1. 主成分分析

R中，与主成分分析有关的函数有：

1. princomp函数

作主成分分析最重要的函数是princomp()函数，其使用格式为：

princomp(formula,data=NULL,subset,na.action,...)

其中formula是没有响应变量的公式，data是数据框。或者

princomp(x,cor=FALSE,scores=TRUE,covmat=NULL,

subset=rep(TRUE,nrow(as.matrix(x)),...)

其中x是用于主成分分析的数据，以数值矩阵或数据框的形式给出。cor是逻辑变量，当cor=TRUE时表示用样本的相关矩阵R作主成分分析，当cor=FALSE（缺省值）表示用样本的协方差阵S作主分分析。covmat是协方差阵，如果数据不由x提供，可由协方差阵提供。

1. summary函数

summary()其目的是提取主成分的信息，其使用格式为：

summary(object, loadings = FALSE, cutff = 0.1, ...)

其中object是由princomp()得到的对象。Loadings是逻辑变量，当loadings = TRUE表示显示loadings的内容（具体含义见下面的loadings ()函数），当loadings = FALSE则不显示。

1. loadings函数

loadings()函数是显示主成分分析或因子分析中loadings（载荷）的内容。在主成分分析中，该内容实际上是主成分对应的各列。loadings()函数的主要使用格式为loadings(x)，其中x是princomp()得到的对象。

1. predict函数

predict()函数是预测主成分的值，其使用格式为：

predict(object, newdata, ...)

其中object是由princomp()得到的对象。Newdata是由预测值构成的数据框，当newdata缺省时，预测已有数据的主成分值。

7.2 样本的主成分

书上**例7.2.1**，R程序如下：

student<-data.frame(

X1=c(148, 139, 160, 149, 159, 142, 153, 150, 151, 139,

140, 161, 158, 140, 137, 152, 149, 145, 160, 156,

151, 147, 157, 147, 157, 151, 144, 141, 139, 148),

X2=c(41, 34, 49, 36, 45, 31, 43, 43, 42, 31,

29, 47, 49, 33, 31, 35, 47, 35, 47, 44,

42, 38, 39, 30, 48, 36, 36, 30, 32, 38),

X3=c(72, 71, 77, 67, 80, 66, 76, 77, 77, 68,

64, 78, 78, 67, 66, 73, 82, 70, 74, 78,

73, 73, 68, 65, 80, 74, 68, 67, 68, 70),

X4=c(78, 76, 86, 79, 86, 76, 83, 79, 80, 74,

74, 84, 83, 77, 73, 79, 79, 77, 87, 85,

82, 78, 80, 75, 88, 80, 76, 76, 73, 78)

)

student.pr<-princomp(student, cor=TRUE)

summary(student.pr,loadings=TRUE)

R程序结果：

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4

Standard deviation 1.8817805 0.55980636 0.28179594 0.25711844

Proportion of Variance 0.8852745 0.07834579 0.01985224 0.01652747

Cumulative Proportion 0.8852745 0.96362029 0.98347253 1.00000000

Loadings:

Comp.1 Comp.2 Comp.3 Comp.4

X1 -0.497 0.543 -0.450 0.506

X2 -0.515 -0.210 -0.462 -0.691

X3 -0.481 -0.725 0.175 0.461

X4 -0.507 0.368 0.744 -0.232

**结果说明：**

1) Standard deviation：表示主成分的标准差，即主成分的方差平方根，即相应特征值的开方；

2) Proportion of Variance：表示方差的贡献率；

3) Cumulative Proportion：表示方差的累计贡献率。

4) 用summmary函数中loadings=TRUE选项列出了主成分对应原始变量的系数, 因此得到前两个主成分是

由于前两个主成分的累计贡献率已经达到96.36%, 所以取前两个主成分来降维。

5) 对于主成分的解释：第1主成分对应的系数符号都相同，其值在0.5左右，它反映学生身材的魁梧程度，身材高大的学生，他的四个尺寸都比较大，则第1主成分的值就较小，绝对值就较大，反之相反，因此我们称第1主成分为大小因子(魁梧因子)；第2主成分是高度与围度的差, 它的值越大表明该学生越“细高”，反之表明该学生越“矮胖”，故称第2主成分为体形因子(或胖瘦因子)。

另外，利用语句predict(student.pr)可以看到各样本的主成分值，并可以在此基础上画出关于第1和第2主成分样本的散点图。

7.3 主成分分析的应用

书上**例7.3.1**，R程序如下：

# 输入数据，按下三角输入，构成向量

x<-c( 1.00,

0.79, 1.00,

0.36, 0.31, 1.00,

0.96, 0.74, 0.38, 1.00,

0.89, 0.58, 0.31, 0.90, 1.00,

0.79, 0.58, 0.30, 0.78, 0.79, 1.00,

0.76, 0.55, 0.35, 0.75, 0.74, 0.73, 1.00,

0.26, 0.19, 0.58, 0.25, 0.25, 0.18, 0.24, 1.00,

0.21, 0.07, 0.28, 0.20, 0.18, 0.18, 0.29,-0.04, 1.00,

0.26, 0.16, 0.33, 0.22, 0.23, 0.23, 0.25, 0.49,-0.34, 1.00,

0.07, 0.21, 0.38, 0.08,-0.02, 0.00, 0.10, 0.44,-0.16, 0.23, 1.00,

0.52, 0.41, 0.35, 0.53, 0.48, 0.38, 0.44, 0.30,-0.05, 0.50, 0.24, 1.00,

0.77, 0.47, 0.41, 0.79, 0.79, 0.69, 0.67, 0.32, 0.23, 0.31, 0.10, 0.62, 1.00,

0.25, 0.17, 0.64, 0.27, 0.27, 0.14, 0.16, 0.51, 0.21, 0.15, 0.31, 0.17, 0.26, 1.00,

0.51, 0.35, 0.58, 0.57, 0.51, 0.26, 0.38, 0.51, 0.15, 0.29, 0.28, 0.41, 0.50, 0.63, 1.00,

0.21, 0.16, 0.51, 0.26, 0.23, 0.00, 0.12, 0.38, 0.18, 0.14, 0.31, 0.18, 0.24, 0.50, 0.65, 1.00)

# 输入变量名称

names<-c("X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8", "X9",

"X10", "X11", "X12", "X13", "X14", "X15", "X16")

# 将向量生成相关矩阵

R<-matrix(0, nrow=16, ncol=16, dimnames=list(names, names))

for (i in 1:16){

for (j in 1:i){

R[i,j]<-x[(i-1)\*i/2+j]; R[j,i]<-R[i,j]

}

}

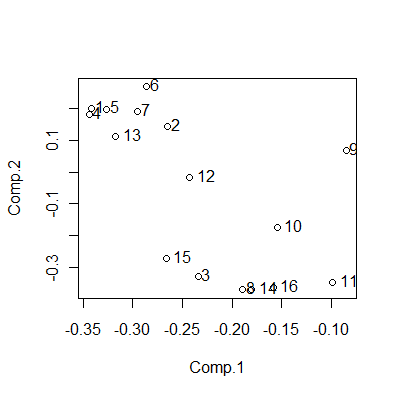
# 作主成分分析

pr<-princomp(covmat=R); load<-loadings(pr)

# 画散点图

plot(load[,1:2]);text(load[,1],load[,2],adj=c(-0.4,0.3))

所得图形：



将上图左上角的点看成一类，它们是“长”类，即身高（）、坐高（）、头高（）、裤长（），下档（）、手长（）、袖长（）。

右下角的点看成一类，它们是“围”类，即胸围（）、领围（）、肩厚（）、肋围（），腰围（）、腿肚（）。

中间的点看成一类，为体形特征，即前胸（）、后背（）、肩宽（）。

R程序如下：

conomy<-data.frame(

x1=c(149.3, 161.2, 171.5, 175.5, 180.8, 190.7, 202.1,

212.4, 226.1, 231.9, 239.0),

x2=c(4.2, 4.1, 3.1, 3.1, 1.1, 2.2, 2.1, 5.6, 5.0, 5.1, 0.7),

x3=c(108.1, 114.8, 123.2, 126.9, 132.1, 137.7,

146.0, 154.1, 162.3, 164.3, 167.6),

y=c(15.9, 16.4, 19.0, 19.1, 18.8, 20.4, 22.7,

26.5, 28.1, 27.6, 26.3)

)

# 作主成分分析

conomy.pr<-princomp(~x1+x2+x3, data=conomy, cor=T)

summary(conomy.pr, loadings=TRUE)

Importance of components:

Comp.1 Comp.2 Comp.3

Standard deviation 1.413915 0.9990767 0.0518737839

Proportion of Variance 0.666385 0.3327181 0.0008969632

Cumulative Proportion 0.666385 0.9991030 1.0000000000

Loadings:

Comp.1 Comp.2 Comp.3

x1 0.706 0.707

x2 -0.999

x3 0.707 -0.707

前两个主成分已达到99%的贡献率。第1主成分是关于国内总产值和总消费，因此称第1主成分为产销因子；第2主成分只与存储量有关，称为存储因子。注意：

= 0.05187378392 = 0.002690889 ≈ 0，变量存在多重共线性，下面做主成分回归。

# 预测样本主成分，并作主成分分析

pre<-predict(conomy.pr)

conomy$z1<-pre[,1]; conomy$z2<-pre[,2]

lm.sol<-lm(y~z1+z2, data=conomy)

summary(lm.sol)

Call:

lm(formula = y ~ z1 + z2, data = conomy)

Residuals:

Min 1Q Median 3Q Max

-0.89838 -0.26050 0.08435 0.35677 0.66863

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 21.8909 0.1658 132.006 1.21e-14 \*\*\*

z1 2.9892 0.1173 25.486 6.02e-09 \*\*\*

z2 -0.8288 0.1660 -4.993 0.00106 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.55 on 8 degrees of freedom

Multiple R-squared: 0.9883, Adjusted R-squared: 0.9853

F-statistic: 337.2 on 2 and 8 DF, p-value: 1.888e-08

回归系数与回归方程均通过检验，而且效果显著。即得到回归方程：

2.9892-0.8288

# 作变换，得到原坐标下的关系表达式

beta<-coef(lm.sol); A<-loadings(conomy.pr)

x.bar<-conomy.pr$center; x.sd<-conomy.pr$scale

coef<-(beta[2]\*A[,1]+ beta[3]\*A[,2])/x.sd

beta0 <- beta[1]- sum(x.bar \* coef)

c(beta0, coef)

(Intercept) x1 x2 x3

-9.13010782 0.07277981 0.60922012 0.10625939

即回归方程为：