A1.

```
nordic = read.table('Nordic.txt', header = TRUE)
nordic.pc = prcomp(nordic[, c(4,5)], scale = TRUE)
nordic.pc
```

```
## Standard deviations (1, .., p=2):
## [1] 1.005286 0.994686
##
## Rotation (n x k) = (2 x 2):
##                      PC1       PC2
## SkiJump      -0.7071068 0.7071068
## CrossCountry  0.7071068 0.7071068
```

A2. In the competition, it is better to have higher score of SkiJump and shorter time of CrossCountry. We can see that PC1 loading value is negatively correlated with SkiJump while positively correlated with CrossCountry. That means PC1 negatively correlates to the overall performance. Hence, PC1 contains the maximum variance. So it suggests that PC1 is a good way of combining the scores.

A3.

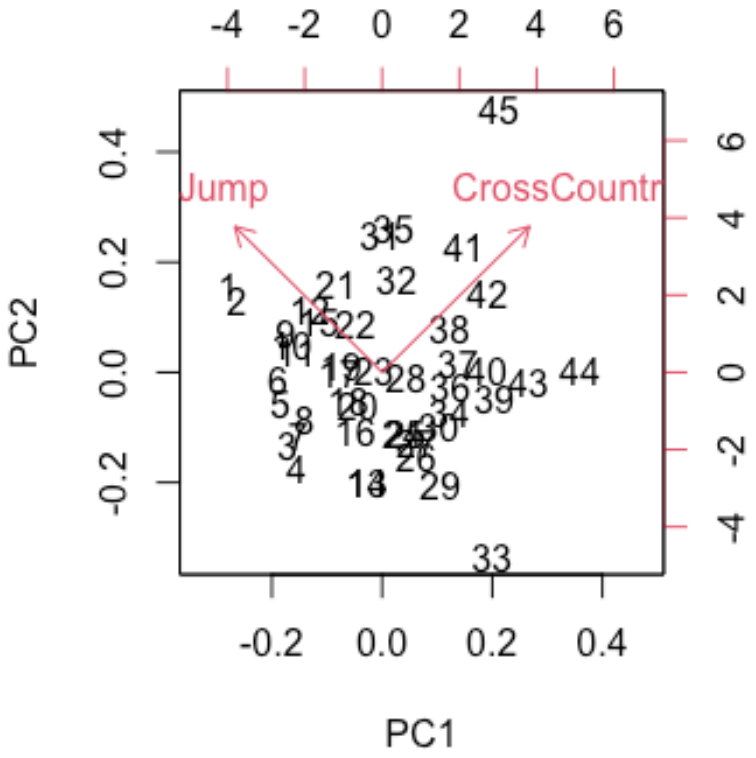
```
scores = nordic.pc$x
nordic[rank(scores[,1])==3,]
```

```
##   Nat    first   Name SkiJump CrossCountry
## 6 GER Johannes RYDZEK   121.2       1406.5
```

We can see: Johannes RYDZEK would have won the bronze medal based on the first principal component.

A4.

```
biplot(nordic.pc)
```



As SkiJump and CrossCountry are equally and positively correlated with PC2 while SkiJump and CrossCountry better has higher and shorter values respectively, PC2 could represent the athletes who are good at one activity but bad at the other.

A5.

```
summary(nordic.pc)

## Importance of components:
##                           PC1    PC2
## Standard deviation     1.0053 0.9947
## Proportion of Variance 0.5053 0.4947
## Cumulative Proportion  0.5053 1.0000
```

The first principal component only summarize 50% of the variance. So one principal component can not summarize the whole data.

A6.

```
cor(nordic[, c(4,5)])

##                   SkiJump CrossCountry
## SkiJump       1.00000000  -0.01059985
## CrossCountry -0.01059985   1.00000000
```

SkiJump and CrossCountry are not really correlated. So it is not reasonable to drop the Nordic combined.

A7.

```
nordic.pcv = prcomp(nordic[, c(4,5)])
scores2 = nordic.pcv$x
nordic[rank(scores2[,1])==1,]

##   Nat      first   Name SkiJump CrossCountry
## 4 ITA Alessandro PITTIN   113.4       1367.5
```

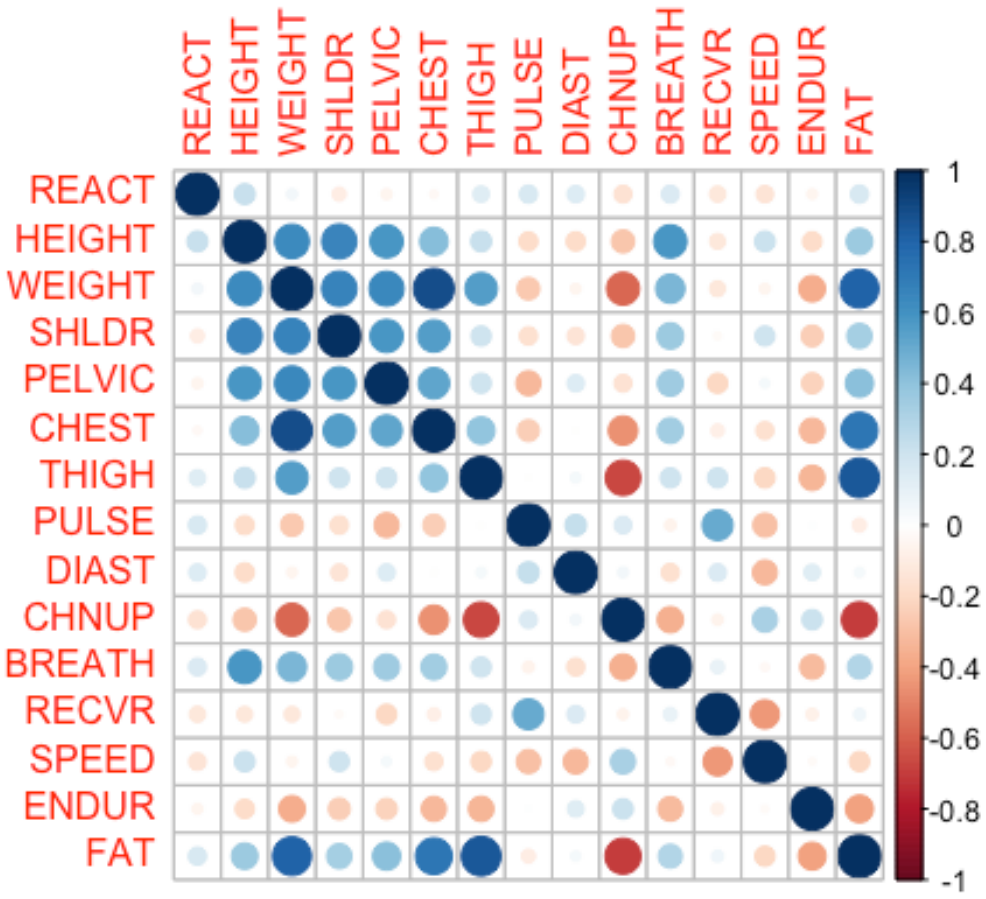
As the value of CrossCountry (more than 1000) is significantly higher than that of SkiJump(a little more than 100), it is better to standardize the data. So running the PCA on the correlation matrix would be better than running it on the covariance matrix. Otherwise, the measurement in CrossCountry will count more in the PCA result. We can see, although Alessandro PITTIN had a relatively lower score in SkiJump, he got the shortest time in CrossCountry. As a result, Alessandro PITTIN would be the gold medallist based on first PCA component on the variance matrix.

B1.

```
police = read.csv("police.csv", header=TRUE)
police.scale <- scale(police, center=TRUE, scale=TRUE)
library(corrplot)

## corrplot 0.84 loaded

corrplot(cor(police))
```




```
round(sapply(1:6, function(i) factanal(police.scale, factors=i)$PVAL), 3)

## objective objective objective objective objective objective
##     0.000     0.000     0.001     0.018     0.071     0.173
```

Using hypotheses testing with p<=0.05, 5 factors can be found.

B2.

```
fa <- factanal(police.scale, factors=5)
apply(fa$loadings[,c(1,2)] > 0.5, 2, function(x) names(police)[x])

## $Factor1
## [1] "WEIGHT" "THIGH"  "FAT"
##
## $Factor2
## [1] "HEIGHT" "WEIGHT" "SHLDR"  "PELVIC" "BREATH"
```

“WEIGHT” “CHEST” “THIGH” “FAT” are grouped by the first factor. “HEIGHT” “WEIGHT” “SHLDR” “PELVIC” “BREATH” are grouped by the second factor.

B3.

```
cor(police$DIAST, police$PULSE)
```

```
## [1] 0.2340876
```

```
fa$loadings["DIAST",]
```

```
##     Factor1     Factor2     Factor3     Factor4     Factor5
##  0.03749942 -0.16572482  0.22973048  0.16647210  0.14203174
```

```
fa$uniquenesses
```

```
##      REACT     HEIGHT     WEIGHT      SHLDR     PELVIC      CHEST      THI
GH
## 0.37042187 0.10916745 0.02760498 0.31282635 0.48487803 0.08118617 0.055173
64
##      PULSE      DIAST      CHNUP     BREATH      RECVR      SPEED      END
UR
## 0.62105039 0.87047648 0.46530671 0.58728093 0.00500000 0.52175456 0.825902
50
##        FAT
## 0.05769859
```

Although DIAST only contributes little to each factor, it has a low correlation(0.2340876) with PULSE which means DIAST can not be replaced by PULSE. Looking at the uniqueness of DIAST(0.87047648), it suggests DIAST is a very unique measurement. Unfortunately, DIAST ie included in any factor when the loading threshold is set 0.5.That is probably one of the drawbacks of factor analysis.

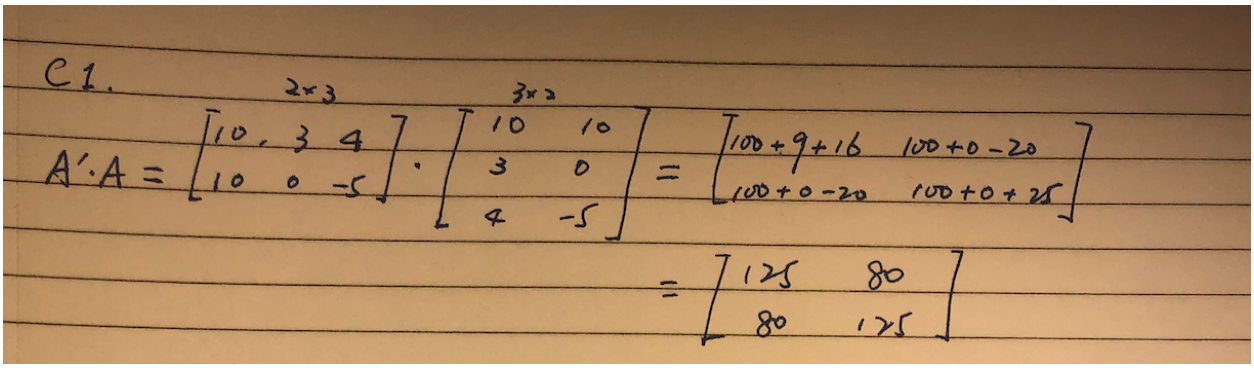
B4.

```
apply(fa$loadings[,c(1,2,3,4,5)] > 0.5, 2, function(x) names(police)[x])
```

```
## $Factor1
## [1] "WEIGHT" "THIGH"  "FAT"
##
## $Factor2
## [1] "HEIGHT" "WEIGHT" "SHLDR"  "PELVIC" "BREATH"
##
## $Factor3
## [1] "PULSE" "RECVR"
##
## $Factor4
## [1] "CHEST"
##
## $Factor5
## [1] "REACT"
```

As athletes always have low RECVR and PULSE measurements while other variables are not guaranteed on athletes, for example, an athlete could be high or short, we can see Factor 3 is a perfect one to use for separating athletic applicants from non-athletic applicants.

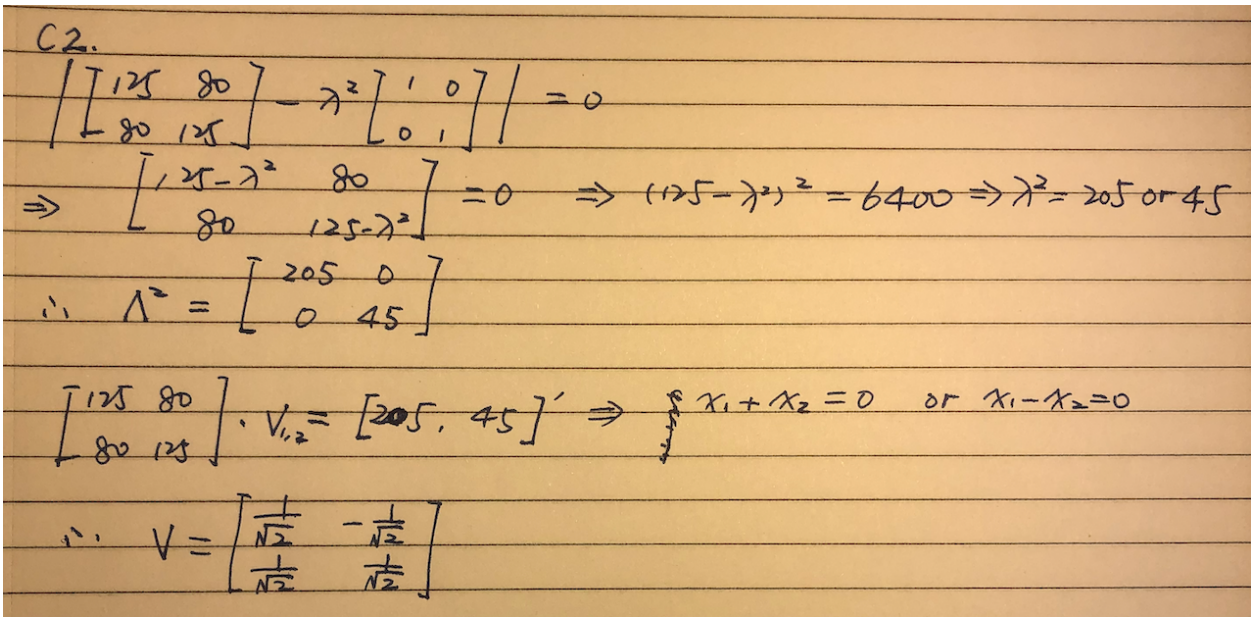
C1.




```
A= matrix(c(10,3,4,10,0,-5), ncol = 2)
crossprod(A)
```

```
##      [,1] [,2]
## [1,]  125   80
## [2,]   80  125
```

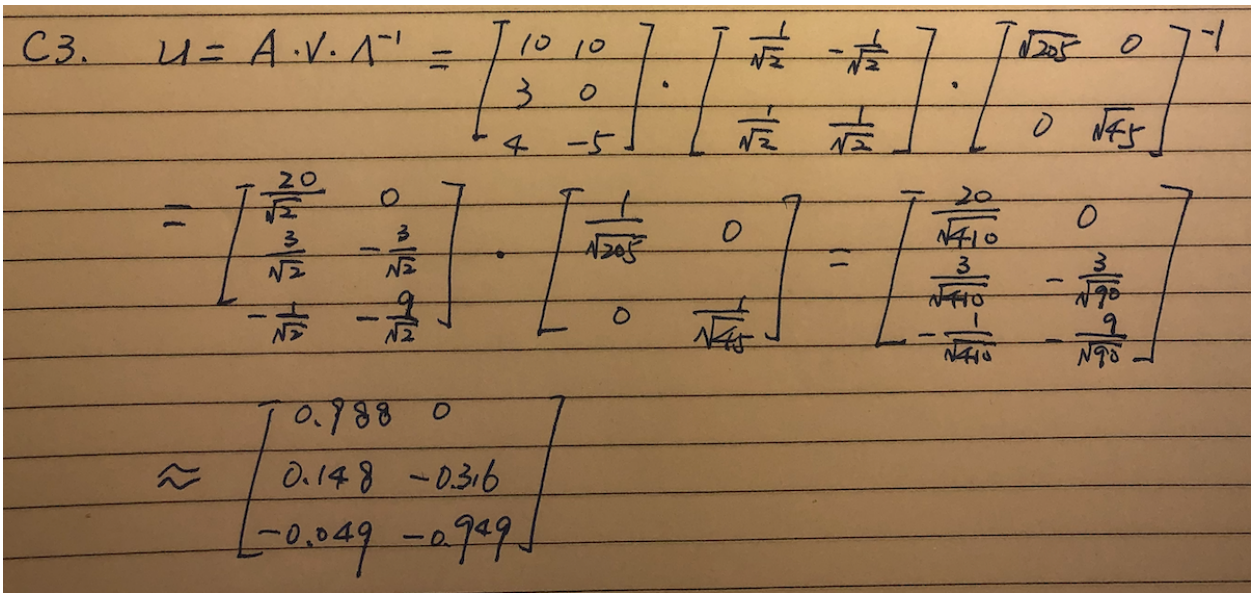
C2.




```
eigen(crossprod(A))
```

```
## eigen() decomposition
## $values
## [1] 205  45
##
## $vectors
##           [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
```

C3.




```
L = matrix(c(sqrt(205),0,0,sqrt(45)), nrow = 2)
L

##          [,1]     [,2]
## [1,] 14.31782 0.000000
## [2,]  0.00000 6.708204

V = eigen(crossprod(A))$vectors
U = A %*% V %*% solve(L)
U

##             [,1]       [,2]
## [1,]  0.98772960  0.0000000
## [2,]  0.14815944 -0.3162278
## [3,] -0.04938648 -0.9486833
```

C4. The 3 observations have the maximum variance in the direction of the eigen vectors. The eigen values explain how much of the variance in the whole sample data is explained by the relative principal component. In this case, the first principal component is at the direction of [0.707, 0.707], with a proportion of 14.3/(14.3 + 6.7) of the whole sample variance. The second principal component is at the direction of [-0.707, 0.707] with a proportion of 6.7/(14.3 + 6.7) of the whole sample variance.