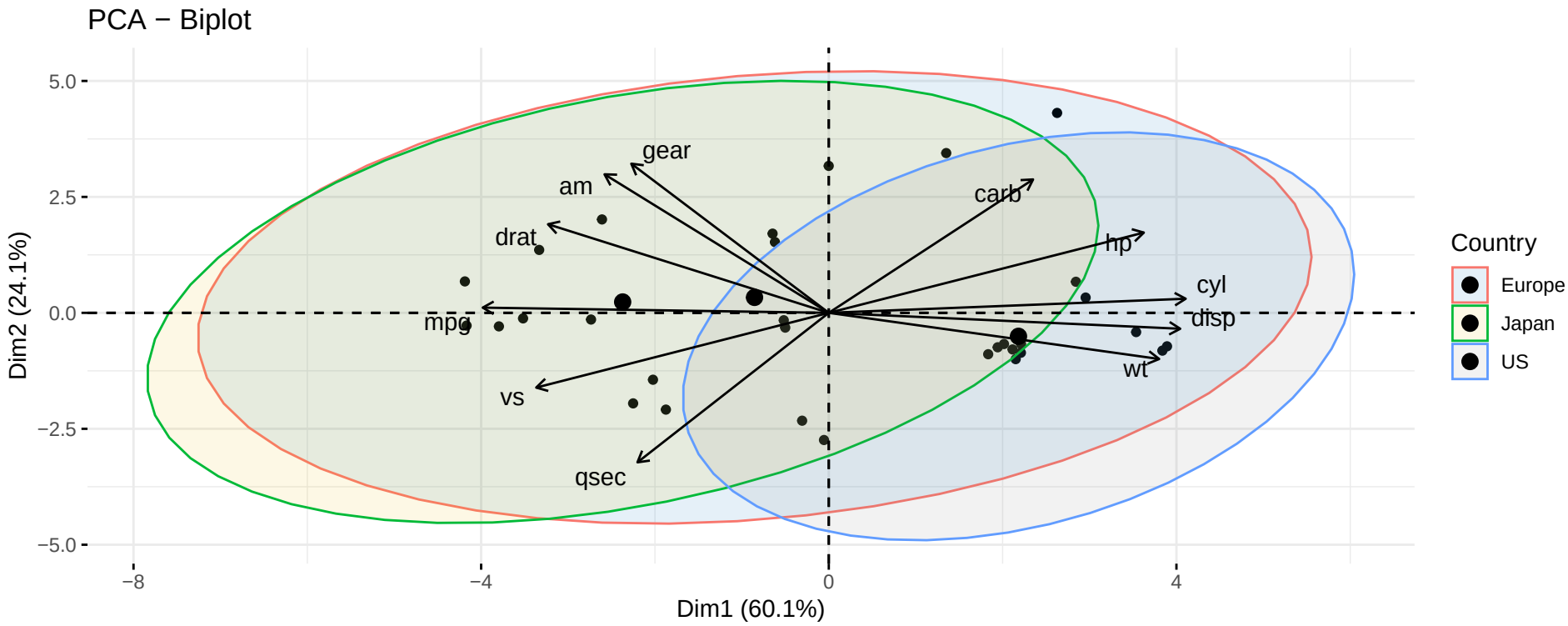
```
plot2
```



# PCA biplot 其它画法

详情见sthda.com.

```
require(factoextra);
fviz_pca_biplot(mtcars.pca, fill.ind = mtcars.country, palette = "jco",
                addEllipses = TRUE, label = "var", col.var = "black",
                repel = TRUE, legend.title = "Country")
```



# PCA 常见分析 2：利用 PCA Scores 进行聚类分析

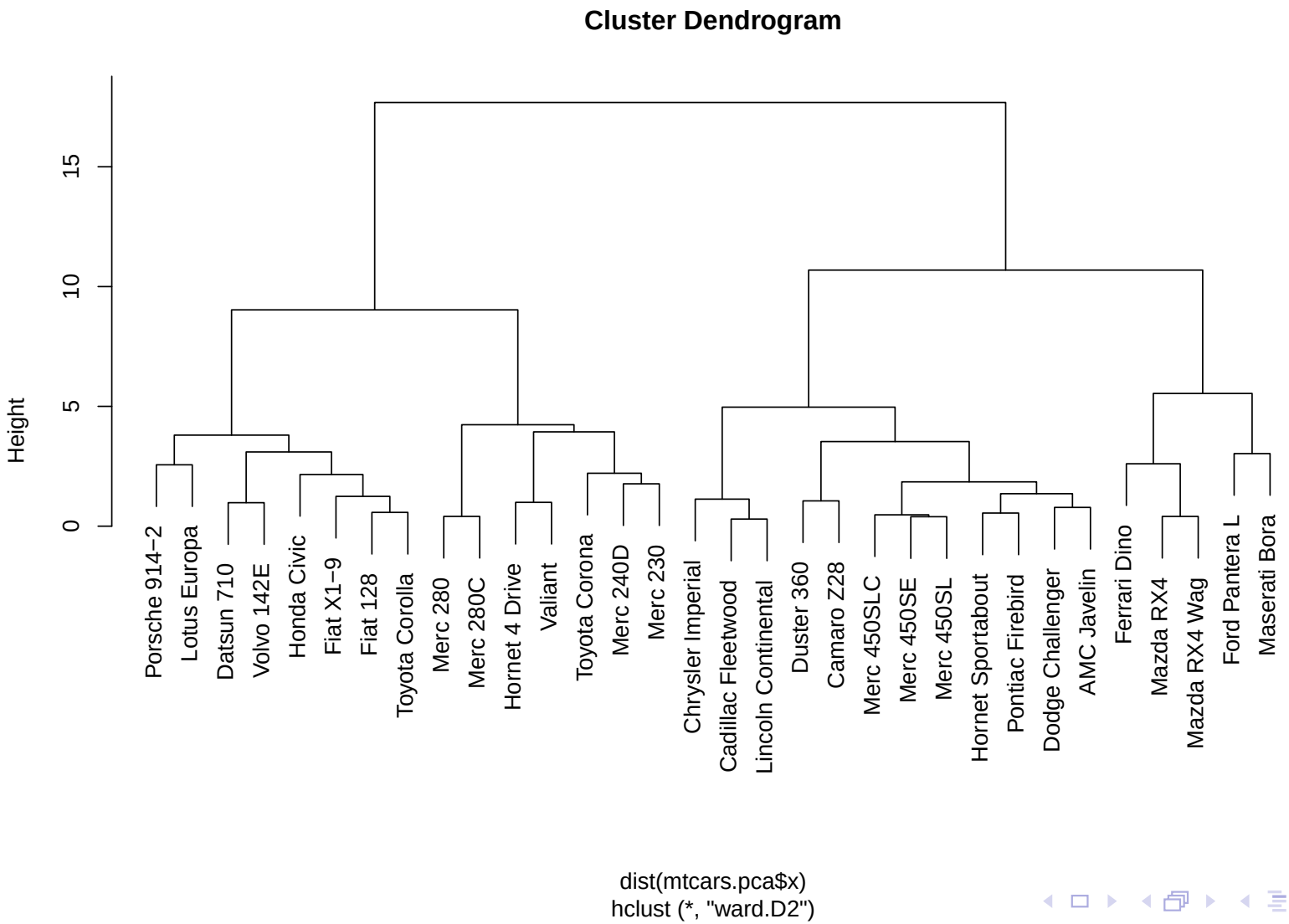
```
knitr::kable( head( round(mtcars.pca$x, digits = 2), n = 3 ) );
```

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 | PC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mazda RX4 | -0.65 | 1.71 | -0.59 | 0.11 | -0.95 | -0.02 | -0.43 | -0.01 | 0.15 | -0.07 | 0. |
| Mazda RX4 Wag | -0.62 | 1.53 | -0.38 | 0.20 | -1.02 | -0.24 | -0.42 | -0.08 | 0.07 | -0.13 | 0. |
| Datsun 710 | -2.74 | -0.14 | -0.24 | -0.25 | 0.40 | -0.35 | -0.61 | 0.59 | -0.13 | 0.05 | -0. |

```
carsHC <- hclust(dist(mtcars.pca$x), method = "ward.D2")
```

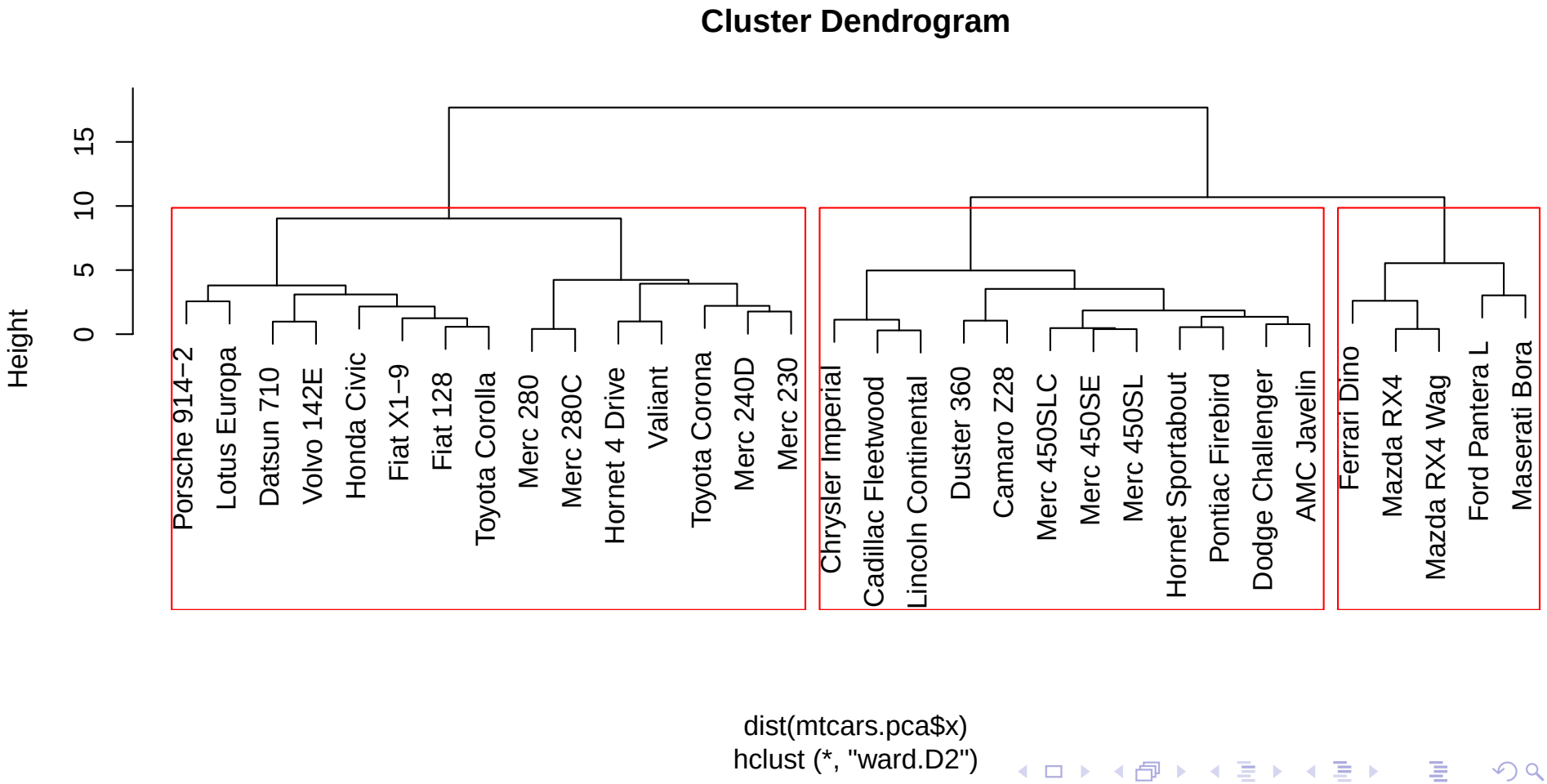
# Cluster Dendrogram

```
plot(carsHC);
```



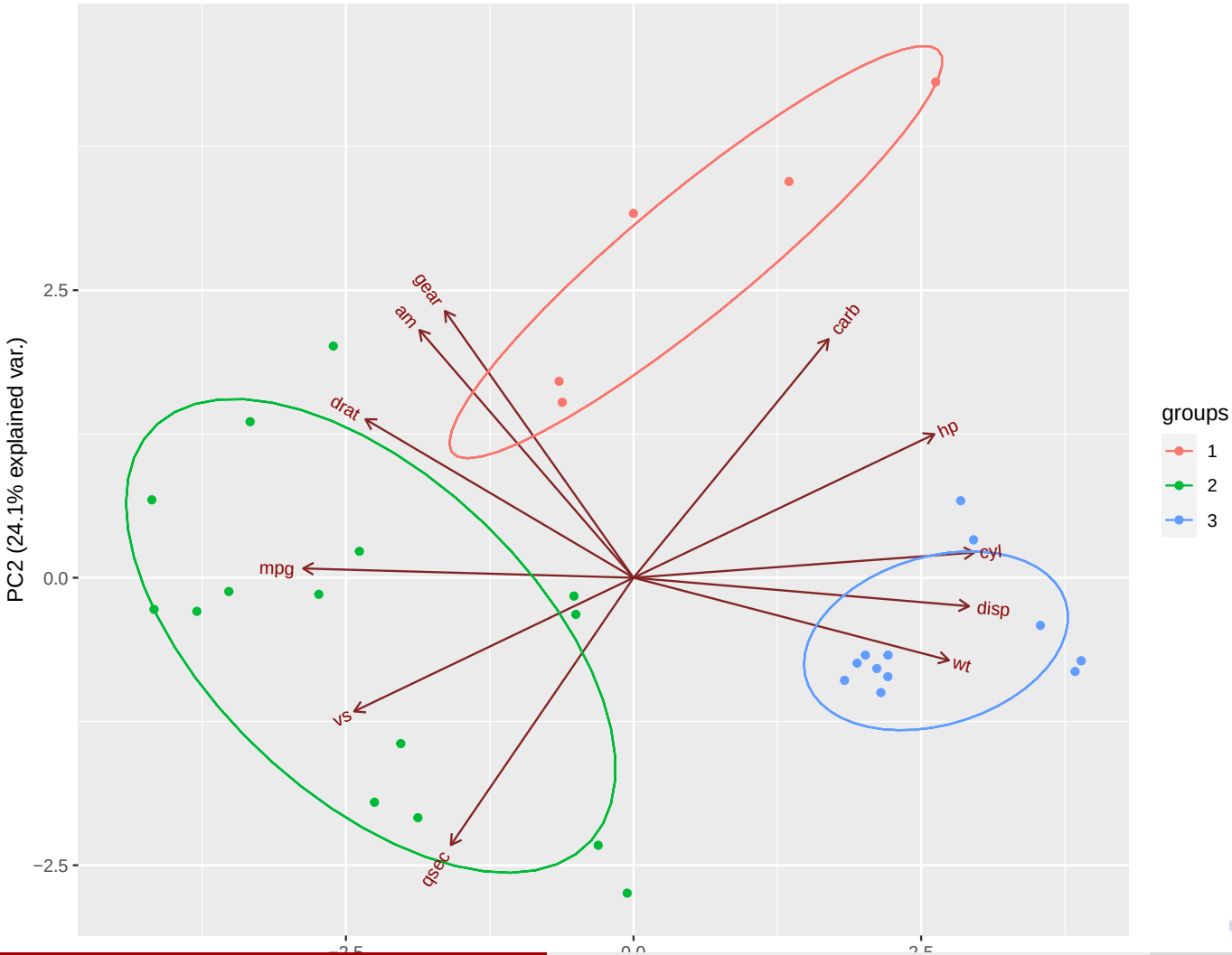
# Cut the Dendrogram

```
carsClusters <- cutree(carsHC, k = 3);
plot(carsHC);
rect.hclust(carsHC, k=3, border="red");
```



# First 2 PCs with Cluster Membership

```
ggbiplot( mtcars.pca, obs.scale = 1, var.scale = 1, groups = factor(carsClusters),
          ellipse = T)
```

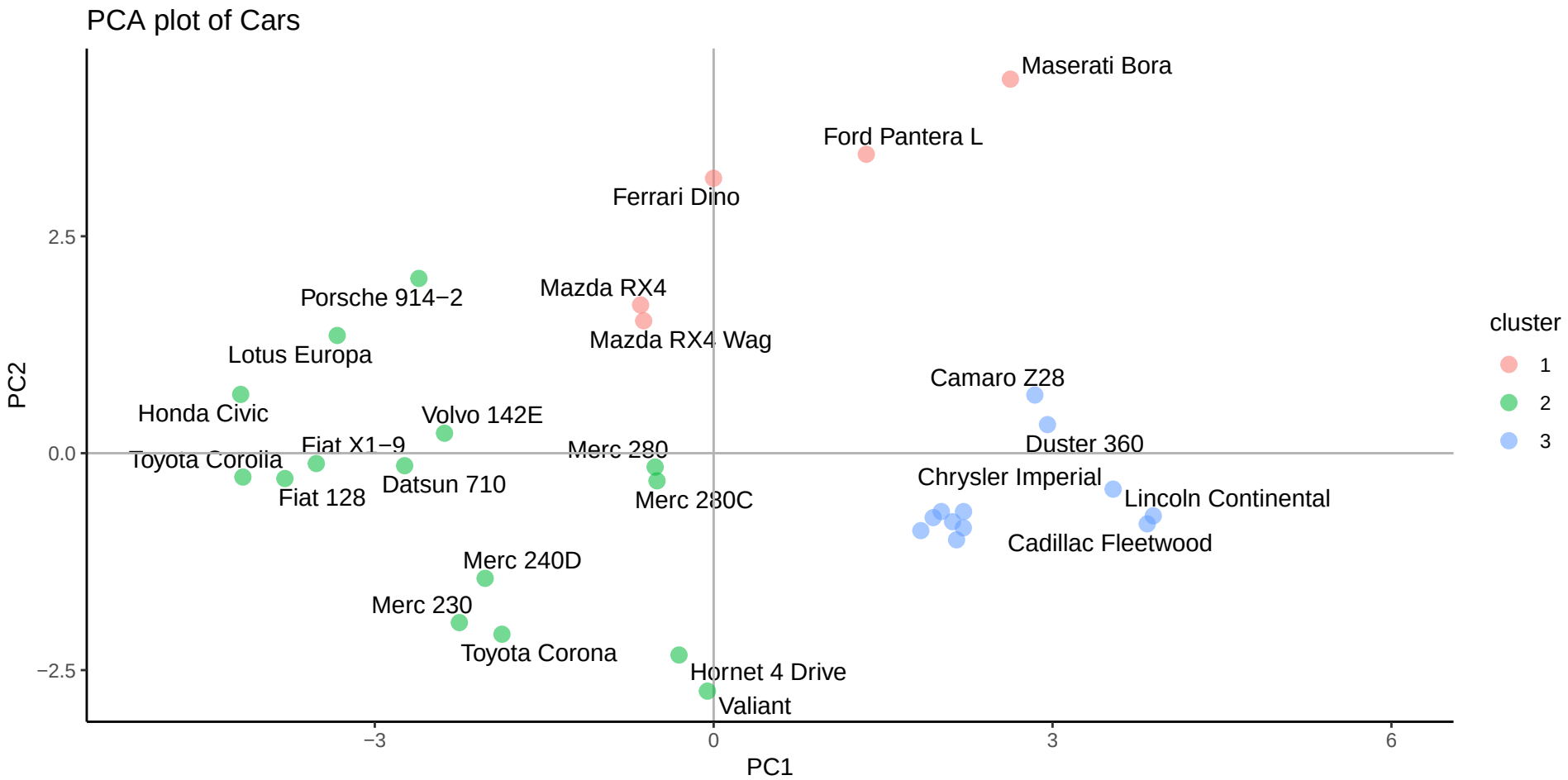


# Plot the Cluster Membership using native ggplot2

```
library(ggplot2)
library(ggrepel)
carsDf <- data.frame( mtcars.pca$x, cluster = factor(carsClusters) );
plot4 <-
  ggplot(carsDf,aes(x=PC1, y=PC2)) +
    geom_text_repel(aes(label = rownames(carsDf))) +
    theme_classic() +
    geom_hline(yintercept = 0, color = "gray70") +
    geom_vline(xintercept = 0, color = "gray70") +
    geom_point(aes(color = cluster), alpha = 0.55, size = 3) +
    xlab("PC1") +
    ylab("PC2") +
    xlim(-5, 6) +
    ggtitle("PCA plot of Cars");
```

# Plot

```
plot4;
```

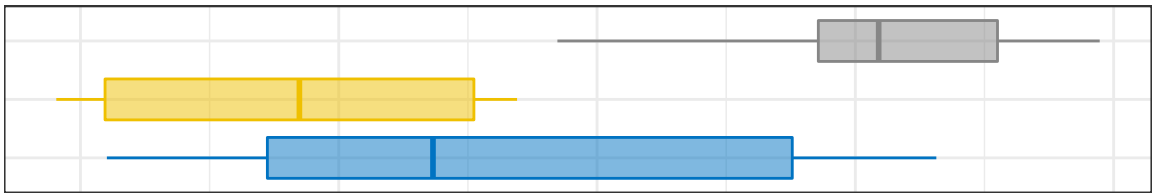

PCA plot of Cars
Maserati Bora
Ford Pantera L
Ferrari Dino
Porsche 914−2
Mazda RX4
Mazda RX4 Wag
Lotus Europa
Camaro Z28
Honda Civic
Volvo 142E
Duster 360
Fiat X1−9
Merc 280
Toyota Corolla
Chrysler Imperial
Lincoln Continental
Datsun 710
Fiat 128
Merc 280C
Cadillac Fleetwood
Merc 240D
Merc 230
Toyota Corona
Hornet 4 Drive
Valiant
2.5
0.0
−2.5
PC2
−3
0
3
6
PC1
cluster
1
2
3

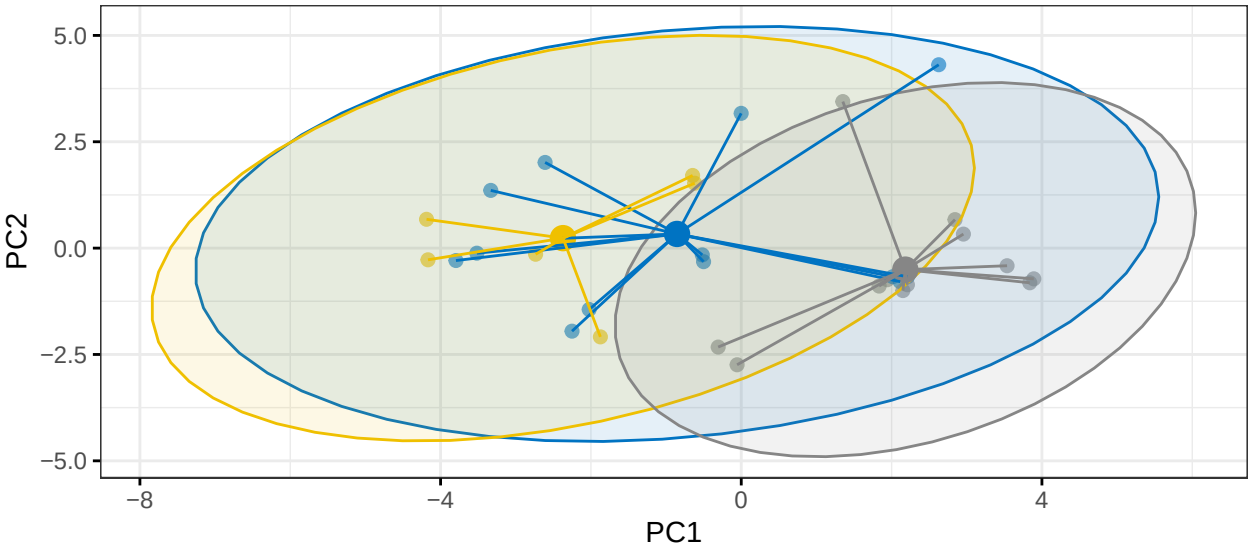
# PCA 进阶分析：添加组间统计分析

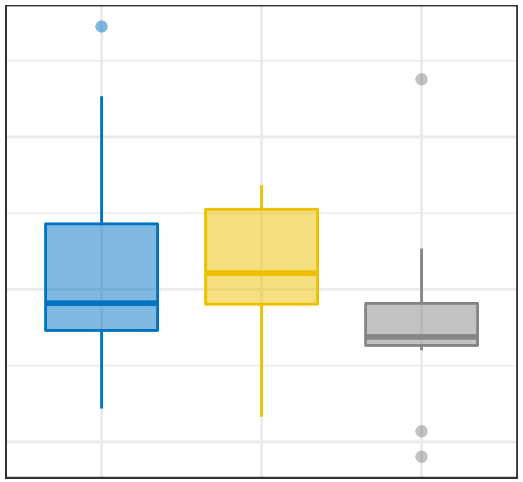
```
library(cowplot);
library("ggExtra");
library(ggpubr);
carsDF2 <- data.frame( mtcars.pca$x, groups = factor(mtcars.country) );

sp <- ggscatter(carsDF2, x = "PC1", y = "PC2",
                color = "groups", palette = "jco",
                ellipse = TRUE, mean.point = TRUE, star.plot = TRUE,
                alpha = 0.6, ggtheme = theme_bw());
# Marginal boxplot of x (top panel) and y (right panel)
xplot <- ggboxplot(carsDF2, x = "groups", y = "PC1",
                   color = "groups", fill = "groups", palette = "jco",
                   alpha = 0.5, ggtheme = theme_bw())+rotate();
yplot <- ggboxplot(carsDF2, x = "groups", y = "PC2",
                   color = "groups", fill = "groups", palette = "jco",
                   alpha = 0.5, ggtheme = theme_bw())
# Cleaning the plots
sp <- sp + rremove("legend");
yplot <- yplot + clean_theme() + rremove("legend")
xplot <- xplot + clean_theme() + rremove("legend")
```

# Plot

```
plot_grid(xplot, NULL, sp, yplot, ncol = 2, align = "hv",
          rel_widths = c(2, 1), rel_heights = c(1, 2))
```





# Section 2: PCoA

# PCA 的局限

PCA 分析存在着自身的局限性，PCA 分析需基于线性模型（linear model）开展，所谓线性模型就是假设物种丰度伴随着环境变量的变化做出线性变化的响应，这种模型使用范围较为有限。

而在实际环境中，**微生物丰度**通常呈现**单峰模型**（unimodal model），该模型假设在一定范围内微生物丰度随环境因素上升而增加，但到达临界值后，若环境因子指标继续增加，微生物丰度则出现下降。如下图所示。

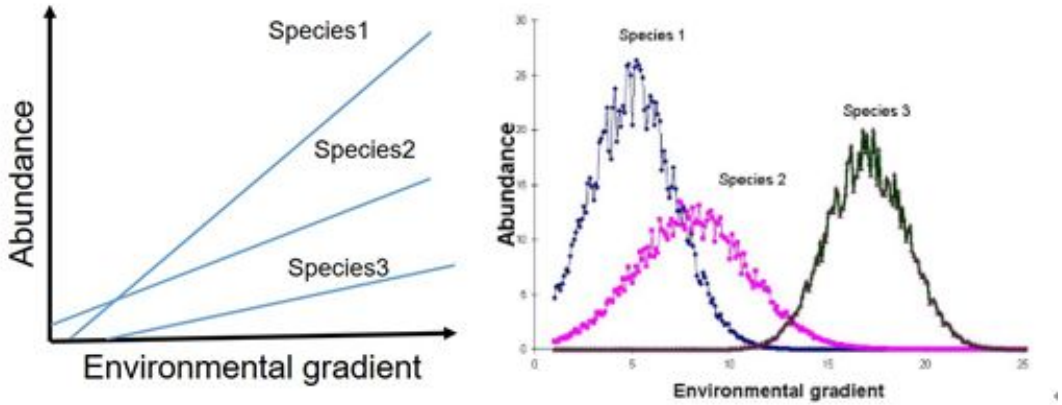



**Figure 1:** 图来源于知乎

# 主坐标分析 | principal co-ordinates analysis，PCoA

PCoA 分析同样采用降维的思想对样本关系进行低维平面的投影。

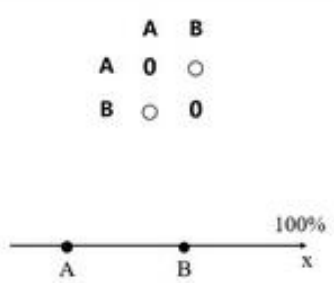
不同的是，PCA 分析是对样本中物种丰度数据的直接投影，而 PCoA 则是将样本数据经过不同距离算法获得样本距离矩阵的投影，在图形中样本点的距离等于距离矩阵中的差异数据距离。

因此，**PCA 是一种同时反映样本与物种信息的分析，而 PCoA 图形则仅对样本距离矩阵进行降维**。

PCoA 常用于微生物　多样性分析中，　多样性的衡量指标是样本相似距离值，相似距离值的算法有很多种，常见的距离类型有：**Jaccard**、**Bray-Curtis**、**Unifrac** 等。

# PCoA 的输入是样本间的距离值

假设有两个样本：

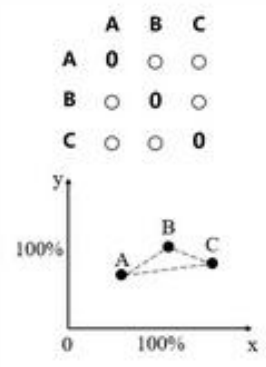



2个样本在一维坐标系中呈现

（上部分为物种组成矩阵，下部分为样本的坐标系呈现）

# 随着样本量增加，维度也增加

三个样本：

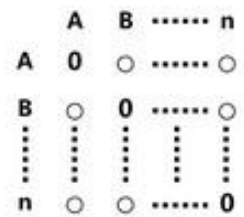



3个样本在二维坐标系中呈现

（上部分为物种组成矩阵，下部分为样本的坐标系呈现）

# N 个样本，N-1 维度

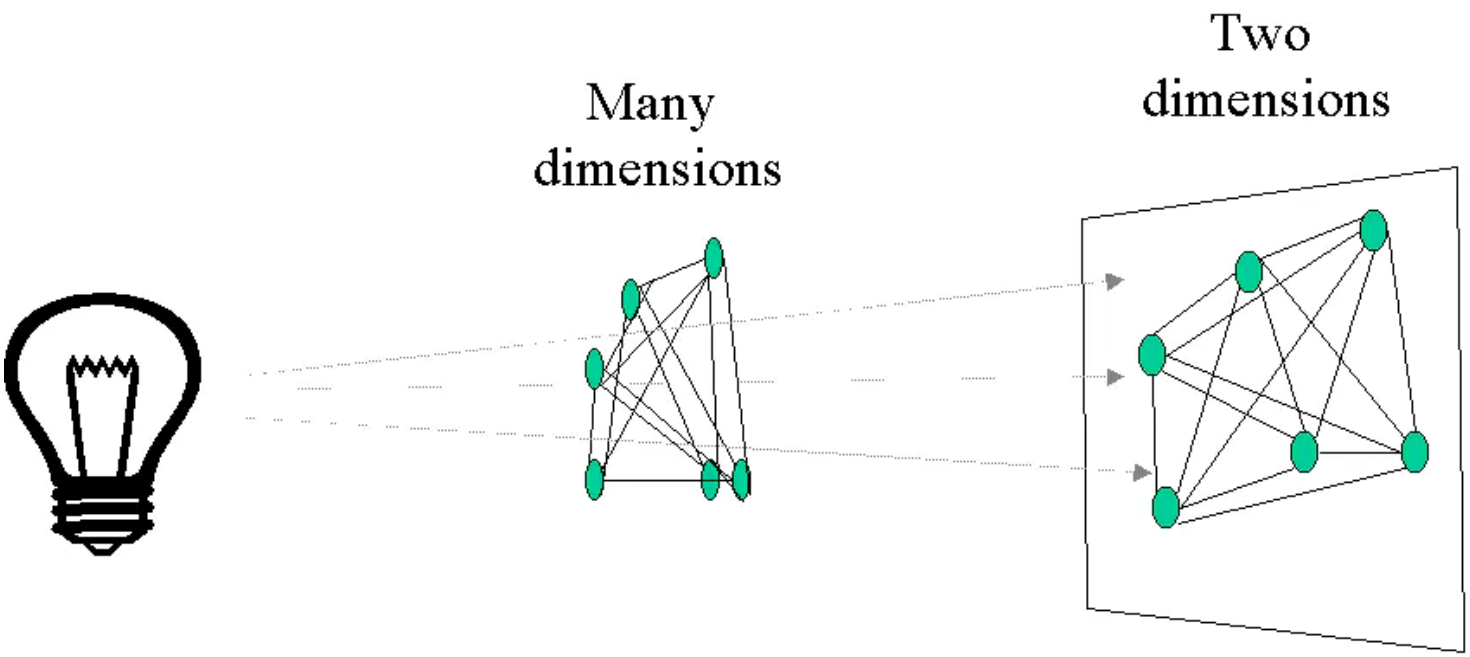
N 个样本：




n个样本的矩阵示意图

以上图来源于 genewiz.com.

# PCoA 的降维处理




因此，和 PCA 一样，PCoA 也有**主成分**，也有信息损失。

# PCoA 分析举例

## 先看一下数据；来源于结直肠癌的宏基因组样本。

```
load("../data/PCA/spe_list_puzi.Rdata"); ## 得到 spe.list 和 crc.meta
names(spe.list);
```

```
## [1] "all.spe.data" "kw.spe.data"  "sig.spe.data"
```

```
knitr::kable( spe.list$all.spe.data[1:3, 1:2] );
```

| | Abiotrophia_defectiva | Acidaminococcus_fermentans |
|---|---|---|
| CCMD10032470ST-11-0 | 0 | 0 |
| CCMD10191450ST-11-0 | 0 | 0 |
| CCMD15562448ST-11-0 | 0 | 0 |

```
knitr::kable( head(crc.meta, n = 3) );
```

| | Sample_ID | Age | Gender | BMI | Country | Project | Group | humanper | |
|---|---|---|---|---|---|---|---|---|---|
| SID31004 | SID31004 | 64 | male | 29.35 | AUT | PRJEB7774 | CRC | 1.18e-05 | 0.001 |
| SID31009 | SID31009 | 68 | male | 32.00 | AUT | PRJEB7774 | CTR | 1.55e-05 | 0.001 |
| SID31021 | SID31021 | 60 | female | 22.10 | AUT | PRJEB7774 | CTR | 2.50e-05 | 0.002 |

# 数据预处理

## 得到项目 PRJEB7774 的菌群丰度和 meta data。

```
## get meta-data for the target project
meta.data <- crc.meta;
meta.data$Group2 <- if_else( meta.data$Group == "CTR", 0, 1 );
meta.data <- meta.data %>% filter( Project == "PRJEB7774" );

## -- get species abundance data
feat.spec.raw <- spe.list$all.spe.data;
feat.spec <- feat.spec.raw[rownames(meta.data), 1:( ncol(feat.spec.raw )-2 )]; ## remove Group

## check if data are valid --
nrow(feat.spec);
```

```
## [1] 107
```

```
table(meta.data$Group);
```

```
##
## CRC CTR
##  46  61
```

# 计算样本间距离

样本间距离由 vegan 包的 vegdist 函数完成；此函数支持多种方法计算距离，请使用 ?vegdist 查看。

```
require(vegan);
dist <- vegdist( feat.spec, method = "bray" );
```

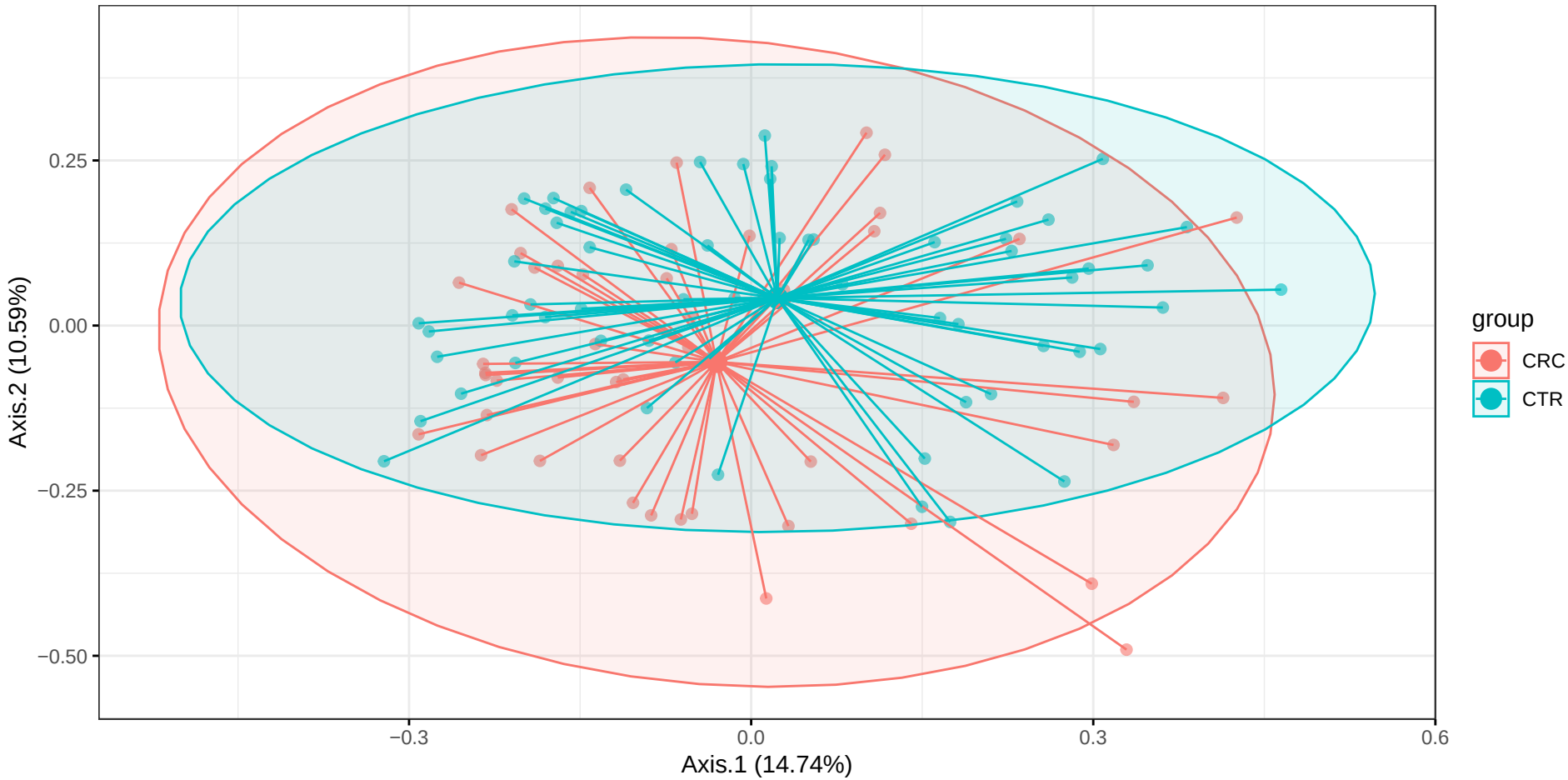
```
> dist
           SID31004  SID31009  SID31021  SID31071  SID31112  SID31129  SID31159
SID31009  0.5952456
SID31021  0.5941267 0.5179527
SID31071  0.4445009 0.5857364 0.5134636
SID31112  0.8715033 0.8359772 0.8213524 0.7386002
SID31129  0.7429998 0.6966337 0.7425318 0.6307419 0.7589381
SID31159  0.8189178 0.7968901 0.7602995 0.6879346 0.4433978 0.7761685
```

# 计算 PCoA

```
require(ape);
PCOA <- pcoa(dist);
rel.eig <- PCOA$values$Relative_eig[1:10];
val <- data.frame( PCOA$vectors[, 1:2], group = factor(meta.data$Group) );

plot_pcoa <-
  ggscatter(val,color = "group", x = "Axis.1", y = "Axis.2",
            ellipse = TRUE, mean.point = TRUE, star.plot = TRUE,
            alpha = 0.6, ggtheme = theme_bw()) +
    xlab( paste0( "Axis.1 (", round( rel.eig[1] * 100, digits = 2), "%)") ) +
    ylab( paste0( "Axis.2 (", round( rel.eig[2] * 100, digits = 2), "%)") );
```

# PCoA plot

```
plot_pcoa;
```



# Are two groups significantly different

```
adonis( dist ~ group, data = val);
```

```
##
## Call:
## adonis(formula = dist ~ group, data = val)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##            Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## group       1    0.5993 0.59927  2.1658 0.02021  0.011 *
## Residuals 105   29.0533 0.27670         0.97979
## Total     106   29.6526                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

adonis 函数可用于分析两组或多组间的差异。更多细节见：
https://www.rpubs.com/roalle/mres_2019

# Section 3: NMDS

# 非度量多维标度分析法 | Non-metric multidimensional scaling，NMDS

NMDS 分析与 PCoA 分析的相同点在于两者都使用样本相似性距离矩阵进行降维排序分析，从而在二维平面上对样本关系做出判断。

不同于 PCoA 分析，NMDS 弱化了对实际距离数值的依赖，更加强调数值间的排名（秩次），例如三个样本的两两相似性距离，（1,2,3）和（10,20,30）在 NMDS 分析上的排序一致，所呈现的效果相同。

详情见知乎页面：**常见分析方法 | PCA、PCoA 和 NMDS 有什么区别？**

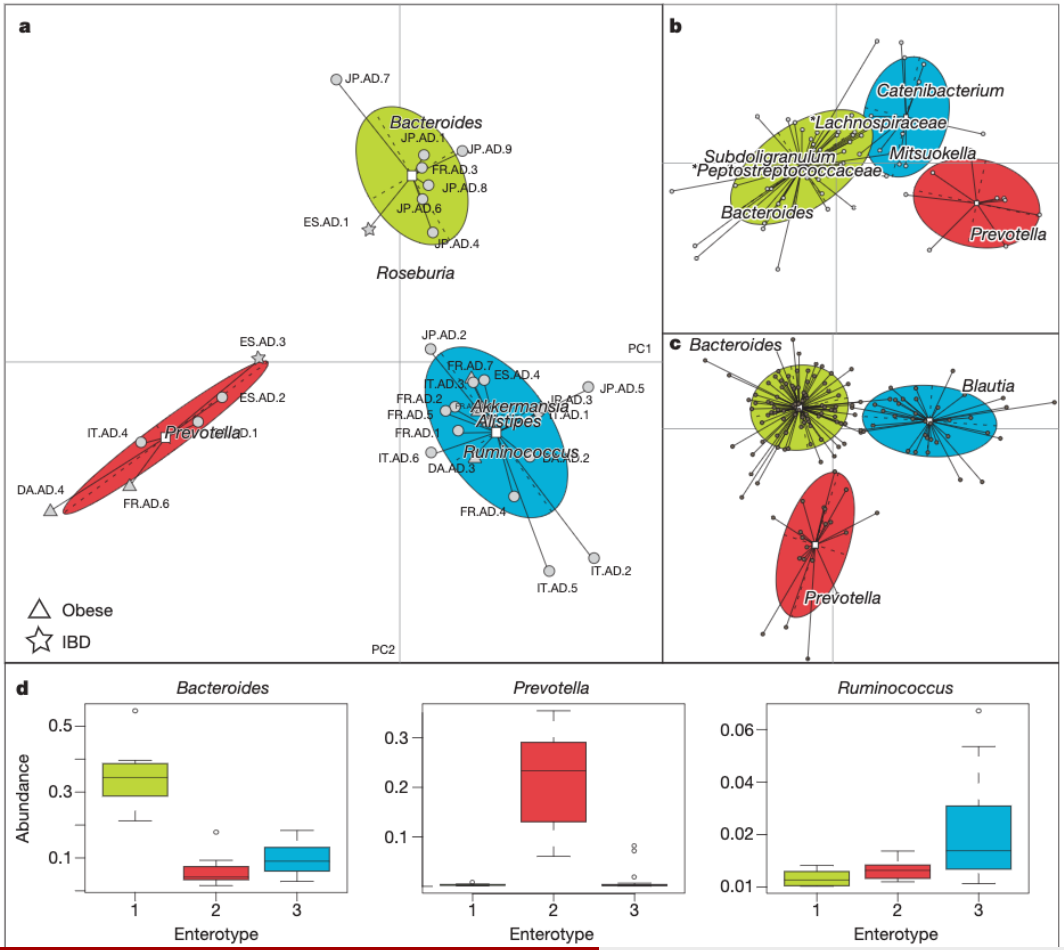
# NMDS 图形

1. 常见分析点：微生物群落研究的　分析。
2. 输入的数据：样本相似性距离表格。
3. 图形类型：散点图。
4. 图形解读：图形中的点代表样本，不同颜色/形状代表样本所属的分组信息。同组样本点距离远近说明了样本的重复性强弱，不同组样本的远近则反应了组间样本距离在秩次（数据排名）上的差异。样本相似性距离计算方式对结果有影响，选择输入不同相似性距离值的矩阵，得到的结果存在着不同程度差异。
5. 横纵坐标轴含义：NMDS 是距离值的秩次（数据排名）信息的评估，图形上样本信息仅反映样本间数据秩次信息的远近，而不反映真实的数值差异，横纵坐标轴并无权重意义，横轴不一定比纵轴更加重要。NMDS 整体降维效果由 Stress 值进行判断。
6. stress 值含义：NMDS 图形通常会给出该模型的 stress 值，用于判断该图形是否能准确反映数据排序的真实分布，stress 值越接近 0 则降维效果越好，一般要求该值 $<0.1$。

# Section 4: 常见生物分析任务

# Enterotype analysis

什么是肠型？In April 2011, the **MetaHIT consortium** published the discovery of enterotypes in the human gut microbiome (**Arumugam, Raes *et al.* 2011**).



# One of the top papers from Peer Bork's group

## Peer Bork

FOLLOWING

EMBL
Verified email at embl.de - Homepage

biology computational biology systems biology bioinformatics evolution

| TITLE | CITED BY | YEAR |
|---|---|---|
| Initial sequencing and analysis of the human genome<br>ES Lander, LM Linton, B Birren, C Nusbaum, MC Zody, J Baldwin, ...<br>Macmillan Publishers Ltd. | 26478 | 2001 |
| A method and server for predicting damaging missense mutations<br>IA Adzhubei, S Schmidt, L Peshkin, VE Ramensky, A Gerasimova, P Bork, ...<br>Nature methods 7 (4), 248-249 | 10597 | 2010 |
| A human gut microbial gene catalogue established by metagenomic sequencing<br>J Qin, R Li, J Raes, M Arumugam, KS Burgdorf, C Manichanh, T Nielsen, ...<br>nature 464 (7285), 59-65 | 8963 | 2010 |
| Initial sequencing and comparative analysis of the mouse genome<br>RH Waterston, L Pachter<br>Nature 420 (6915), 520-562 | 7355 | 2002 |
| STRING v10: protein–protein interaction networks, integrated over the tree of life<br>D Szklarczyk, A Franceschini, S Wyder, K Forslund, D Heller, ...<br>Nucleic acids research 43 (D1), D447-D452 | 6931 | 2015 |
| Functional organization of the yeast proteome by systematic analysis of protein complexes<br>AC Gavin, M Bösche, R Krause, P Grandi, M Marzioch, A Bauer, J Schultz, ...<br>Nature 415 (6868), 141-147 | 5701 | 2002 |
| Enterotypes of the human gut microbiome<br>M Arumugam, J Raes, E Pelletier, D Le Paslier, T Yamada, DR Mende, ...<br>nature 473 (7346), 174-180 | 5536 | 2011 |

Taken on April 13, 2021

# 肠型分析所需数据

下载地址和 tutorial：
https://enterotype.embl.de/enterotypes.html

```
data <- read.table("../data/PCA/MetaHIT_SangerSamples.genus.txt",
                   header=T, row.names=1, dec=".", sep="\t");
data <- data[-1,];
dim(data);
```

```
## [1] 248  33
```

```
knitr::kable( data[1:3, 1:5] ); ## genus level data
```

| | AM.F10.T1 | AM.F10.T2 | DA.AD.1 | DA.AD.2 | DA.AD.3 |
|---|---|---|---|---|---|
| Bacteria | 0 | 0 | 0 | 1.33e-05 | 0 |
| Prosthecochloris | 0 | 0 | 0 | 0.00e+00 | 0 |
| Chloroflexus | 0 | 0 | 0 | 0.00e+00 | 0 |

# Clustering

```
source("../rscripts/dist.JSD.R"); ## load function dist.JSD() ...
data.dist <- dist.JSD(data);
```

We used the Partitioning around medoids (PAM) clustering algorithm to cluster the abundance profiles. PAM derives from the basic k-means algorithm, but has the advantage that it supports any arbitrary distance measure and is more robust than k-means. It is a supervised procedure, where the predetermined number of clusters is given as input to the procedure, which then partitions the data into exactly that many clusters.

# PAM

```
pam.clustering <- function(x,k) { # x is a distance matrix and k the number of clusters
                          require(cluster)
                          cluster = as.vector(pam(as.dist(x), k, diss=TRUE)$clustering)
                          return(cluster)
}

data.cluster <- pam.clustering(data.dist, k=3);
head(data.cluster); ## 显示每个样本的分组情况;
```

```
## [1] 1 1 2 1 1 2
```

```
length(data.cluster); ## 共有 33 个样本
```

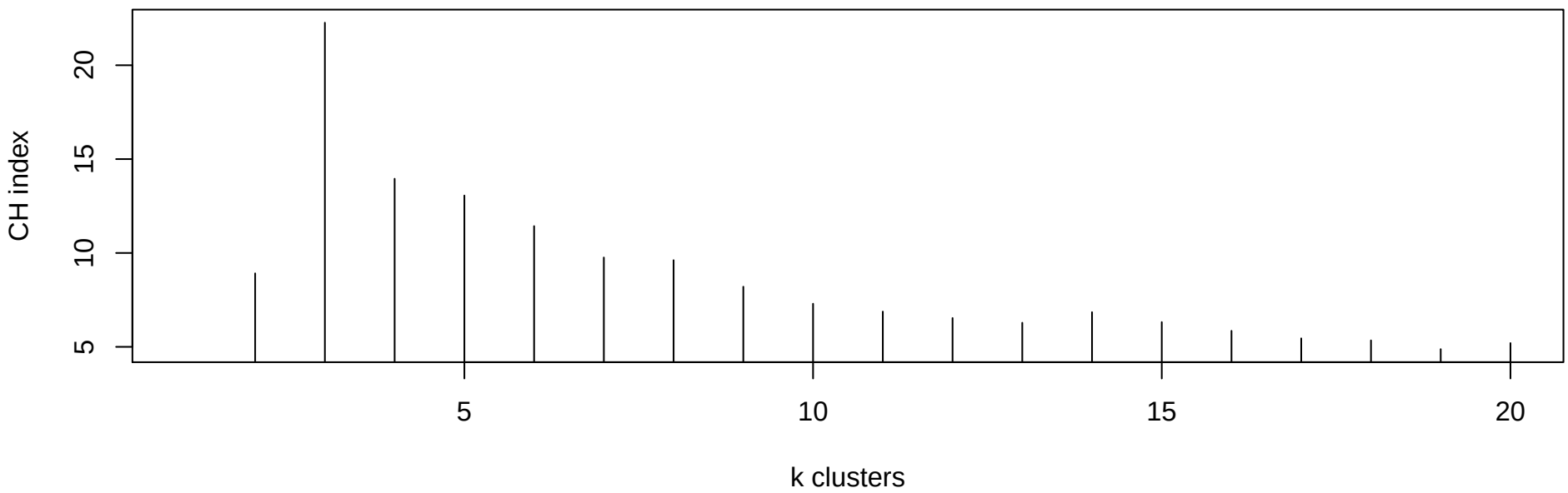
```
## [1] 33
```

# Decide the optimal number of clusters ...

```
require(clusterSim)
nclusters=NULL

  for (k in 1:20) {
      if (k==1) {
          nclusters[k]=NA
      } else {
          data.cluster_temp=pam.clustering(data.dist, k)
          nclusters[k]=index.G1(t(data),data.cluster_temp,  d = data.dist,
          centrotypes = "medoids")
      }
  }
```

# Plot the results

```
plot(nclusters, type="h", xlab="k clusters", ylab="CH index");
```



# Noise removal

```
noise.removal <- function(dataframe, percent=0.01, top=NULL){
    dataframe->Matrix
    bigones <- rowSums(Matrix)*100/(sum(rowSums(Matrix))) > percent
    Matrix_1 <- Matrix[bigones,]
    print(percent)
    return(Matrix_1)
}

data.denoized <- noise.removal(data, percent=0.01);
```

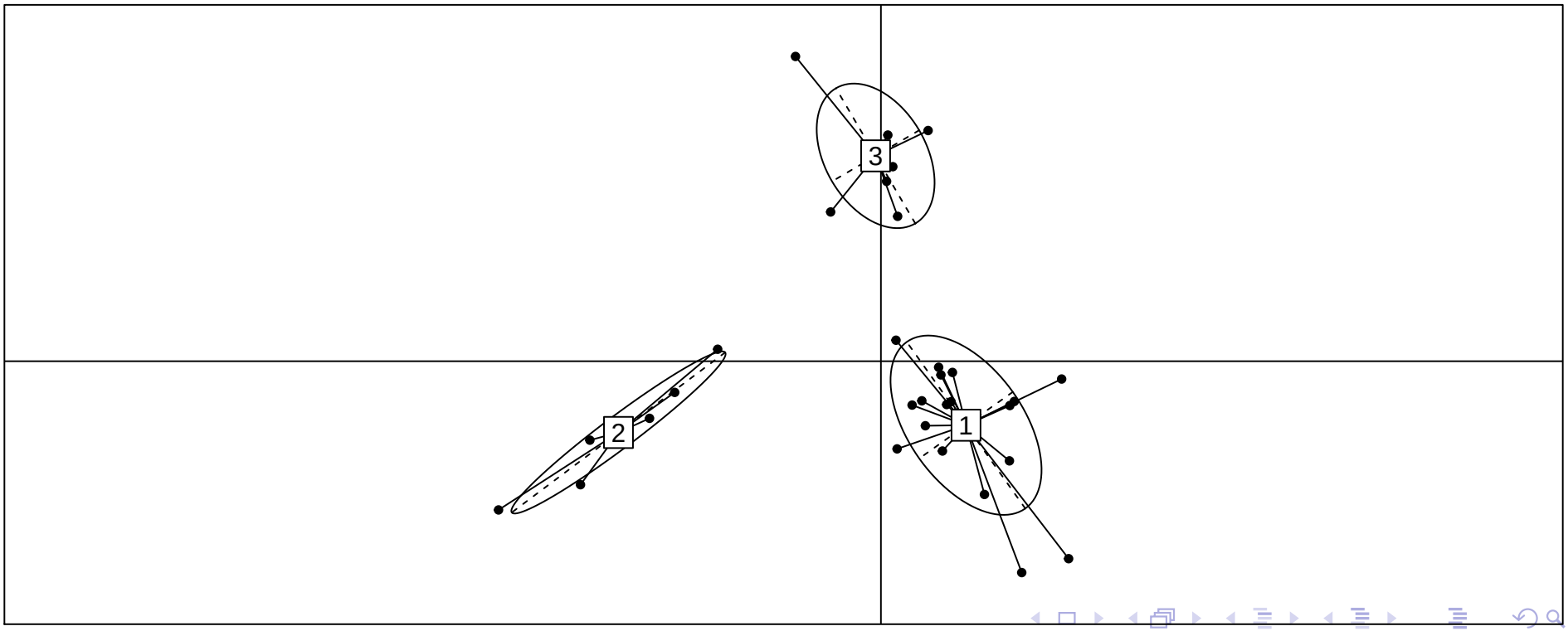
```
## [1] 0.01
```

# Plot enterotype

```
require(ade4);
```

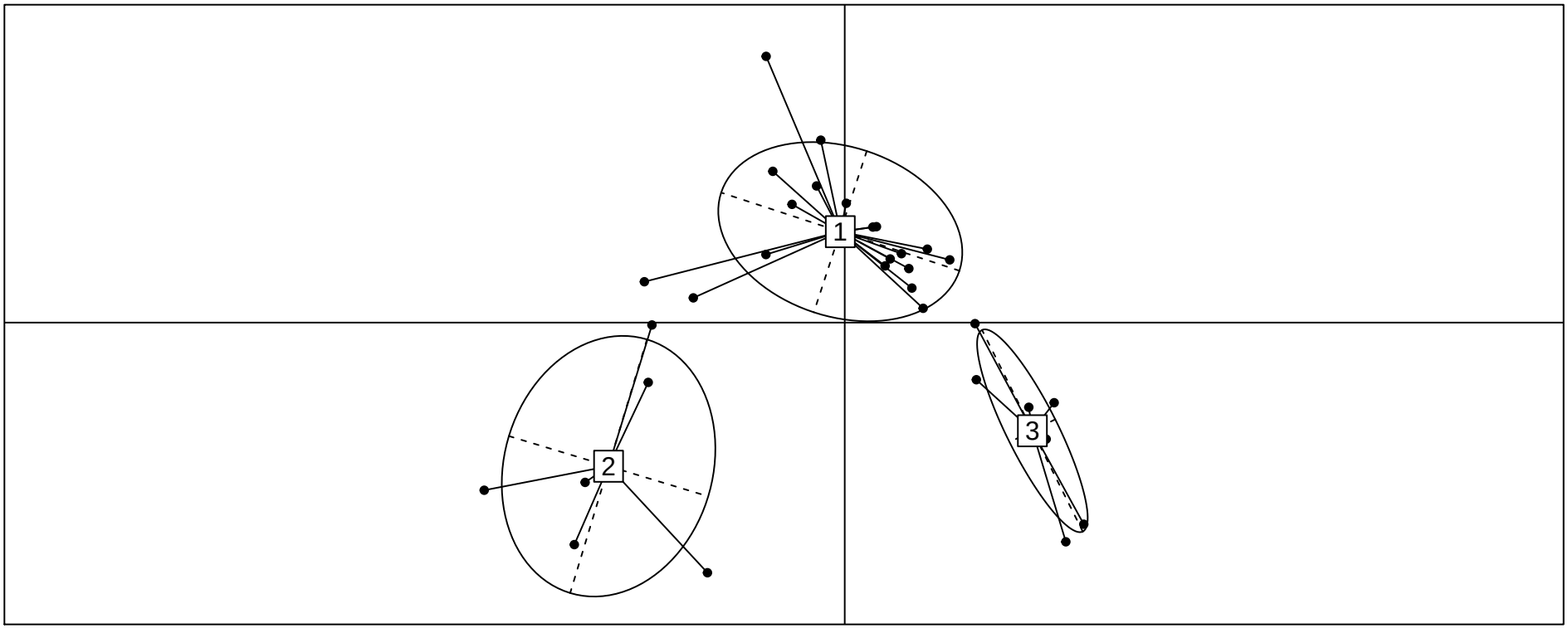
```
## Loading required package: ade4
```

```
obs.pca <- dudi.pca(data.frame(t(data)), scannf=F, nf=10)
obs.bet <- bca(obs.pca, fac=as.factor(data.cluster), scannf=F, nf=k-1)
s.class(obs.bet$ls, fac=as.factor(data.cluster), grid=F)
```



# PCoA instead of PCA

```
obs.pcoa <- dudi.pco(data.dist, scannf=F, nf=3)
s.class(obs.pcoa$li, fac=as.factor(data.cluster), grid=F)
```



# Section 5: Homework

# Homework

- 鉴定每个 enterotype 的富集菌（可以用 LEfSe 分析）
- 显示富集菌在不同 enterotype 样本中的分布情况

# 延伸阅读

- 科学网博客：主成分分析 PCA
- ggpubr: Publication Ready Plots
- PCA - Principal Component Analysis Essentials
- **INTRODUCTION TO ORDINATION**
- Beta-diversity Analysis
- 微生物　多样性常用计算方法比较
- 常见分析方法 | PCA、PCoA 和 NMDS 有什么区别？