《多元统计分析》第二次上机作业
利用R软件实现多元数据进行可视化和参数估计

3180103000 许乐乐

- 实验目的与要求：

  通过本试验，实现下列目标：

  1. 多元数据分析图示，轮廓图、雷达图、调和曲线图和散布图矩阵；
  2. 能够利用R求解多元正态随机向量的均值、协方差、样本相关矩阵等参数的极大似然估计；
  3. 能够对简单时间序列参数估计的效果进行估计。

习题一

一、附表中的数据sample.xls进行分析。记X_1=BMI, X_2=FPG, X_3=SBP, X_4=DBP, X_5=TG, X_6=HDL-C，并构成一个向量，X=(X_1,X_2,X_3,X_4,X_5)。

(1)分析X各变量之间的相关性?

(2)分析患代谢综合症的比例有没有性别差异，与吸烟或喝酒是否有关?

(3)利用多元数据分析图给出20~30年龄段，X各个指标的分布情况。

(4)给出总体X的均值、协方差矩阵和相关矩阵的估计。

注意：由于是原始数据，数据中有可能有缺失，有数据读入的错误等，在分析之前，要进行数据预处理，如查异常值，分析数据分布形态，检查数据缺失情况，并在实验报告中说明你是如何处理缺失数据的。

- 读取数据

  1. 法一：先在Excel软件中把数据表转存为CSV格式， 然后用readr扩展包的`read_csv()`读取。

     注：文本转换为数值col_double，和因子值col_factor。跳过第一行文本。

```
library(readr)
d <- read_csv(
  "hw2_sample.csv", skip = 1, col_types=cols(
    .default = col_double(),
    性别=col_factor(levels=c("男", "女")),
    吸烟=col_factor(levels=c("是", "否")),
    饮酒=col_factor(levels=c("是", "否")) ))
```

	体检编号	性别	年龄	收缩压	舒张压	体重	身高	吸烟	饮酒	葡萄糖（GLU）	甘油三酯（TG）	高密度脂蛋白胆固醇(HDL)
1	201208090010	男	56	95	60	66.3	171.0	是	是	5.01	0.74	1.79
2	201208090011	女	27	112	58	54.5	172.0	否	NA	4.44	0.75	2.17
3	201208090014	男	NA	131	92	74.9	170.5	NA	NA	5.95	0.66	1.65
4	201208090015	女	NA	98	68	67.3	158.0	NA	NA	5.60	0.75	1.72
5	201208090016	男	64	141	81	59.9	159.0	否	是	5.93	1.76	1.91
6	201208090017	男	43	100	69	59.5	167.5	否	NA	5.68	5.03	1.20

  2. 法二：直接用readxl扩展包的`read_excel()`读取。

注：指定数据类型为数值numeric和文本text。跳过第一行文本。

```
library(readxl)
d=read_excel("hw2_sample.xls", sheet = 2, col_names = TRUE, na = "",  skip = 1,
             col_types = c("numeric","text","numeric","numeric","numeric","numeric","numeric","text","text","numeric","numeric","numeric")
             )
```

	体检编号	性别	年龄	收缩压	舒张压	体重	身高	吸烟	饮酒	葡萄糖（GLU）	甘油三酯（TG）	高密度脂蛋白胆固醇(HDL)
1	201208090010	男	56	95	60	66.3	171.0	是	是	5.01	0.74	1.79
2	201208090011	女	27	112	58	54.5	172.0	否	无	4.44	0.75	2.17
3	201208090014	男	NA	131	92	74.9	170.5	NA	NA	5.95	0.66	1.65
4	201208090015	女	NA	98	68	67.3	158.0	NA	NA	5.60	0.75	1.72
5	201208090016	男	64	141	81	59.9	159.0	否	是	5.93	1.76	1.91
6	201208090017	男	43	100	69	59.5	167.5	否	无	5.68	5.03	1.20

- 修改变量名

  用dplyr扩展包的`rename()`修改数据框变量名

```
d<-d%>%
  dplyr::rename(
    葡萄糖="葡萄糖（GLU）",
    甘油三酯="甘油三酯（TG）",
    胆固醇="高密度脂蛋白胆固醇(HDL)"
  )
```

	体检编号	性别	年龄	收缩压	舒张压	体重	身高	吸烟	饮酒	葡萄糖	甘油三酯	胆固醇
1	201208090010	男	56	95	60	66.3	171.0	是	是	5.01	0.74	1.79
2	201208090011	女	27	112	58	54.5	172.0	否	NA	4.44	0.75	2.17
3	201208090014	男	NA	131	92	74.9	170.5	NA	NA	5.95	0.66	1.65
4	201208090015	女	NA	98	68	67.3	158.0	NA	NA	5.60	0.75	1.72
5	201208090016	男	64	141	81	59.9	159.0	否	是	5.93	1.76	1.91
6	201208090017	男	43	100	69	59.5	167.5	否	无	5.68	5.03	1.20

- 自变量之间的相关性分析

  构建随机变量X

```
attach(d)
X1<-(体重/身高)^2*100
X2<-葡萄糖
X3<-收缩压
X4<-舒张压
X5<-甘油三酯
X6<-胆固醇
X = cbind(X1,X2,X3,X4,X5,X6)
X<-na.omit(X)
```

用`cor()`计算各数值型自变量之间的相关系数。

```
cor1<-cor(X)
```

	X1	X2	X3	X4	X5	X6
X1	1.000000000	-0.002733831	-0.01049344	-0.03115495	0.008276911	0.02687150
X2	-0.002733831	1.000000000	0.19675054	0.27125487	0.388356186	-0.20163401
X3	-0.010493437	0.196750539	1.00000000	0.80588140	0.193984431	-0.07084859
X4	-0.031154950	0.271254865	0.80588140	1.00000000	0.313722801	-0.12591467
X5	0.008276911	0.388356186	0.19398443	0.31372280	1.000000000	-0.33593348
X6	0.026871501	-0.201634007	-0.07084859	-0.12591467	-0.335933482	1.00000000

用corrplot扩展包的`corrplot()`来绘制相关系数矩阵热图。

```
library(corrplot)
corrplot(cor1, tl.col ="black", tl.srt = 0.5, order = "hclust")
# 绘制相关系数矩阵图
# hclust表示按照层次聚类顺序
```

由相关系数矩阵热图，很明显的可以看出：

- 收缩压与舒张压有很强的正相关性
- 甘油三酯与葡萄糖有较强的正相关性
- 甘油三酯与胆固醇有较强的负相关性

- 判断体检者是否患病

用dplyr扩展包的`mutate()`计算数据框中新变量，运用magrittr扩展包的管道连接符`%>%`处理数据框

```
library(magrittr)
library(dplyr)
d.y<-d%>%
  mutate(
    BMI1=ifelse((体重/身高)^2*100>=25,1,0),
    FPG1=ifelse(葡萄糖>=6.1,1,0),
    SDBP1=ifelse(收缩压>=140|舒张压>=90,1,0),
    TGHDL1=ifelse((甘油三酯>=1.7)|(胆固醇<0.9&性别=="男")|(胆固醇<1&性别
=="女"),1,0),
    flag=BMI1+FPG1+SDBP1+TGHDL1,
    患病=ifelse(flag>=3,"是","否")
  )
```

	体检编号	性别	年龄	收缩压	舒张压	体重	身高	吸烟	饮酒	葡萄糖	甘油三酯	胆固醇	BMI1	FPG1	SDBP1	TGHDL1	flag	患病
1	201208090010	男	56	95	60	66.3	171.0	是	是	5.01	0.74	1.79	0	0	0	0	0	否
2	201208090011	女	27	112	58	54.5	172.0	否	NA	4.44	0.75	2.17	0	0	0	0	0	否
3	201208090014	男	NA	131	92	74.9	170.5	NA	NA	5.95	0.66	1.65	0	0	1	0	1	否
4	201208090015	女	NA	98	68	67.3	158.0	NA	NA	5.60	0.75	1.72	0	0	0	0	0	否
5	201208090016	男	64	141	81	59.9	159.0	否	是	5.93	1.76	1.91	0	0	1	1	2	否
6	201208090017	男	43	100	69	59.5	167.5	否	NA	5.68	5.03	1.20	0	0	0	1	1	否

- 修改变量类型

```
d.y$患病<-factor(d.y$患病,levels=c("是","否"))
```

- 列联表独立性卡方检验

1. 患病&性别

```
table.sex<-table(d.y$性别,d.y$患病)  # 2*2列联表
table.sex<-addmargins(table.sex) # 添加求和项
table.sex%>%
  knitr::kable() # 制作成表格
chisq.test(table.sex) # 求卡方值
```

	患病	不患病	Sum
男	8	152	160
女	2	90	92
Sum	10	242	252

```
    Pearson's Chi-squared test

data:  table.sex
X-squared = 1.2242, df = 4, p-value = 0.8741
```

由于p值为0.8741>0.05，所以我们接受原假设，认为患代谢综合症的比例没有性别差异。

2. 患病&吸烟

```
table.smoke<-table(d.y$吸烟,d.y$患病)
table.smoke<-addmargins(table.smoke)
table.smoke%>%
  knitr::kable()
chisq.test(table.smoke)
```

	患病	不患病	Sum
吸烟	4	56	60
不吸烟	4	125	129
Sum	8	181	189

```
    Pearson's Chi-squared test

data:  table.smoke
X-squared = 1.2846, df = 4, p-value = 0.864
```

由于p值为0.864>0.05，所以我们接受原假设，认为患代谢综合症与吸烟无关。

3. 患病&喝酒

```
table.drunk<-table(d.y$饮酒,d.y$患病)
table.drunk<-addmargins(table.drunk)
table.drunk%>%
  knitr::kable()
chisq.test(table.drunk)
```

	患病	不患病	Sum
饮酒	4	72	76
不饮酒	0	11	11
Sum	4	83	87

```
    Pearson's Chi-squared test

data:  table.drunk
X-squared = 0.60685, df = 4, p-value = 0.9623
```

由于p值为0.9623>0.05，所以我们接受原假设，认为患代谢综合症与饮酒无关。

- 筛选符合条件的观测值

  用`filter()`选择行子集

```
d.age<-d%>%
  filter(年龄<=30&年龄>=20)
```

体检编号	性别	年龄	收缩压	舒张压	体重	身高	吸烟	饮酒	葡萄糖	甘油三酯	胆固醇
201208090011	女	27	112	58	54.5	172.0	否	NA	4.44	0.75	2.17
201208100114	女	27	93	61	48.2	165.0	NA	NA	4.73	0.53	1.55
201208110104	女	22	NA	NA	NA	NA	NA	NA	5.05	0.93	2.06
201208110111	女	30	97	55	52.2	165.0	否	NA	5.29	0.59	2.47
201208110254	女	21	95	65	51.7	165.5	否	NA	4.68	1.06	1.53
201208140014	男	28	98	58	70.7	168.0	否	NA	5.13	0.66	1.54
201208140026	女	22	96	61	53.1	163.0	否	NA	4.41	0.61	1.88
201208140106	男	27	104	70	68.6	169.5	NA	NA	5.92	0.75	1.54
201208140155	男	22	106	73	75.2	179.5	否	NA	4.62	1.24	1.90
201208140237	男	21	141	85	101.5	178.0	否	是	5.19	3.64	1.39
201208150138	男	29	118	70	97.0	175.5	否	NA	5.43	1.89	1.06
201208150162	女	29	97	64	53.5	160.5	否	NA	4.92	0.73	1.31
201208150345	女	27	121	76	46.0	147.0	否	NA	5.62	1.65	1.30
201208160012	男	29	115	67	53.6	163.0	NA	NA	4.56	1.07	2.07
201208160018	男	25	120	70	90.6	166.0	NA	NA	5.23	2.96	1.52
201208160026	女	29	115	78	59.0	163.5	NA	NA	4.95	1.55	1.99
201208160028	女	26	107	65	52.6	157.0	否	NA	5.13	0.30	1.89
201208160029	女	27	102	60	45.4	162.5	NA	NA	5.45	0.40	1.23

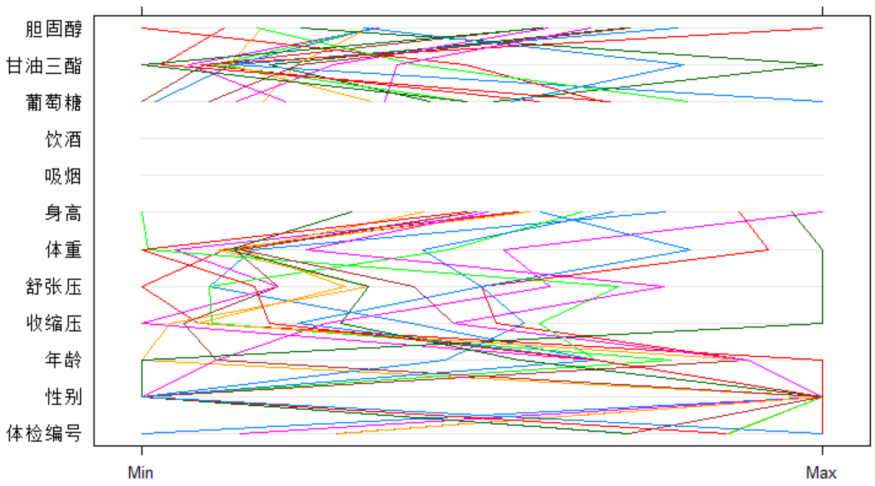
- 多元数据分析图示

  1. 轮廓图

     星相图和脸谱图适合观测值较少的情况，轮廓图更适合观测值较多的情况；星相图和脸谱图没有分类或分类情况未知时适用，轮廓图则在观测值分为几个大类时更适用。

     用`parallelplot()`来绘制轮廓图

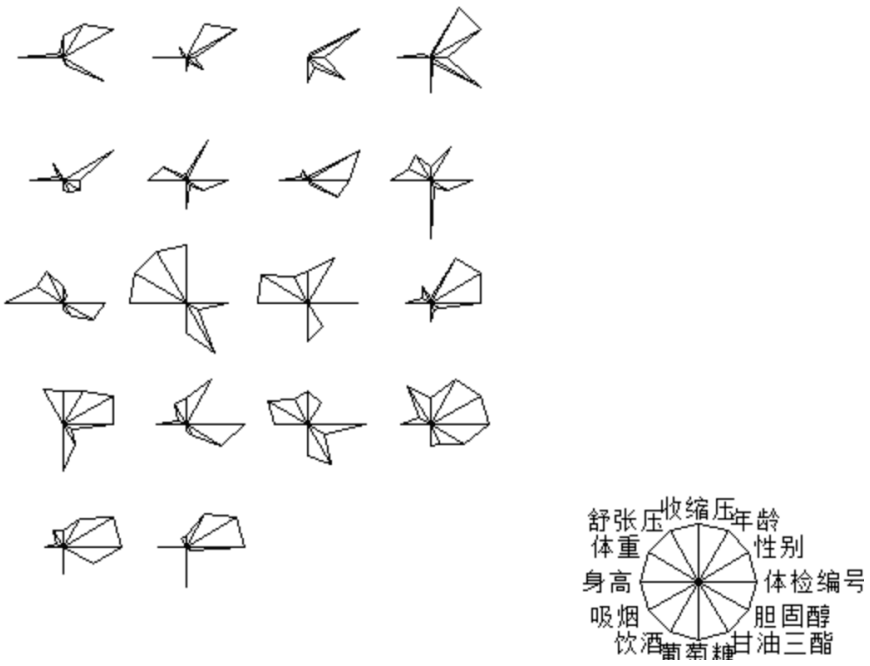
```
library(lattice)
parallelplot(d.age)
```



2. 雷达图

星相图是雷达图的多元表达形式，用`stars()`来绘制星象图

```
stars(d.age,key.loc=c(13,1.5))
# key.loc用来调整右下角图例的位置的函数
```



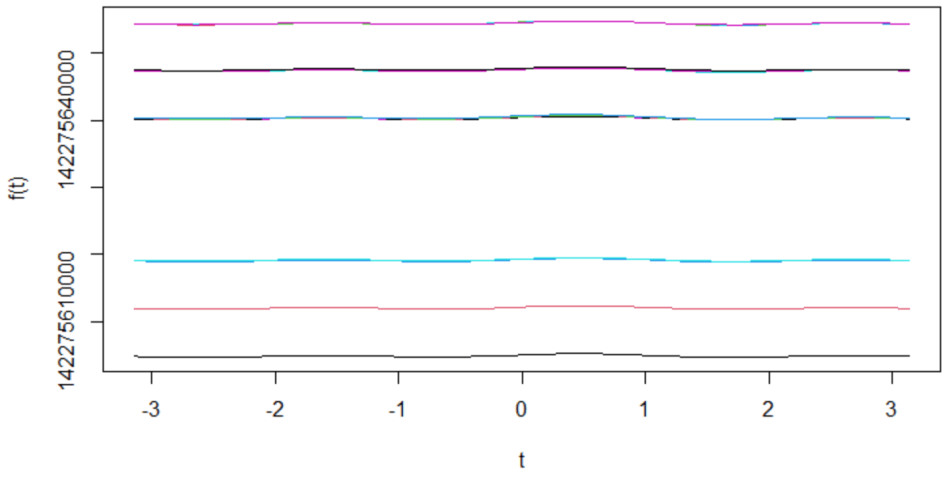
3. 调和曲线图

调和曲线公式：能将多维变量映射到二维平面上

调和曲线图只能对比观测值之间的相似程度，而不能对比变量之间的相似程度，也表现不出哪些变量有助于观测值分类、哪些变量反而混淆了观测值类别。调和曲线图在聚类分析中有重要作用。

用MSG扩展包的`andrews_curve()`来绘制调和曲线图

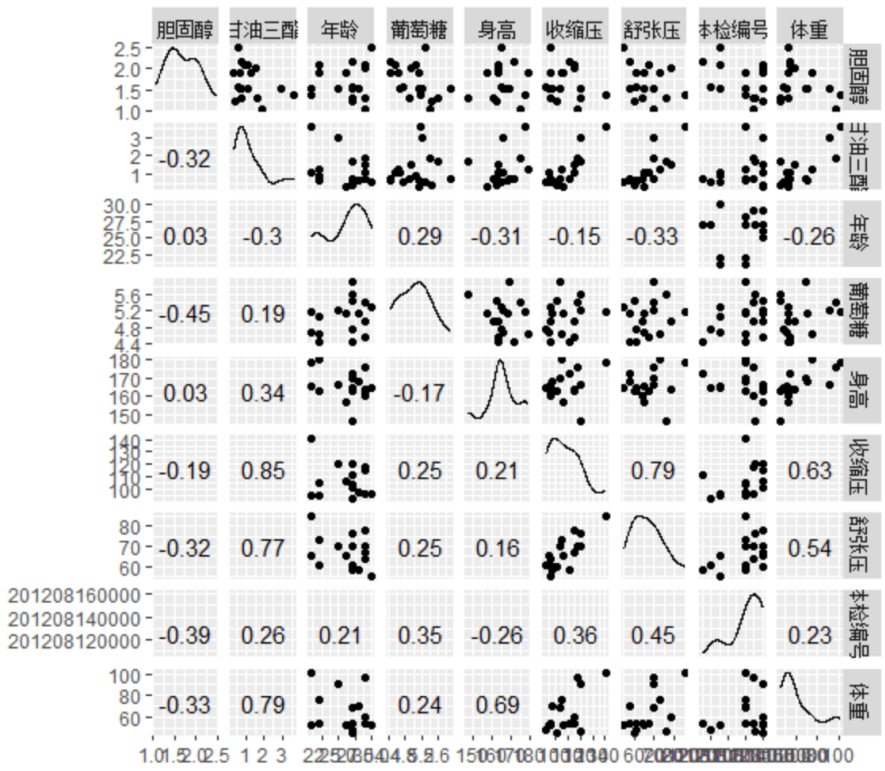
```
library(MSG)
d.age.num=which(sapply(d.age,is.numeric))
andrews_curve(d.age[,d.age.num])
```



4. 散布图矩阵

   多个变量之间的关系经常用散点图矩阵表示。基础R图形中提供了`pairs()`函数作散点图矩阵， GGally扩展包提供了一个`ggscatmat()`函数作散点图矩阵。

```
library(GGally)
ggscatmat(data=d.age,columns=d.age.num)
```



   排列成矩阵的各个小图的下三角位置是两个变量的散点图， 对角线位置是单个变量的核密度估计， 上三角位置是两个变量的相关系数。

- 总体X的均值、协方差矩阵和相关矩阵的估计
  - 删除含有缺失值NA的行
    1. 用`na.omit()`删除数据框中含有缺失值NA的行

```
d.na <- na.omit(d)
```

    2. 用tidyr扩展包的`drop_na()`去除指定的变量有缺失值的行

```
library(tidyr)
d.na<-d%>%
  drop_na()
```

- 选择符合要求的列子集

  用dplyr扩展包的`select()`选择列子集

```
d.na<-d%>%
  select(-性别,-吸烟,-饮酒)%>%  # 扣除性别、吸烟、饮酒这三列（只留下数值型变量）
  drop_na() # 删除有缺失值的行
```

- 用样本均值、样本协方差矩阵、样本相关矩阵来估计

```
options(scipen = 200) # scipen表示在200个数字以内都不使用科学计数法
d.mean<-apply(d.na,2,mean) # 样本均值 2表示列
d.mean<-t(d.mean) # 行列转置
```

	体检编号	年龄	收缩压	舒张压	体重	身高	葡萄糖	甘油三酯	胆固醇
1	201208131535	46.57085	118.6761	76.01619	66.3502	165.5615	5.557611	1.944291	1.608462

```
d.cov<-cov(d.na) # 样本协方差阵
```

	体检编号	年龄	收缩压	舒张压	体重	身高	葡萄糖	甘油三酯	胆固醇
体检编号	423877744.13436	-23237.77770646	36516.6063329	30347.7066752	11592.371181	-17576.338211	-761.5205049	-2866.6044587	-52.11126016
年龄	-23237.77771	121.96955334	57.7791383	31.9907179	14.292770	-6.762914	2.0907594	1.4222963	0.08734522
收缩压	36516.60633	57.77913828	286.0084921	156.3142095	63.031367	19.204159	3.2272720	4.9796883	-0.40257348
舒张压	30347.70668	31.99071788	156.3142095	131.7477042	53.138208	23.889650	3.0067055	5.5406619	-0.49489368
体重	11592.37118	14.29277015	63.0313666	53.1382081	198.714136	17.145638	3.3263643	7.6996699	-1.68950782
身高	-17576.33821	-6.76291432	19.2041588	23.8896498	17.145638	108.465872	1.4732045	3.1635275	-1.06223827
葡萄糖	-761.52050	2.09075936	3.2272720	3.0067055	3.326364	1.473205	0.9488280	0.5710400	-0.06641060
甘油三酯	-2866.60446	1.42229634	4.9796883	5.5406619	7.699670	3.163528	0.5710400	2.2792490	-0.17134622
胆固醇	-52.11126	0.08734522	-0.4025735	-0.4948937	-1.689508	-1.062238	-0.0664106	-0.1713462	0.11572283

```
d.cor<-cor(d.na) # 样本相关阵
```

	体检编号	年龄	收缩压	舒张压	体重	身高	葡萄糖	甘油三酯	胆固醇
体检编号	1.000000000	-0.10219945	0.10487709	0.1284204	0.03994272	-0.08197125	-0.03797236	-0.09222562	-0.007440494
年龄	-0.102199452	1.00000000	0.30935410	0.2523636	0.09180711	-0.05879787	0.19435002	0.08530382	0.023248979
收缩压	0.104877092	0.30935410	1.00000000	0.8052625	0.26439478	0.10903333	0.19590808	0.19503691	-0.069975528
舒张压	0.128420407	0.25236362	0.80526248	1.0000000	0.32841350	0.19984427	0.26892156	0.31973799	-0.126745009
体重	0.039942716	0.09180711	0.26439478	0.3284135	1.00000000	0.11678652	0.24224883	0.36179440	-0.352319197
身高	-0.081971246	-0.05879787	0.10903333	0.1998443	0.11678652	1.00000000	0.14521876	0.20120055	-0.299823640
葡萄糖	-0.037972356	0.19435002	0.19590808	0.2689216	0.24224883	0.14521876	1.00000000	0.38830848	-0.200416808
甘油三酯	-0.092225619	0.08530382	0.19503691	0.3197380	0.36179440	0.20120055	0.38830848	1.00000000	-0.333633089
胆固醇	-0.007440494	0.02324898	-0.06997553	-0.1267450	-0.35231920	-0.29982364	-0.20041681	-0.33363309	1.000000000