统计计算与软件包第一次作业

姓名：王韦力　　学号：2019302010009

2022 年 3 月 9 日

说明： 本文大部分 Rcode 在拷贝时丢失了缩进，但不影响代码的使用.

3.3

分析过程 由题中给出的累积分布函数

$$F(x) = 1 - \left(\frac{b}{x}\right)^a, \quad x \geq b > 0, a > 0,$$

易得逆变换为 $F^{-1}(U) = b(1-U)^{-\frac{1}{a}}, 0 \leq U < 1$.

为用逆变换法模拟一个服从 Pareto(2,2) 分布的样本，令 $a = b = 2$，得累积密度函数与逆变换分别为 $F(x) = 1 - \left(\frac{2}{x}\right)^2, x \geq 2$ 和 $F^{-1}(U) = 2(1-U)^{-\frac{1}{2}}, 0 \leq U < 1$.

为了画出样本的密度直方图并叠加密度函数，对 Pareto(2,2) 的分布函数求导，得密度函数为 $f(x) = 8x^{-3}, x \geq 2$.

参照课本例 3.2 的方法得到服从 Pareto(2,2) 分布的样本，并画出样本的密度直方图和密度函数曲线.

Rcode

```
n <- 1000
u <- runif(n)
x <- 2/sqrt(1-u)
# density histogram of sample
hist(x, prob = TRUE, xlim = c(0,40), xaxt = ``n'', yaxt = ``n'',
breaks = 20, main = bquote(f(x)==1-(2/x)^2))
axis(side = 1, at = c(2,10,20,30,40), labels = c(2,10,20,30,40))
axis(side = 2, at = c(0,0.1,0.2,0.3,0.4), labels = c(0,0.1,0.2,0.3,0.4),
las = 1)
# density curve f(x)
y <- seq(2, 50, 0.05)
lines(y, 8/(y^3))
```

运行结果 R 输出的图形如图 1所示.

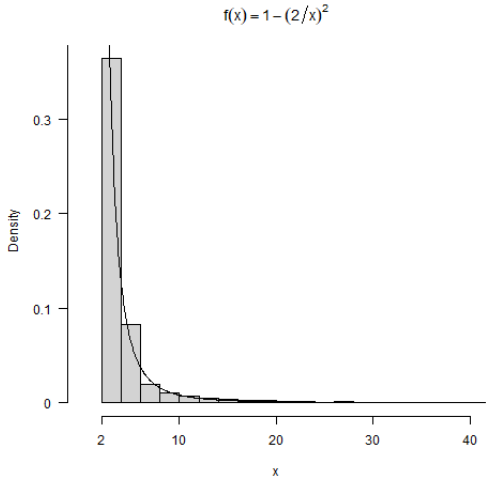

图 1: 练习 3.3 中样本的密度直方图，其中叠加了 Pareto(2,2) 的密度函数曲线

结论 (1) 逆变换为 $F^{-1}(U) = b(1-U)^{-\frac{1}{a}}, 0 \le U < 1$;

(2) 从图 1可以看出，经验分布和理论分布基本吻合.

3.4

分析过程 考虑采用接受拒绝法生成服从 Rayleigh(σ) 分布的随机变量. 对密度函数为

$$f(x) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}, \quad x \ge 0, \sigma > 0$$

的 Rayleigh(σ) 分布，考虑密度函数为

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\sigma)^2}{2\sigma^2}}, \quad \sigma > 0$$

的正态分布 $N(\sigma, \sigma^2)$. 我们证明，存在常数 c 使得 $\frac{f(t)}{g(t)} \le c$ 对所有满足 $f(t) > 0$ 的 t 都成立.

事实上，$f(t) > 0$ 当且仅当 $t > 0$. 此时令 $h(t) = \frac{f(t)}{g(t)}$，则 $h(t) = \frac{\sqrt{2\pi}t}{\sigma} e^{\frac{1}{2}-\frac{t}{\sigma}}$，$h'(t) = \frac{\sqrt{2\pi}(\sigma-t)}{\sigma^2} e^{\frac{1}{2}-\frac{t}{\sigma}}$，于是 $t = \sigma$ 时，$h'(t) = 0$，$h(\sigma) = \frac{\sqrt{2\pi}}{e}$，即对任意 $t > 0$，都有

$$\frac{f(t)}{g(t)} \le \frac{\sqrt{2\pi}}{e}.$$

因此令 $c = \frac{\sqrt{2\pi}}{e}$，对任意 $x > 0$，只要满足

$$\frac{f(x)}{cg(x)} = \frac{ex}{\sigma} e^{\frac{1}{2}-\frac{x}{\sigma}} > u, \quad u \sim U(0,1),$$

一个服从 $g(x)$ 分布的随机数就是被接受的. 根据以上推导过程，生成服从 Rayleigh(σ) 分布的随机变量 Y 的算法如下:

1. 生成一个随机变量 x，使其服从正态分布 $N(\sigma, \sigma^2)$;

2. 生成一个随机变量 u，使其服从 $U(0,1)$;

3. 如果 $u < \frac{ex}{\sigma} e^{\frac{1}{2}-\frac{x}{\sigma}}$，则令 $y = x$，否则拒绝 x，返回步骤 2.

用 R 实现以上算法，分别选取 $\sigma = 0.5, 1$ 和 2，生成 Rayleigh(σ) 样本，并画出频数直方图.

Rcode（以 $\sigma = 0.5$ 为例，其他情况完全类似）

```
sigma <- 0.5
e <- exp(1)
n <- 1000
k <- 0 # counter for accepted
y <- numeric(n)
while (k < n) {
u <- runif(1)
x <- rnorm(1, sigma, sigma^2) # random variate from g
if(e*x*exp(0.5-x/sigma)/sigma > u){ # accept x
k <- k+1
y[k] <- x
}
}
hist(y, xlim = c(0,1.5), xaxt = ``n'', yaxt = ``n'', main = ``y=f(x),
σ=0.5'')
axis(side = 1, at = c(0,0.5,1,1.5), labels = c(0,0.5,1,1.5))
axis(side = 2, at = c(0,50,100,150), labels = c(0,50,100,150),
las = 1)
```

运行结果 $\sigma = 0.5$ 时，结果如图 2所示. 可见样本的众数确实在 $\sigma = 0.5$ 附近. $\sigma = 1$ 和 $\sigma = 2$ 的情形分别如图 3和图 4所示，结论是相似的.

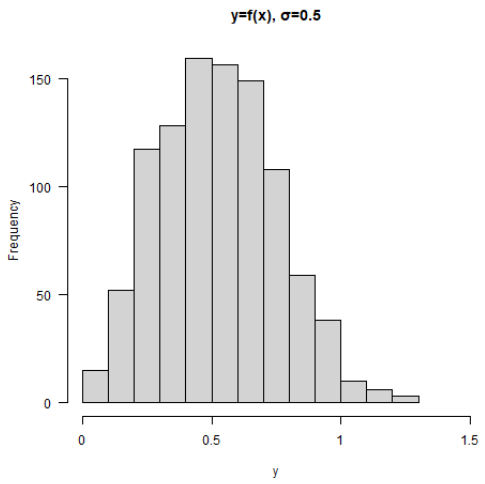

图 2: 练习 3.4 中 Rayleigh(0.5) 样本的频数直方图

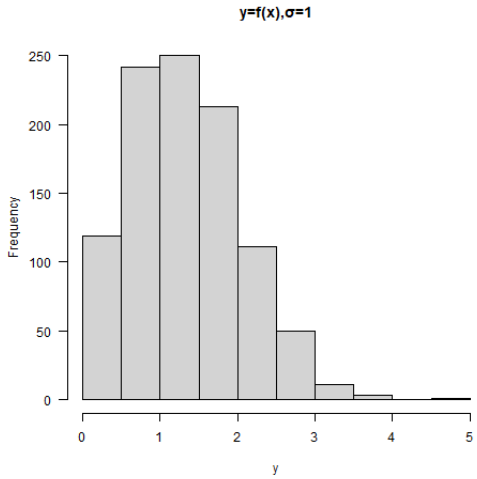

图 3: 练习 3.4 中 Rayleigh(1) 样本的频数直方图

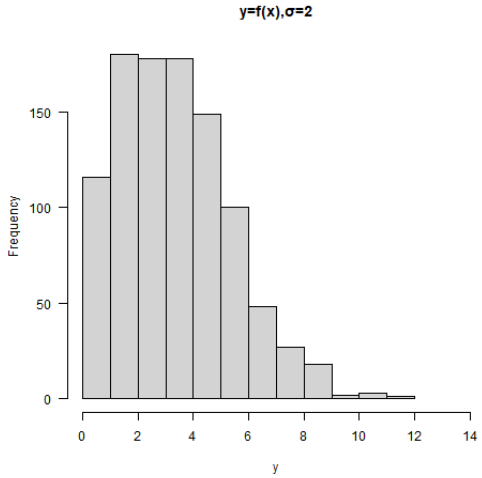

图 4: 练习 3.4 中 Rayleigh(2) 样本的频数直方图

结论 (1) 用接受拒绝法可以生成 Rayleigh(σ) 分布的随机样本，具体算法见“分析过程”部分；

(2) 从图 2，图 3和图 4均可看出，生成样本的众数都接近理论众数 σ.

3.9

分析过程 直接按照题中算法用 R 写出函数并绘制直方图即可.

Rcode

```
fe <- function(n){
U <- numeric(n)
j <- 1 # iterations
while (j <= n) {
u <- runif(3, -1, 1)
if (abs(u[3]) >= abs(u[2]) && abs(u[3]) >= abs(u[1])) U[j] = u[2]
```

```
else U[j] = u[3]
j <- j + 1
}
return(U)
}
n <- 1000
U <- fe(n)
hist(U, prob = TRUE)
```

运行结果 密度直方图如图 5所示.

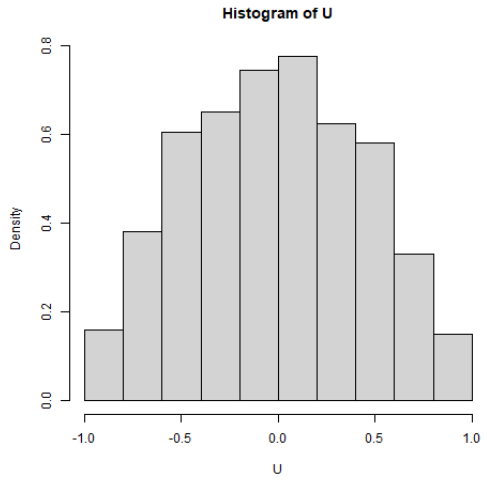

图 5: 练习 3.9 中样本的密度直方图

结论 见“运行结果”部分.

3.11

分析过程 模拟混合变量的步骤如下:

1. 生成一个随机数 $u \sim U(0,1)$;

2. 如果 $u < p_1$, 生成一个来自 $N(0,1)$ 分布的 x; 否则生成一个来自 $N(3,1)$ 分布的 x.

除 $p_1 = 0.5$ 的情形外, 我们还分别令 $p1 = 0.4, 0.6, 0.25$ 和 0.75, 观察混合变量的经验分布是否是双峰分布.

Rcode

```
f <- function(n, mu1, mu2, sigma1, sigma2, p1){
x1 <- rnorm(n, mu1, sigma1)
x2 <- rnorm(n, mu2, sigma2)
u <- runif(n)
k <- as.integer(u < p1)
x <- k*x1 + (1 - k)*x2
```

```
return(x)
}
n <- 1000
mu1 <- 0
mu2 <- 3
sigma1 <- 1
sigma2 <- 1
x0.5=f(n, mu1, mu2, sigma1, sigma2, 0.5)
hist(x0.5, prob = TRUE, xlim = c(-4, 6), ylim = c(0, 0.23),
main = "Histogram of x p1=0.5")
lines(density(x))
par(mfrow = c(2,2))
x0.4=f(n, mu1, mu2, sigma1, sigma2, 0.4)
hist(x0.4, prob = TRUE, xlim = c(-4, 6), ylim = c(0, 0.28),
main = "Histogram of x p1=0.4")
lines(density(x0.4))
x0.6=f(n, mu1, mu2, sigma1, sigma2, 0.6)
hist(x0.6, prob = TRUE, xlim = c(-4, 6), ylim = c(0, 0.28),
main = "Histogram of x p1=0.6")
lines(density(x0.6))
x0.25=f(n, mu1, mu2, sigma1, sigma2, 0.25)
hist(x0.25, prob = TRUE, xlim = c(-4, 6), ylim = c(0, 0.28),
main = "Histogram of x p1=0.25")
lines(density(x0.25))
x0.75=f(n, mu1, mu2, sigma1, sigma2, 0.75)
hist(x0.75, prob = TRUE, xlim = c(-4, 6), ylim = c(0, 0.28),
main = "Histogram of x p1=0.75")
lines(density(x0.75))
par(mfrow = c(1,1))
```

运行结果 R 输出的图形如图 6和图 7所示.

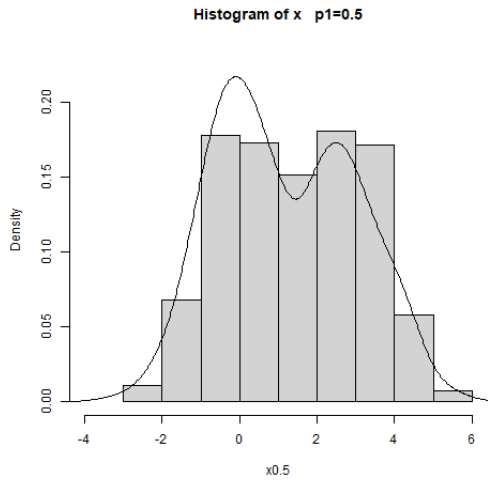

图 6: 练习 3.11 中 $p_1 = 0.5$ 时叠加了的密度曲线的直方图

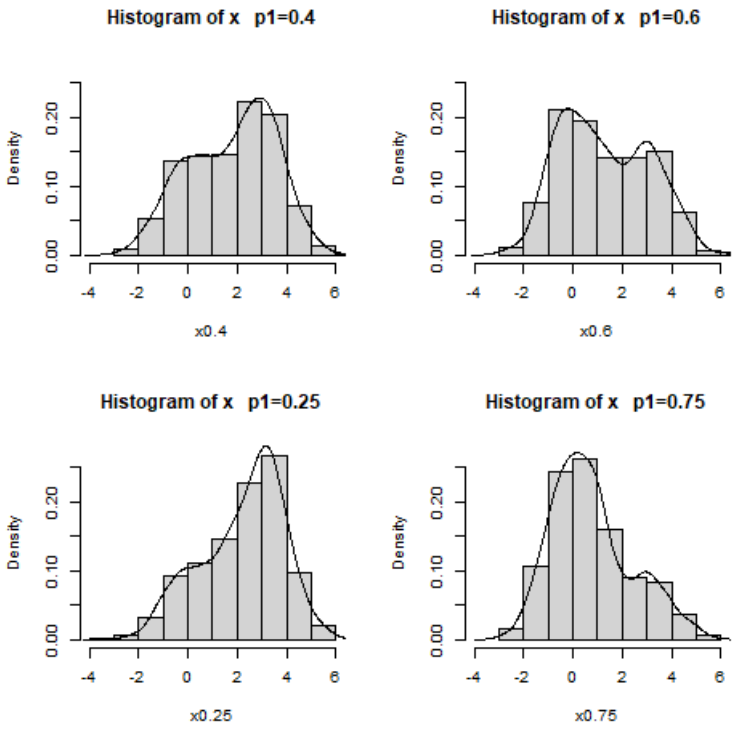

图 7: 练习 3.11 中 $p_1 = 0.4, 0.6, 0.25$ 和 0.75 时叠加了的密度曲线的直方图

结论 由图 6和图 7可知，当 $p_1 \approx 0.5$ 时，混合变量的经验分布大致服从双峰分布；当 p_1 比较接近于 0 或 1 时，混合变量的经验分布没有明显服从双峰分布的倾向.

3.13

分析过程 按照练习 3.12 的要求，先生成服从 Gamma(r, β) 分布的观测值 λ，再由 $y = \lambda e^{-\lambda y}$ 即可生成题中指数-伽马混合变量的观测值.

因为 Pareto 分布的累积分布函数为

$$F(y) = 1 - \left(\frac{\beta}{\beta + y}\right)^r, \quad y \geq 0,$$

所以它的密度函数为

$$f(y) = F'(y) = \frac{r\beta^r}{(\beta + y)^{r+1}}, \quad y \geq 0.$$

由此即可在样本的密度直方图上叠加经验分布和理论分布的密度曲线.

Rcode

```
n <- 1000
r <- 4
beta <- 2
lambda <- rgamma(n, r, beta) # lambda~Gamma(r,beta)
y <- rexp(n, lambda) # the mixture
hist(y, prob = TRUE, breaks = 20, xlim = c(0, 10))
lines(density(y, from = 0)) # from = 0: draw the density line when
x>0
x <- seq(0, 10, 0.05)
lines(x, r*beta^r/(beta+x)^(r+1), col = "red") # density line of Pareto
```

运行结果 R 输出的图形如图 8所示.

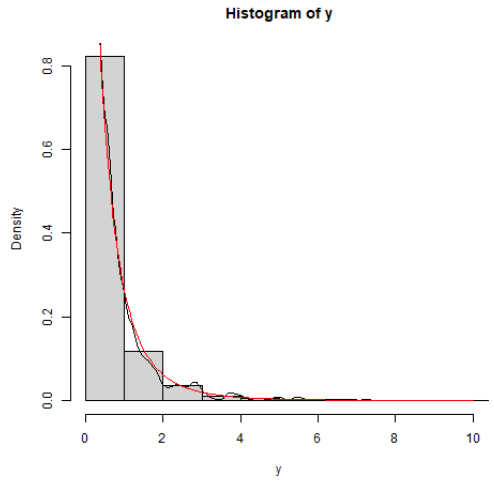

图 8: 练习 3.13 中样本的密度直方图，其中叠加了经验分布（黑色曲线）和理论（Pareto）分布（红色曲线）的密度曲线

结论 由图 8可以看出，经验和理论（Pareto）分布基本吻合.

3.14

分析过程 仿照课本例 3.18 给出的 rmvn.Choleski 函数，分别令 $n = 200, \mu = (0, 1, 2)', \Sigma = \begin{pmatrix} 1.0 & -0.5 & 0.5 \\ -0.5 & 1.0 & -0.5 \\ 0.5 & -0.5 & 1.0 \end{pmatrix}$

即可.

Rcode

```
rmvn.chol <- function(n, mu, Sigma) {
d <- length(mu)
Q <- chol(Sigma)
Z <- matrix(rnorm(n*d), nrow = n, ncol = d)
X <- Z%*%Q + matrix(mu, n, d, byrow = TRUE)
return(X)
}
n <- 200
mu <- 0:2
Sigma <- matrix(c(1, -0.5, 0.5, -0.5, 1, -0.5, 0.5, -0.5, 1), nrow =
3)
X <- rmvn.chol(n, mu, Sigma)
pairs(X)
```

运行结果 散点图如图 9所示.

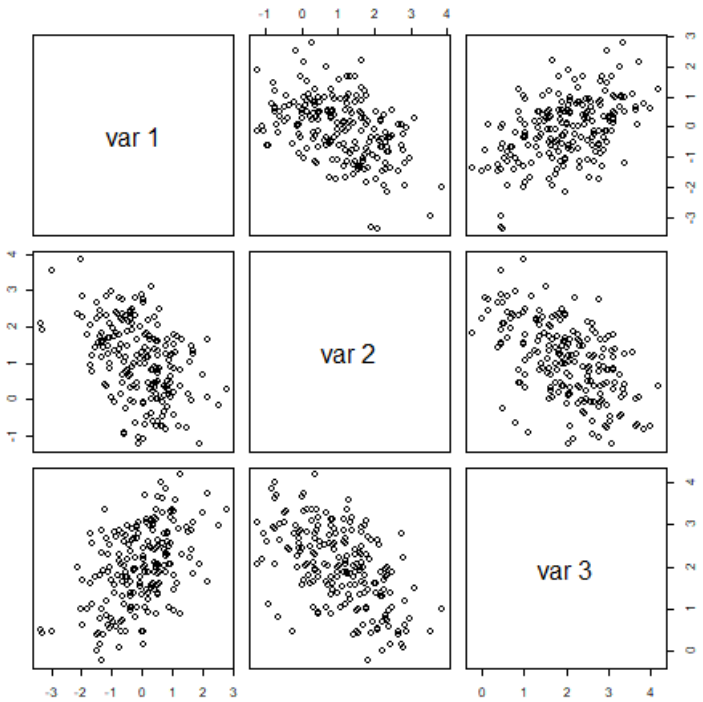

图 9: 练习 3.14 中模拟的三维多元正态随机样本的二元边缘分布散点图

结论 从直观上来说，图 9中每一对边缘分布的联合分布都大致显示出多元正态分布的椭圆对称性，且第一个分量与第三个分量呈正相关，其余两对呈负相关，这与协方差矩阵吻合.

3.16

分析过程 依练习 15 中的定义，将样本矩阵 X（X 的每一行都是一组样本观测值）标准化即对 X 作某种变换，使得变换后新的样本矩阵 Z 的样本均值为零向量，样本协方差阵为单位矩阵. 因为 X 的样本协方差阵 S_X 是正定阵，所以可以对 S_X 作 Choleski 分解，使得 $S_X = Q'Q$. 下面我们证明，

$$Z = (X - J\bar{X}')Q^{-1}$$

即为对 X 作标准化变换后新的样本矩阵，其中 J 为 n 维全一向量（n 为 X 的行数），$\bar{X}$ 为样本均值向量.

由多元统计知识，$\bar{X} = \frac{1}{n}X'J$，$S_X = \frac{1}{n-1}(X'X - n\bar{X}\bar{X}') = \frac{1}{n-1}X'\left(I - \frac{1}{n}JJ'\right)X$. 于是 Z 的样本均值为

$$\bar{Z} = \frac{1}{n}Z'J = \frac{1}{n}(Q')^{-1}(X' - \bar{X}J')J = 0, \tag{1}$$

(1) 式的最后一个等号成立是因为 $(X' - \bar{X}J')J = X'J - \bar{X}J'J = X'J - \frac{1}{n}X'JJ'J = X'J - X'J = 0$. Z 的样本方差为

$$\begin{aligned} S_Z &= \frac{1}{n-1}(Z'Z - n\bar{Z}\bar{Z}') \\ &= \frac{1}{n-1}Z'Z \\ &= \frac{1}{n-1}(Q')^{-1}(X' - \bar{X}J')(X - J\bar{X}')Q^{-1}, \end{aligned} \tag{2}$$

而

$$\begin{aligned} &(X' - \bar{X}J')(X - J\bar{X}') \\ =&X'X - X'J\bar{X}' - \bar{X}J'\bar{X} + \bar{X}J'J\bar{X}' \\ =&X'X - \frac{1}{n}\bar{X}'JJ'\bar{X} - \frac{1}{n}\bar{X}'JJ'\bar{X} + \frac{1}{n^2}X'JJ'JJ'X \\ =&X'X - \frac{1}{n}X'JJ'X \\ =&(n-1)S_X, \end{aligned} \tag{3}$$

把 (3) 式的结果代入 (2) 式，即得

$$\begin{aligned} S_Z &= \frac{1}{n-1}(Q')^{-1}(n-1)S_XQ^{-1} \\ &= (Q')^{-1}Q'QQ^{-1} \\ &= I, \end{aligned} \tag{4}$$

即 Z 的样本协方差阵为单位矩阵. 这样就证明了 Z 即为将 X 标准化后的矩阵.

为了方便用 R 对给定的样本实现上述变换，我们编写一个 norm.std 函数，用这个函数分别把学生的考试成绩样本 (X_1, X_2)（闭卷）和 (X_3, X_4, X_5)（开卷）标准化. 此外，在实际操作中，由于计算精度的限制，变换后样本矩阵的协方差阵往往并不是一个严格的单位阵. 协方差阵的非对角元有时会出现一些微小的量（一般小于 10^{-10} 量级）. 为了使最后输出的结果更为简洁，我们编写一个 rm.minor 函数用于剔除绝对值过小的量. 具体的 R 语言实现过程见“Rcode”部分.

Rcode

```
norm.std <- function(X) { # standardize sample matrix X to Z so that
# Z can be treated as a sample matrix from N(0,I)
```

```
n <- nrow(X) # number of samples
Xbar <- apply(X, 2, mean) # sample mean
J <- rep(1,n) # vector of ones with dimension n
S <- cov(X) # sample covariance matrix
Q <- chol(S) # Choleski decomposition: S=Q'Q
Z <- (X-J%*%t(Xbar))%*%solve(Q)
return(Z)
}
rm.minor <- function(X, min = 1e-10) { # remove the values whose
# absolute values are too small(< min)
X[which(abs(X)<min)] <- 0
return(X)
}
mydata <- read.csv("scor.csv")
X <- as.matrix(mydata)
X12 <- X[,1:2]
Z12 <- norm.std(X12) # standardize (X1,X2)
X345 <- X[,3:5]
Z345 <- norm.std(X345) # standardize (X3,X4,X5)
Z <- cbind(Z12,Z345)
rm.minor(apply(Z, 2, mean))
rm.minor(cov(Z))
```

运行结果 运行代码，最后两行输出的结果为

```
> rm.minor(apply(Z, 2, mean))
   mec  vec  alg  ana  sta
     0    0    0    0    0
> rm.minor(cov(Z))
                mec         vec       alg          ana           sta
mec    1.00000000  0.00000000  0.5633315  0.02946846  -0.00445176
vec    0.00000000  1.00000000  0.3558405  0.06883101   0.03154961
alg    0.56333148  0.35584053  1.0000000  0.00000000   0.00000000
ana    0.02946846  0.06883101  0.0000000  1.00000000   0.00000000
sta   -0.00445176  0.03154961  0.0000000  0.00000000   1.00000000
```

结论 rm.minor(apply(Z, 2, mean)) 给出的结果即考试成绩样本的均值向量，rm.minor(cov(Z)) 给出的结果即考试成绩样本的协方差矩阵. 可见 (X_1, X_2)（对应 mec 和 vec）和 (X_3, X_4, X_5)（对应 alg，ana 和 sta）对应的样本确实已分别标准化，因为它们的均值向量为 0，在样本的协方差矩阵中对应的子矩阵也都是单位阵.