STAT571 HW#9
Tianqi Wu

1.(a)

```
> library(msos)
> school = read.csv("schooltest.csv",header=T)
> Y = school[,-(1:2)]
> Bigten = as.matrix(subset(Y,school$PAC12 == 0))
> Pac12 = as.matrix(subset(Y,school$PAC12 == 1))
> dfbig = nrow(Bigten)-1
> dfpac = nrow(Pac12)-1
> dfbig
[1] 12
> dfpac
[1] 11
```

1.(b)

```
> covbig = cov(Bigten)
> covpac = cov(Pac12)
> covpool=(covbig*dfbig+covpac*dfpac)/(dfbig+dfpac)
> det(covbig)
[1] 694862652
> det(covpac)
[1] 37075419
> det(covpool)
[1] 1248333751
```

1.(c)

```
> chi_square = (dfbig+dfpac)*log(det(covpool))-
dfbig*log(det(covbig))-dfpac*log(det(covpac))
> chi_square
[1] 45.71293
> q = ncol(Bigten)
> dfchi_square = q*(q+1)/2
> dfchi_square
[1] 21
> p_value=1-pchisq(chi_square,dfchi_square)
> p_value
[1] 0.001394944
```

$2\ln(LR) = 45.71293$. Degree of freedom is 21. Since p_value = 0.00139 < 0.05, we reject the null hypothesis and conclude that we have found a significant difference between the covariance matrices.

1.(d)

```
> tr(covbig)
[1] 9323.846
> tr(covpac)
[1] 12320.89
> cov2cor(covbig)
           SATERW25  SATERW75 SATMath25 SATMath75     ACT25     ACT75
SATERW25  1.0000000 0.9498267 0.9381554 0.8968812 0.9610584 0.9308565
SATERW75  0.9498267 1.0000000 0.8830933 0.8940852 0.8819031 0.9042136
SATMath25 0.9381554 0.8830933 1.0000000 0.9692639 0.8407831 0.8483448
SATMath75 0.8968812 0.8940852 0.9692639 1.0000000 0.7817800 0.8142107
ACT25     0.9610584 0.8819031 0.8407831 0.7817800 1.0000000 0.9396409
ACT75     0.9308565 0.9042136 0.8483448 0.8142107 0.9396409 1.0000000
> cov2cor(covpac)
           SATERW25  SATERW75 SATMath25 SATMath75     ACT25     ACT75
SATERW25  1.0000000 0.9734579 0.9780136 0.9377345 0.9722390 0.9828367
SATERW75  0.9734579 1.0000000 0.9490785 0.9707431 0.9280283 0.9768370
SATMath25 0.9780136 0.9490785 1.0000000 0.9506063 0.9612023 0.9672796
SATMath75 0.9377345 0.9707431 0.9506063 1.0000000 0.9228812 0.9682407
ACT25     0.9722390 0.9280283 0.9612023 0.9228812 1.0000000 0.9823818
ACT75     0.9828367 0.9768370 0.9672796 0.9682407 0.9823818 1.0000000
```

The trace of covariance matrix for the Pac-12 scores is larger than that for the Big Ten scores. From the sample correlation matrices, we see that Pac-12 has overall higher correlation than that of Big Ten.

2.(a)

```
> grades=data.frame(grades)
> male = grades[grades$Gender==0,2:6]
> female = grades[grades$Gender==1,2:6]
> dfmale = nrow(male)-1
> dffemale = nrow(female)-1
> Spool = (cov(male)*dfmale+cov(female)*dffemale)/
(dfmale+dffemale)
> sigma11 = Spool[1:3,1:3]
> sigma22 = Spool[4,4]
> sigma = Spool[1:4,1:4]
> det(sigma11)
[1] 5760642
> sigma22
[1] 85.68988
> det(sigma)
[1] 400373347
> chisq = 105*(log(det(sigma11))+log(sigma22)-log(det(sigma)))
> chisq
```

```
[1] 21.98554
> dfchisq = 3*1
> dfchisq
[1] 3
> p_value=1-pchisq(chisq,dfchisq)
> p_value
[1] 6.568465e-05
```

2ln(LR) = 21.98554 and degree of freedom is 3. Since p_value = 6.568465e-05 < 0.05, we reject the null and conclude that the two sets of scores are not independent.

2.(b)

```
> condS = sigma-(Spool[1:4,5]/Spool[5,5])%*%t(Spool[1:4,5])
> cond_sigma11 = condS[1:3,1:3]
> cond_sigma22 = condS[4,4]
> cond_sigma = condS[1:4,1:4]
> det(cond_sigma11)
[1] 4699937
> cond_sigma22
[1] 55.14605
> det(cond_sigma)
[1] 242785402
> cond_chisq = (105-1)*(log(det(cond_sigma11))
+log(cond_sigma22)-log(det(cond_sigma)))
> cond_chisq
[1] 6.797055
> dfcon_chisq = 3*1
> dfcon_chisq
[1] 3
> p_value=1-pchisq(cond_chisq,dfcon_chisq)
> p_value
[1] 0.07865548
```

2ln(LR) = 6.797055 and degree of freedom is 3. Since p_value = 0.07865548 > 0.05, we accept the null and conclude that the two sets of scores are conditional independent on final score.

2.(c)

```
> cov2cor(Spool[1:4,1:4])
                HW      Labs   InClass  Midterms
HW       1.0000000 0.7810389 0.2785350 0.4108988
Labs     0.7810389 1.0000000 0.4205645 0.3793224
InClass  0.2785350 0.4205645 1.0000000 0.2410711
Midterms 0.4108988 0.3793224 0.2410711 1.0000000
> cov2cor(condS[1:4,1:4])
                HW      Labs   InClass  Midterms
HW       1.0000000 0.7472753 0.1952600 0.2359618
Labs     0.7472753 1.0000000 0.3629685 0.2287113
InClass  0.1952600 0.3629685 1.0000000 0.1061060
Midterms 0.2359618 0.2287113 0.1061060 1.0000000
```

The unconditional and conditional correlation matrices are given above. We can see that the overall correlation for the conditional one is smaller than that for the unconditional one. Thus it appears that the final scores explain some relationships among the other variables.

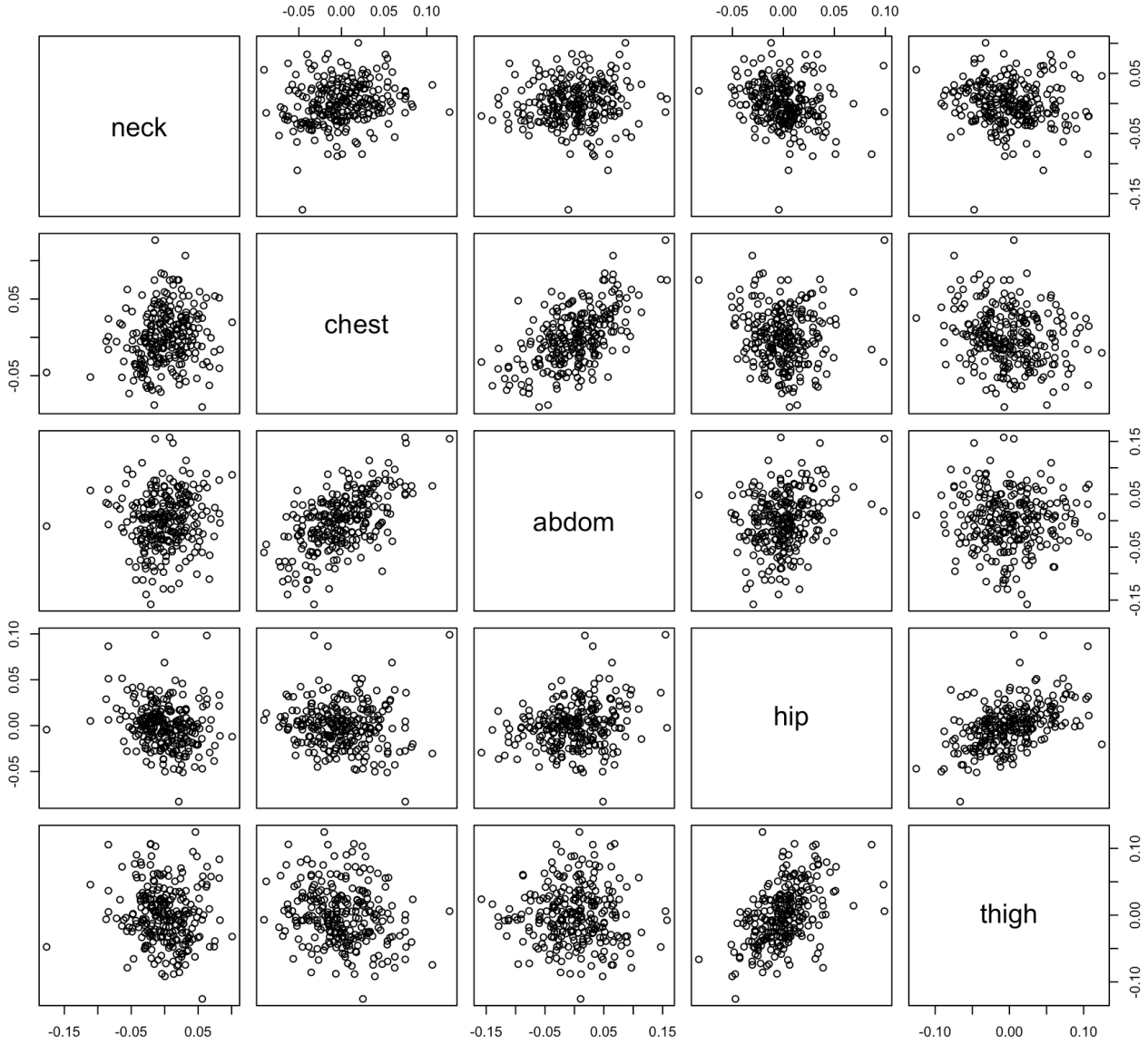
3.(a)

Since we have 5 variables in total, to make ((5-p)^2-5-p)/2 >= 0, at most p = 2 can be estimated.

3.(b)

```
> library(faraway)
> data(fat)
> Y =
cbind(log(fat$neck),log(fat$chest),log(fat$abdom),log(fat$hip),l
og(fat$thigh))
> lm_model= lm(Y~log(fat$weight))
>
pairs(lm_model$residuals,labels=c(‘neck','chest','abdom','hip','
thigh'))
```

From the plot generated below, we can see that chest and abdom may have positive linear relationship. Also, hip and thigh may have positive linear relationship.



3.(c)

```
> sigmahat = cov(lm_model$residuals)
> v = 252-5
> p=1
> q=5
> f = factanal(covmat=sigmahat,factors=1,n.obs=v+1)
> corr0 = f$loadings%*%t(f$loadings) + diag(f$uniquenesses)
> chisq = (v - (2*q+5)/6-2*p/3)*log(det(corr0)/det(f$corr))
> chisq
[1] 104.4932
> dfchisq = ((q - p)^2 - p - q)/2
> dfchisq
[1] 5
> p_value=1-pchisq(chisq,dfchisq)
```

```
> p_value
[1] 0
```

2ln(LR) = 104.4932 and degree of freedom is 5. Since p_value = 0 < 0.05, we reject the null and conclude that one factor is not enough.

3.(d)

```
> f2 = factanal(covmat=sigmahat,factors=2,n.obs=v+1)
> f2$uniquenesses
[1] 0.9234104 0.1872022 0.5070444 0.0050000 0.7309094
> print(f2$loadings,cutoff=0)

Loadings:
     Factor1 Factor2
[1,] -0.240   0.139
[2,] -0.263   0.862
[3,]  0.089   0.696
[4,]  0.973   0.222
[5,]  0.512  -0.081

                Factor1 Factor2
SS loadings       1.343   1.304
Proportion Var    0.269   0.261
Cumulative Var    0.269   0.529
```

“Neck” has the highest uniqueness (0.9234). Also, it has fairly low loadings on both factors. The first factor loads highly on “hip” and “thigh”. The second factor loads highly on “chest” and “abdom”. Hence, we can see first factor describing the lower middle part of the body and second factor describing the upper middle part of the body.

3.(e)

```
> p=2
> corr0 = f2$loadings%*%t(f2$loadings) + diag(f2$uniquenesses)
> chisq = (v - (2*q+5)/6-2*p/3)*log(det(corr0)/det(f2$corr))
> chisq
[1] 0.9035375
> dfchisq = ((q - p)^2 - p - q)/2
> dfchisq
[1] 1
> p_value=1-pchisq(chisq,dfchisq)
> p_value
[1] 0.341835
```

2ln(LR) = 0.9035375 and degree of freedom is 1. Since p_value = 0.341835 > 0.05, we accept the null and conclude that two-factor model fits fine.