Lab1: 多元数据的可视化

1. 内容: 练习多元数据的可视化手段
2. 作业提交: 完成后面的作业, 现场演示给助教并解释结果.

1 三维以下数据的可视化

R包**lattice**提供了一些(如 xyplot(), splom()和cloud()等) 绘制二维散点图, 三维散点图与曲面图, 三维密度图等工具.

xyplot函数提供了灵活的二维散点图: 下面以*iris*数据为例, 展示了三种花的$Petal.Length$和$Peta.Width$变量之间的关系, 并使用非参数拟合方法对各类花两个变量之间的关系进行了拟合.

```
xyplot(Petal.Length ~ Petal.Width, data = iris, #两个变量,一个分类变量
  groups = Species,
  type = c("p", "smooth"), span=.75,
  auto.key =list(title = "Iris Data",  #控制legend
             x = .15, y=.85, corner = c(0,1),
             border = TRUE, lines = TRUE))

#按类别绘制格子图
xyplot(Petal.Length ~ Sepal.Length | Species, iris, type=c("p","smooth"))

#按变量Depth分类绘制格子图
Depth <- equal.count(quakes$depth, number=8, overlap=.1)
xyplot(lat ~ long | Depth, data = quakes)
update(trellis.last.object(), #调整图形
       strip = strip.custom(strip.names = TRUE, strip.levels = TRUE),
       par.strip.text = list(cex = 0.75),
       aspect = "iso")
```

xyplot可以视为是基本绘图函数**plot**的扩展. 阅读xyplot的帮助文档和参考例子, 练习其使用方法, 以及**dotplot, barchart, stripplot,bwplot** 等.

对一个多元数据集, 可以通过研究任意两个变量之间的散点图来展示其特征. **splom**函数是一种增强的绘制散点图阵方法:

```
super.sym <- trellis.par.get("superpose.symbol")
splom(~iris[1:4], groups = Species, data = iris,
          pch =c(8,19,24), col = super.sym$col[1:3],
          panel = panel.superpose,
          key = list(title = "Three Varieties of Iris",    #控制legend
                     columns = 3,
                     points = list(pch =c(8,19,24),
                     col = super.sym$col[1:3]),
                     text = list(c("Setosa", "Versicolor", "Virginica"))))
```

对比基础函数**pairs**.

练习 1. *使用课本表1.6数据, 绘制x_2和x_4的二维散点图, 以及x_3与x_5的散点图, 并使用不同符号和颜色区分两类人群. 对图形进行合适的注解.*

对三维数据来说, 可以直接通过三维散点图或者曲面图来展示数据的特征. **lattice**包里提供的**cloud**函数和**wildframe**提供了三维散点图和曲面图制图功能.

```
super.sym <- trellis.par.get("superpose.symbol")
cloud(Sepal.Length ~ Petal.Length * Petal.Width, data = iris,
         group=Species, pch =c(8,19,24), col = super.sym$col[1:3],
           screen = list(x =-90, y = -20), distance = 0.4, zoom = .7)
trellis.focus("panel",1,1)
panel.identify.cloud(iris[,3],iris[,4],iris[,1],labels=
            paste(rep(c("Se","Ve","Vi"),each=50),rep(1:50,3),sep=""))



g <- expand.grid(x = 1:10, y = 5:15, gr = 1:2)
g$z <- log((g$x^g$gr + g$y^2) * g$gr)
wireframe(z ~ x * y, data = g, groups = gr,
          scales = list(arrows = FALSE),
          drape = TRUE, colorkey = TRUE,
          screen = list(z = 30, x = -60))
```

绘制三维图像的时候, 常常需要选择合适的观察角度. **TeachingDemos**包提供了对**cloud**, **wireframe**等的旋转展示功能, 参考**rotate.cloud**, **rotate.wireframe**.

对三维数据进行展示时候, 交互化绘图包**rgl**可以方便地从不同角度展示数据.

```
library(rgl)
open3d()
x <- sort(rnorm(1000))
y <- rnorm(1000)
z <- rnorm(1000) + atan2(x,y)
plot3d(x, y, z, col=rainbow(1000))
identify3d(x,y,z)
#press ESC to quit identifying points
rgl.snapshot("file1")
```

以及三维曲面

```
x <- seq(-10, 10, length= 30)
y <- x
f <- function(x,y) { r <- sqrt(x^2+y^2); 10 * sin(r)/r }
z <- outer(x, y, f)
z[is.na(z)] <- 1
open3d()
bg3d("white")
material3d(col="black")
persp3d(x, y, z, aspect=c(1, 1, 0.5), col = "lightblue",
        xlab = "X", ylab = "Y", zlab = "Sinc( r )")
```

练习 2. *对课本表1.9数据的长跑变量1500米,3000米和马拉松三个变量绘制三维散点图,是否存在异常点? 在图上标出其名称, 使用不同颜色突出显示.*

2 高维数据的可视化

当数据维数大于三维但不太大时候,脸谱图, 星图以及平行坐标图, Andrew曲线等工具可以用来展示高维数据的特征.

```
utilities <- read.table(file = "T12-4.dat")
names(utilities) <- c("Fixed-charge coverage", "Rate of return on capital",
                      "Cost per kW capacity in place", "Annual load factor",
                      "Peak kWh demand growth from 1974",
                      "Sales (kWh use per year)", "Percent nuclear",
```

```
                    "Total fuel cost (cents per kWh)", "Company")

library(TeachingDemos)
faces(utilities[,-9], labels = as.character(utilities[,9]))
faces2(utilities[,-9], labels = utilities[,9]) # another function

stars(utilities[,-9], labels = as.character(utilities[,9]))
```

对每个观测点的脸谱图或星图进行人工视觉检查, 将类似的划分成一类. 借此发现数据内的特征.

练习 3. 对数据集***USairpollution***, 使用脸谱图或者星图对其进行展示, 并指出是否有城市是类似的.

平行坐标图, Andrew曲线和其他投影方法能够有效展示数据的模式. 而且这些方法适用于变量个数中等大小的场合.

```
parallelplot(~iris[1:4], iris, groups = Species,
                horizontal.axis = FALSE, scales = list(x = list(rot = 90)))

library(andrews)
andrews(iris, type = 4, clr = 5, ymax = 2, main = "Type = 4")

source("starcoord.R") #见课件
starcoord(iris,class = T)

source("mmnorm.R")    #见课件
source("circledraw.R")
source("radviz2d.R")

radviz2d(iris)
```

练习 4. 对包***tourr***里的数据集***flea***, 使用*Andrews*曲线, 平行坐标图以及*star coordinate, Radiz*方法展示其特征, 对几种方法的结果进行比较.