STA437 Assignment #2

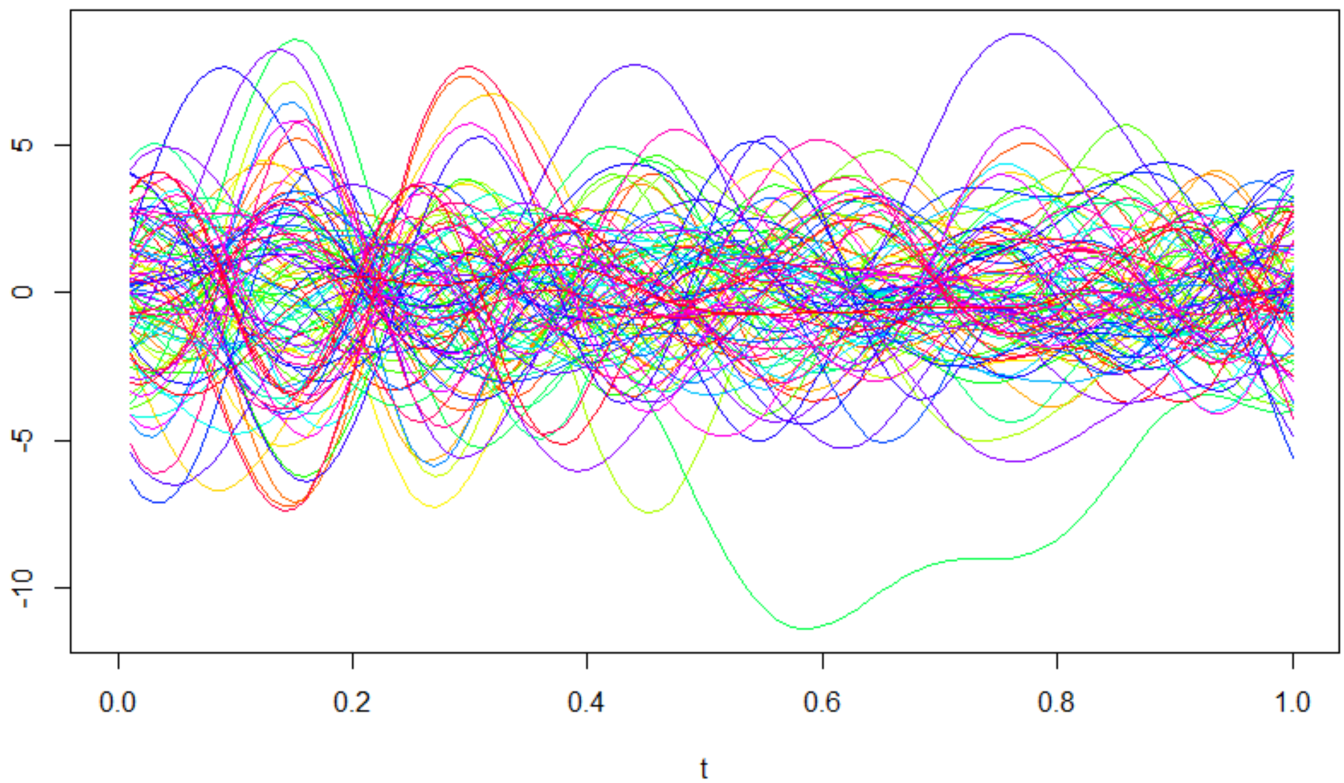
Rui Qiu #999292509

2016-02-25

Problem 1

(a) Solution:

```
> source("andrews.txt")
> data <- read.csv(file="./testdata.txt",head=FALSE,sep=" ")
> r <- andrews(data,scale=T)
```



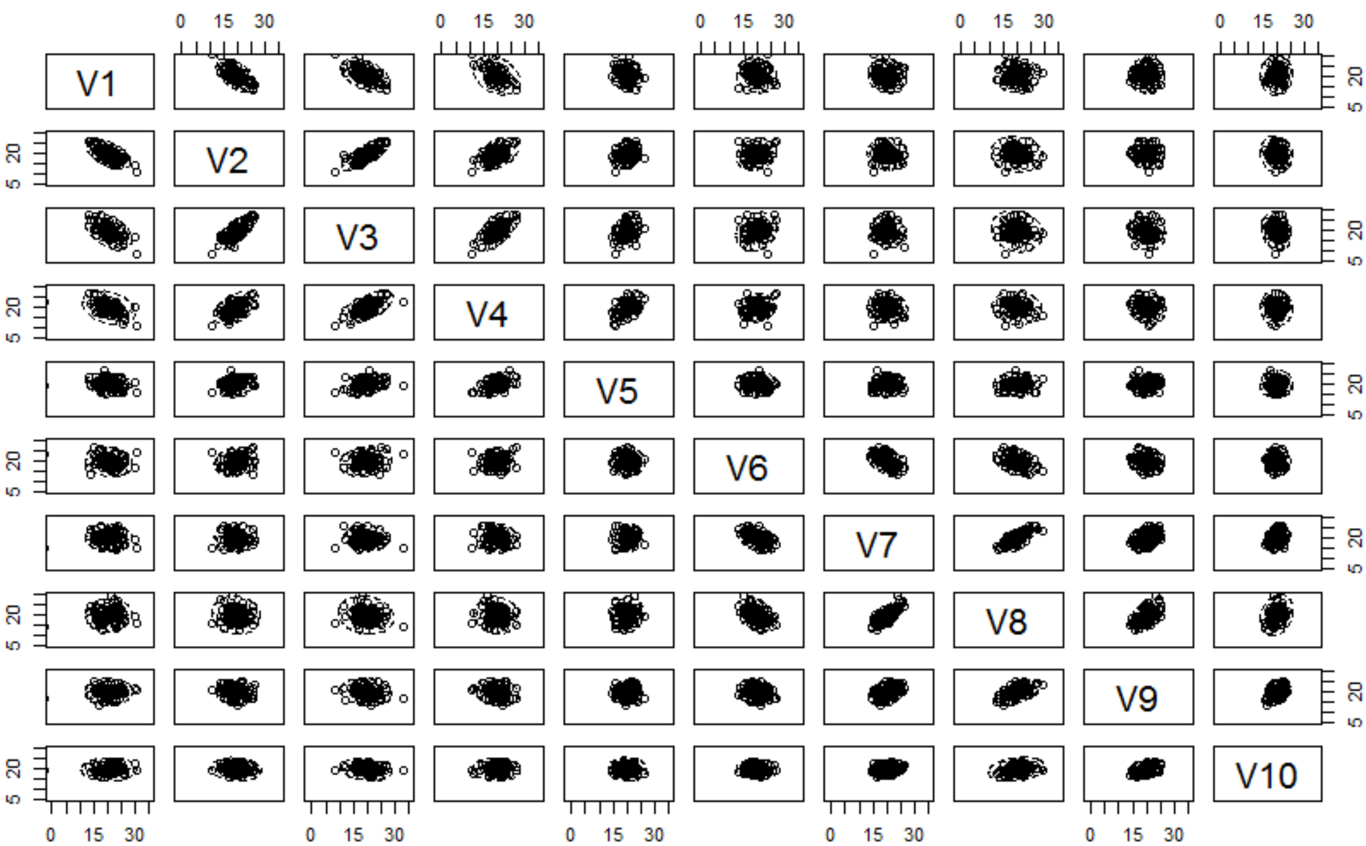
By observing the Andrew curves, I think there are two outliers:

- the purple curve on top right which is clearly higher than other curves,
- the green curve below.

(b) Solution:

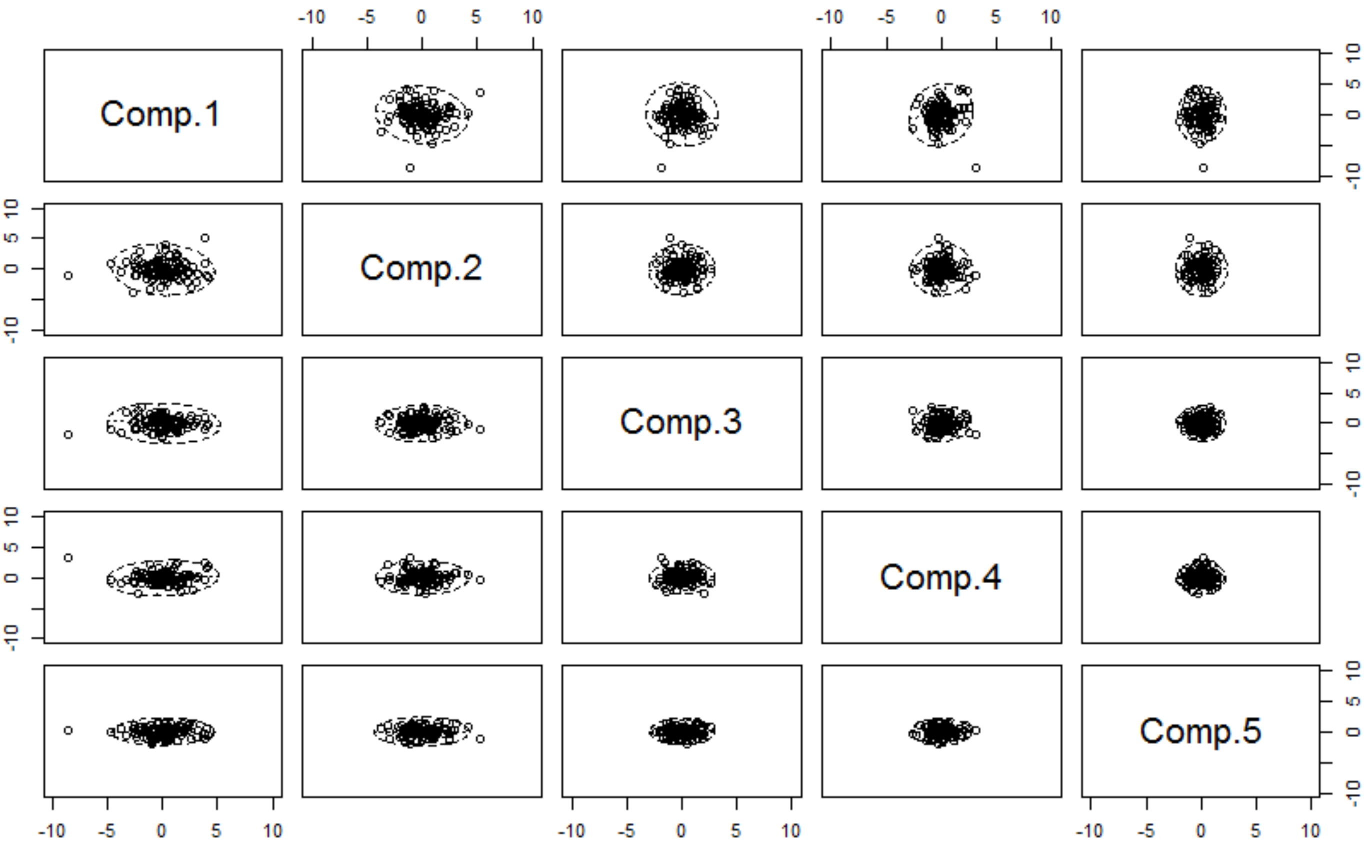
We use package `MVA` as a helper.

```
> library(MVA)
> pairs(data,xlim=c(-1,35),ylim=c(5,30),panel=function(x,y,...)
{bvbox(cbind(x,y), add=TRUE)})
```




```
> data2 <- princomp(data,cor=T)
> summary(data2,loading=T)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5     Comp.6
Standard deviation     1.8828405 1.5342699 1.0573374 1.0096311 0.78564604 0.68918510
Proportion of Variance 0.3545088 0.2353984 0.1117962 0.1019355 0.06172397 0.04749761
Cumulative Proportion  0.3545088 0.5899073 0.7017035 0.8036390 0.86536296 0.91286057
                           Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     0.60997372 0.46189758 0.40842699 0.34520191
Proportion of Variance 0.03720679 0.02133494 0.01668126 0.01191644
Cumulative Proportion  0.95006737 0.97140230 0.98808356 1.00000000

Loadings:
    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
V1   0.390  0.257  0.149 -0.333                0.477 -0.532  0.294  0.222
```

```
V2  -0.432 -0.228 -0.140  0.298        -0.139        -0.334  0.229  0.674
V3  -0.451 -0.235                      -0.118  0.325 -0.467        -0.626
V4  -0.373 -0.249        -0.356  0.251  0.205  0.574  0.420         0.229
V5  -0.165 -0.279  0.298 -0.702 -0.171        -0.505 -0.122  0.109       
V6  -0.243  0.300 -0.466 -0.141 -0.452  0.625                0.116       
V7   0.259 -0.465  0.150  0.258         0.414         0.110  0.643 -0.152
V8   0.284 -0.474               -0.181  0.386        -0.274 -0.638  0.149
V9   0.247 -0.343 -0.378 -0.106 -0.567 -0.459  0.227  0.268  0.109       
V10  0.155 -0.195 -0.692 -0.261  0.578        -0.128 -0.191              
> pairs(data2$scores[,1:5],xlim=c(-10,10),ylim=c(-10,10),panel=function(x,y,...)
{bvbox(cbind(x,y),add=TRUE)})
```



The pairwise scatterplots with 10 variables is too messy to observe, so we take PCA (pick the most important 5 components). We want to count the number of data points that lie outside the circle (that's why we use `MVA`). And we tend to believe the scatterplots of variables with smaller number (such as V1 versus V2), so there are 2 data points clearly outside the circle.

So we believe there are 2 outliers.

Problem 2

(a) Solution:
Suppose $\{g_i(t)\}$ are the Andrew curves defined in problem 1, so we have:

$$g_i(t) = \frac{1}{\sqrt{2}} x_{i1} + x_{i2} \sin(2\pi t) + x_{i3} \cos(2\pi t) + x_{i4} \sin(4\pi t) + x_{i5} \cos(4\pi t) + \cdots$$
$$g_j(t) = \frac{1}{\sqrt{2}} x_{j1} + x_{j2} \sin(2\pi t) + x_{j3} \cos(2\pi t) + x_{j4} \sin(4\pi t) + x_{j5} \cos(4\pi t) + \cdots$$

$$\begin{aligned} 2\int_0^1 [g_i(t) - g_j(t)]^2 dt = 2\int_0^1 [&\frac{1}{\sqrt{2}} (x_{i1} - x_{j1}) + (x_{i2} - x_{j2}) \sin(2\pi t) + (x_{i3} - x_{j3}) \cos(2\pi t) \\ &+ (x_{i4} - x_{j4}) \sin(4\pi t) + (x_{i5} - x_{j5}) \cos(4\pi t) + \cdots]^2 dt \cdots \cdots \cdots (\star) \end{aligned}$$

Before fully expand the RHS and compute, we can take a bite at the first two terms, and try to find some pattern to simplify our calculation.

First term times itself:

$$\begin{aligned} 2\int_0^1 [\frac{1}{\sqrt{2}} (x_{i1} - x_{j1})]^2 dt &= 2\int_0^1 \frac{1}{2} (x_{i1} - x_{j1})^2 dt \\ &= (x_{i1} - x_{j1})^2 \end{aligned}$$

First term times the second term:

$$\begin{aligned} 2\int_0^1 \frac{1}{\sqrt{2}} (x_{i1} - x_{j1})(x_{i2} - x_{j2}) \sin(2\pi t) dt &= \sqrt{2}(x_{i1} - x_{j1})(x_{i2} - x_{j2}) \int_0^1 \sin(2\pi t) \\ &= 0 \end{aligned}$$

Second term times the second term:

$$
\begin{aligned}
2\int_0^1 [(x_{i2}-x_{j2})\sin(2\pi t)]^2 dt &= 2(x_{i2}-x_{j2})^2 \int_0^1 \sin^2(2\pi t)dt \\
&= 2(x_{i2}-x_{j2})^2 \int_0^1 [1-\cos(4\pi t)]dt \\
&= 2(x_{i2}-x_{j2})^2 \frac{1}{2} \\
&= (x_{i2}-x_{j2})^2
\end{aligned}
$$

So now we can see the basic pattern:

- if the product has a $\sin(2\pi t)$ term or $\cos(2\pi t)$ term in calculation, we integrate this term over $[0, 1]$, always get a 0 because $\int_0^1 \sin(2\pi t) = \int_0^1 \cos(2\pi t) = 0$.
- also note that if we have $\int_0^1 \sin(2\pi t)\cos(2\pi t) = \frac{1}{2}\int_0^1 \sin(4\pi t) = 0$,as well.
- if the product has a $\sin^2(2\pi t)$ term or $\cos^2(2\pi t)$ term, then integrated over $[0, 1]$, the result is always $\frac{1}{2}$.

Therefore, if we expand the equation $(\star)$, we will get $p \times p$ terms, but only p terms from producting with itself will remain, the others will be 0.

Hence,

$$
\begin{aligned}
(\star) &= (x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + (x_{i3}-x_{j3})^2 + \cdots + (x_{ip}-x_{jp})^2 \\
&= \sum_{k=1}^{P}(x_{ik}-x_{jk})^2
\end{aligned}
$$

(b) Solution:

$$
\bar{\boldsymbol{x}} = \begin{pmatrix} \bar{\boldsymbol{x_1}} \\ \bar{\boldsymbol{x_2}} \\ \vdots \\ \bar{\boldsymbol{x_p}} \end{pmatrix} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x_i} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_{i1} \\ \frac{1}{n}\sum_{i=1}^n x_{i2} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n x_{ip} \end{pmatrix}
$$

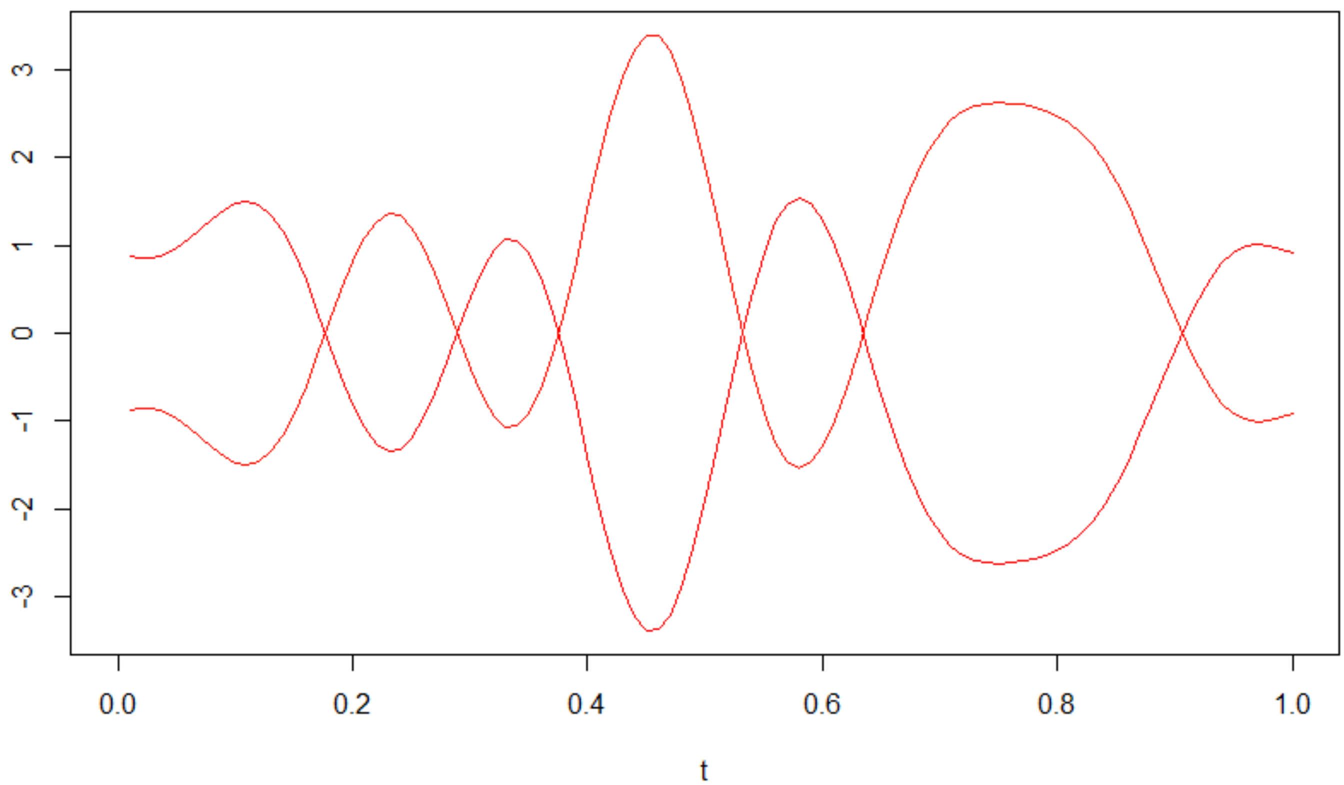
So the Andrew curve of $\bar{\boldsymbol{x}}$ can be represented as:

$$
\begin{aligned}
g(t) &= \frac{1}{\sqrt{2}}\bar{\boldsymbol{x_1}} + \bar{\boldsymbol{x_2}}\sin(2\pi t) + \bar{\boldsymbol{x_3}}\cos(2\pi t) + \bar{\boldsymbol{x_4}}\sin(4\pi t) + \bar{\boldsymbol{x_5}}\cos(4\pi t) + \cdots \\
&= \frac{1}{\sqrt{2}}\frac{1}{n}\sum_{i=1}^n x_{i1} + \frac{1}{n}\sum_{i=1}^n x_{i2}\sin(2\pi t) + \frac{1}{n}\sum_{i=1}^n x_{i3}\cos(2\pi t) + \frac{1}{n}\sum_{i=1}^n x_{i4}\sin(4\pi t) + \frac{1}{n}\sum_{i=1}^n x_{i5}\cos(4\pi t) + \cdots \\
&= \frac{1}{n}\left(\frac{1}{\sqrt{2}}\sum_{i=1}^n x_{i1} + \sum_{i=1}^n x_{i2}\sin(2\pi t) + \sum_{i=1}^n x_{i3}\cos(2\pi t) + \sum_{i=1}^n x_{i4}\sin(4\pi t) + \sum_{i=1}^n x_{i5}\cos(4\pi t) + \cdots\right)
\end{aligned}
$$

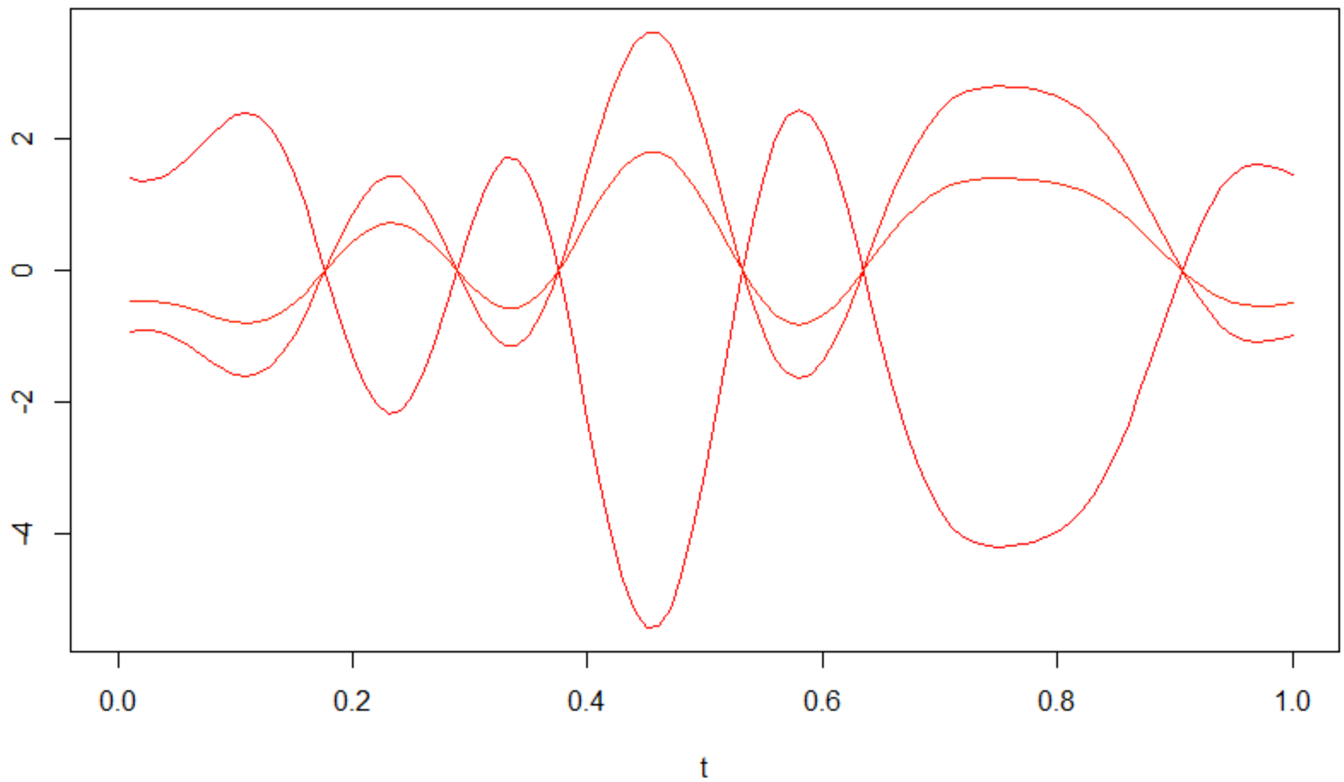
(c) Solution:

If we pick a $x_k = \lambda x_i + (1 - \lambda) x_j$ between x_i and x_j with $0 < \lambda < 1$ from the original data, and draw the Andrew curves:

```
> data3 <- data[56:57,]
> r <- andrews(data3,scale=T)
```




```
> data3[3,] <- 0.2*data3[1,]+0.8*data3[2,] # set lambda as 0.2
> r <- andrews(data3,scale=T)
```



By observing the two plots, we can find that the Andrew curve of x_k goes right between the Andrew curves of x_i and x_j.

Problem 3

(a) Solution:

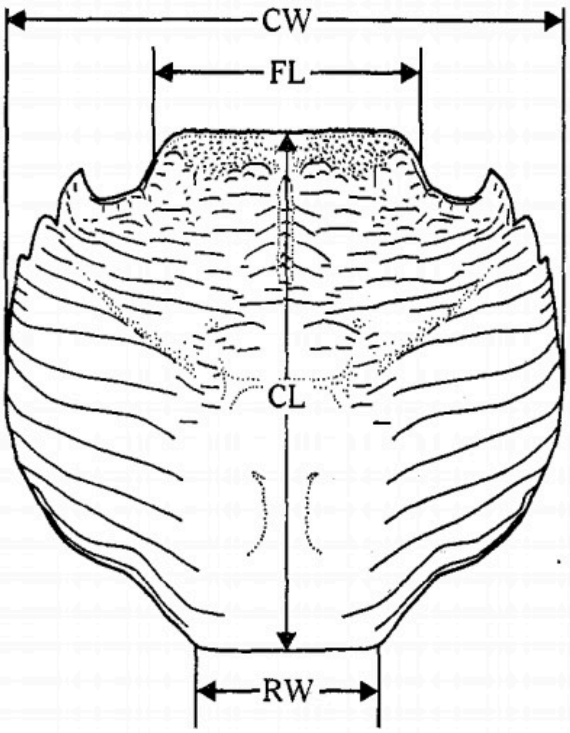


Fig. 1. Dorsal view of carapace of *Leptograpsus*, showing measurements taken. *FL*, width of frontal region just anterior to frontal tubercles. *RW*, width of posterior region. *CL*, length along midline. *CW*, maximum width. The body depth was also measured; in females but not in males the abdomen was first displaced.

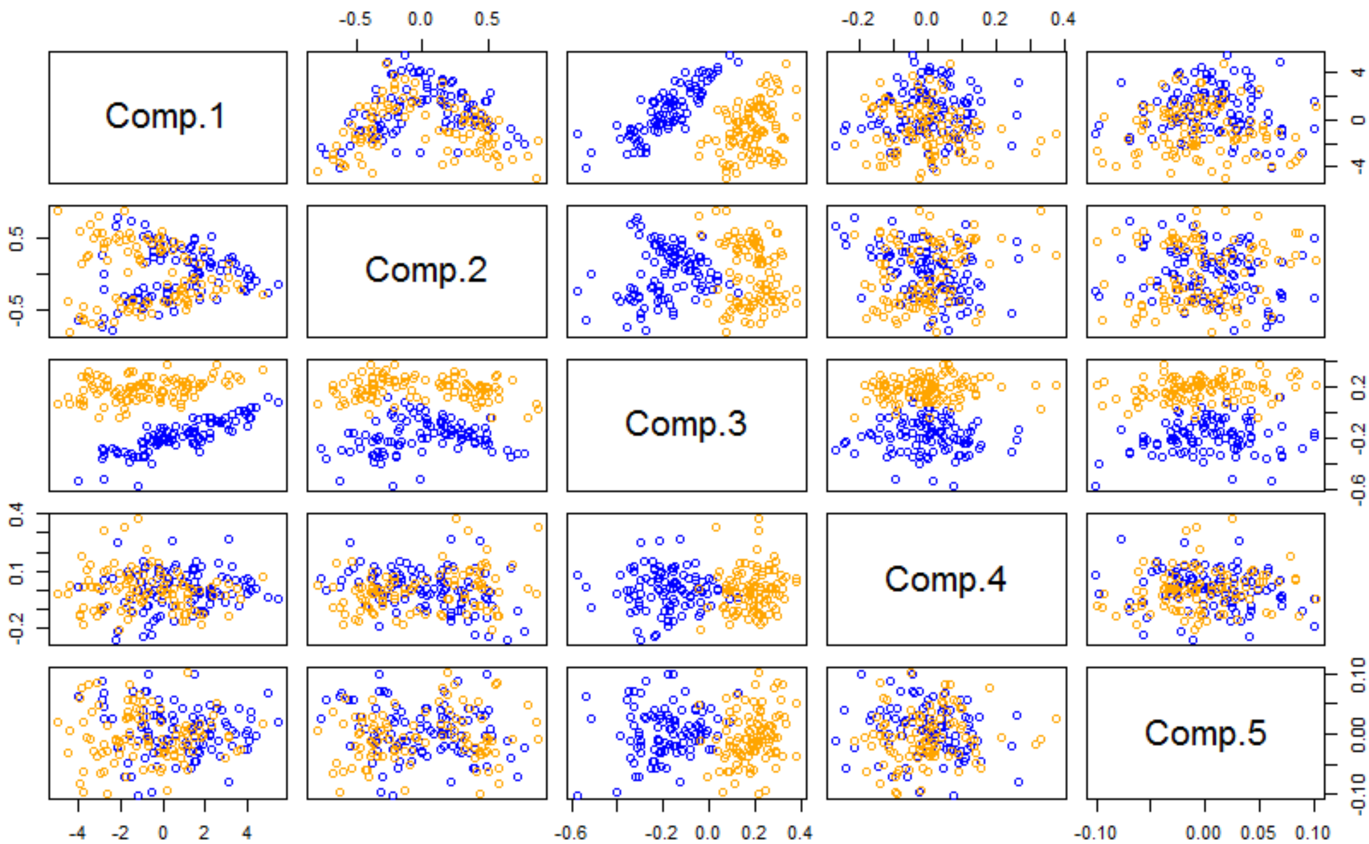
```
> crabs <- scan("crabs(1).txt",skip=1,what=list("c","c",0,0,0,0,0,0))
Read 200 records
> colour1 <- ifelse(crabs[[1]]=="B","blue","orange") # species colours
> colour2 <- ifelse(crabs[[2]]=="M","black","red") # sex colours
> sex <- crabs[[2]]
> FL <- crabs[[4]]
> RW <- crabs[[5]]
> CL <- crabs[[6]]
> CW <- crabs[[7]]
> BD <- crabs[[8]]
> r <- princomp(~FL+RW+CL+CW+BD,cor=T)
> summary(r,loadings=T)
Importance of components:
                         Comp.1     Comp.2      Comp.3      Comp.4       Comp.5
Standard deviation     2.188341 0.38946785 0.215946693 0.105524202 0.0413724263
Proportion of Variance 0.957767 0.03033704 0.009326595 0.002227071 0.0003423355
Cumulative Proportion  0.957767 0.98810400 0.997430593 0.999657664 1.0000000000

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
FL -0.452 -0.138  0.531  0.697
RW -0.428  0.898
CL -0.453 -0.268 -0.310        -0.792
CW -0.451 -0.181 -0.653         0.575
```

```
BD -0.451 -0.264  0.443 -0.707  0.176
```

- The first principal component is a measure of overall size of the crabs.
  - It covers more than 95% of the total variance.
- The second principal component increases only when the variable `RW` increases. So we are focusing on the width of posterior region which varies a lot between male and female crabs. It measures the comparison of `RW` and `(FL, CL, CW, BD)`.
  - It is only responsible for 3% of total variation. This means that the crabs are uniform in shape, but varying on size with some relationship between dimensions within each colour and sex.

(b) Solution:

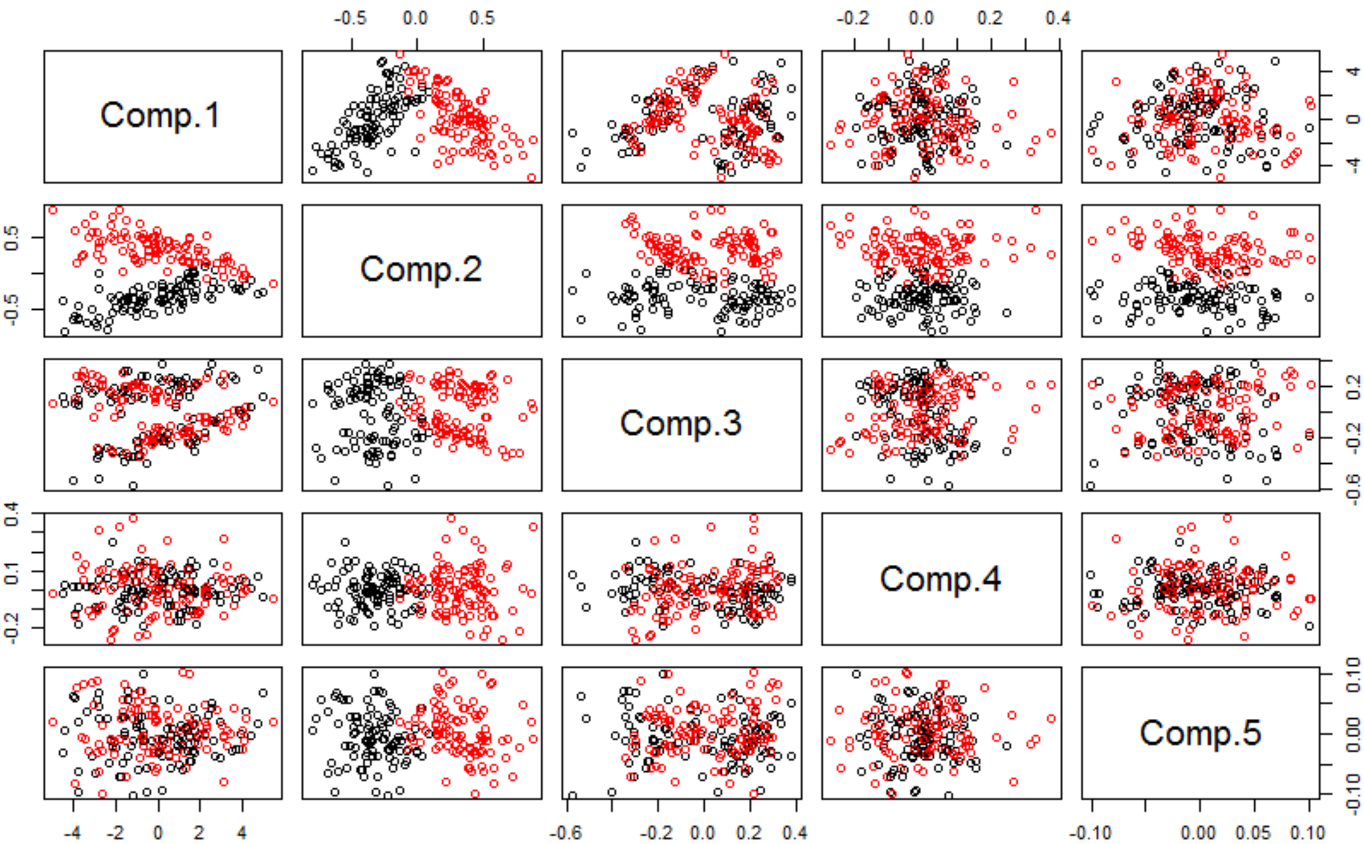
```
> pairs(r$scores,col=colour1)
```



- Component 3 vs Component 1 clearly separetes two species since there is nearly no overlap, and there is even a clear branching.
- Other pairs between Component 3 vs Component 2/4/5 are good enough as well, only a few overlaps.

(c) Solution:

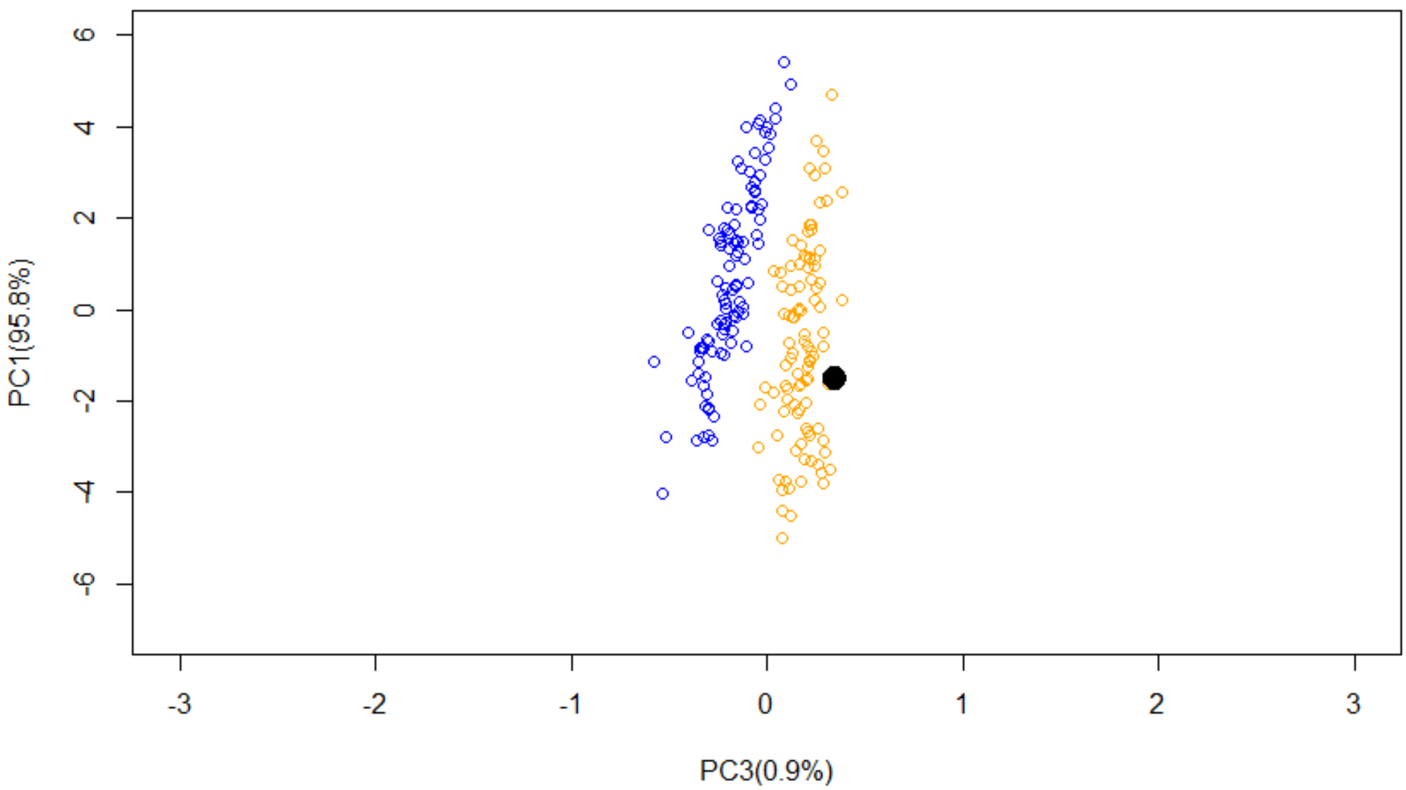
```
> pairs(r$scores,col=colour2)
```



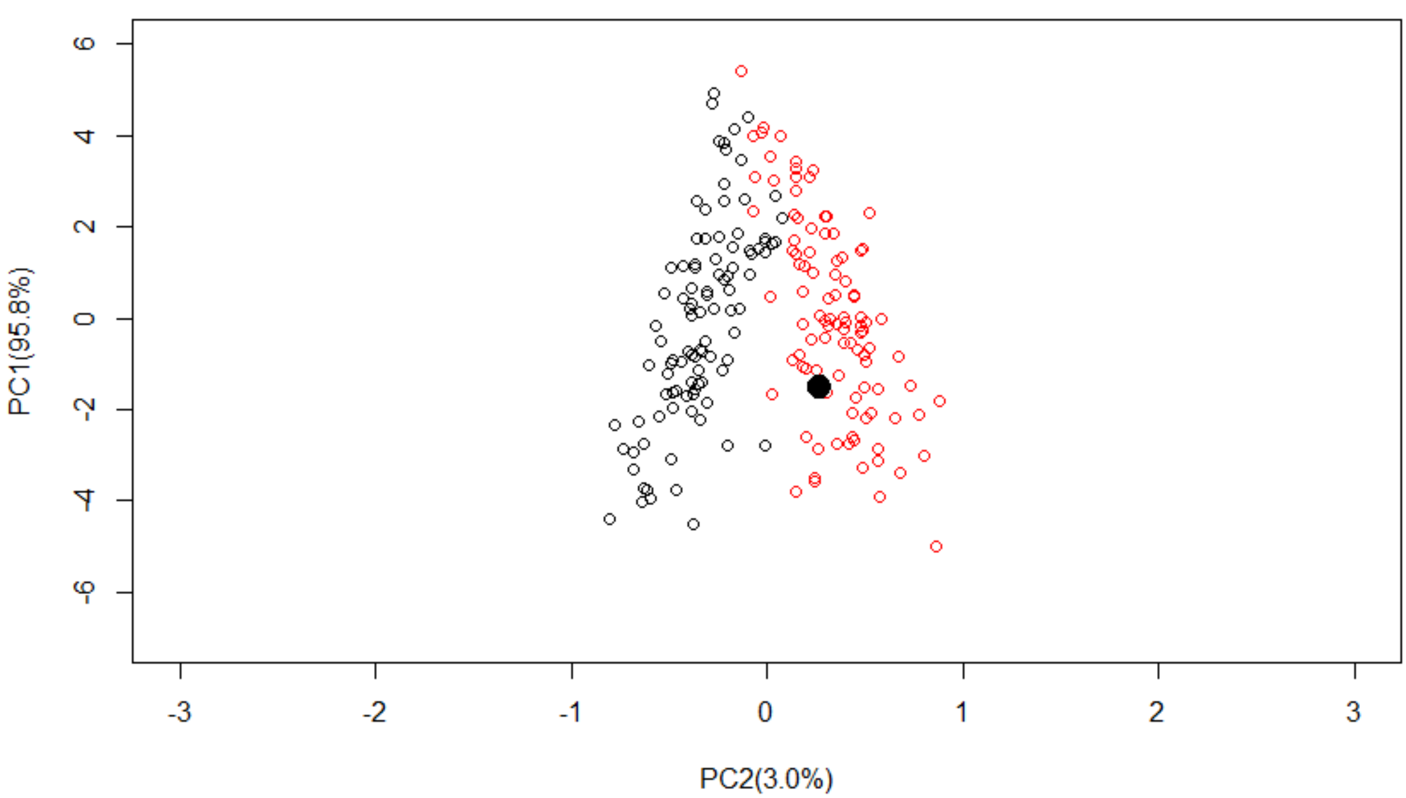
- Component 2 vs Component 1 separates two sexes with a clear branching.
- Other pairs Component 2 vs Component 3/4/5 are good as well, but with some overlaps and no branching.

(d) Solution:

```
mrcrab <- c(18.7,15.0,35.0,40.3,16.6)

p <- function(i){
  co <- r$loadings[,i]
  mv <- colMeans(cbind(FL,RW,CL,CW,BD)) # mean vector of 5 variables
  disp = apply(cbind(FL,RW,CL,CW,BD),2,sd) # dispersion
  return(sum(co*(mrcrab-mv)/disp))
}

plot(r$scores[,3],r$scores[,1],col=colour1,xlim=c(-3,3),ylim=c(-7,6),xlab="PC3(0.9%)",ylab=
"PC1(95.8%)")
points(p(3),p(1),pch=20,cex=3)
plot(r$scores[,2],r$scores[,1],col=colour2,xlim=c(-3,3),ylim=c(-7,6),xlab="PC2(3.0%)",ylab=
"PC1(95.8%)")
points(p(2),p(1),pch=20,cex=3)
```



PC1(95.8%)
6
4
2
0
-2
-4
-6
-3
-2
-1
0
1
2
3
PC3(0.9%)



PC1(95.8%)
6
4
2
0
-2
-4
-6
-3
-2
-1
0
1
2
3
PC2(3.0%)

Therefore, we predict that the target crab is a female "Orange" crab.

Problem 4

(a) Solution:
We know that $X_1, X_2 \sim N(0,1)$, so

$$E(X_1) = E(X_2) = 0, Var(X_1) = Var(X_2) = 1.$$

For expectation:

$$E(a_1X_1 + a_2X_2) = E(a_1X_1) + E(a_2X_2) = a_1E(X_1) + a_2E(X_2) = 0 + 0 = 0.$$

For variance, we know that:

$$Var(X) = E(X^2) - [E(X)]^2,$$
$$Cov(X, Y) = E(XY) - E(X)E(Y).$$

And $Cov(X_1, X_2) = 0$ because of independence.
Also know that $a_1^2 + a_2^2 = 1$.

$$\begin{aligned}
Var(a_1X_1 + a_2X_2) &= E[(a_1X_1 + a_2X_2)^2] - [E(a_1X_1 + a_2X_2)]^2 \\
&= E(a_1^2X_1^2 + a_2^2X_2^2 + 2a_1a_2X_1X_2) + 0 \\
&= a_1^2E(X_1^2) + a_2^2E(X_2^2) + 2a_1a_2E(X_1X_2) \\
&= a_1^2(Var(X_1) + [E(X_1)]^2) + a_2^2(Var(X_2) + [E(X_2)]^2 + 2a_1a_2[Cov(X_1, X_2) + E(X_1)E(X_2)] \\
&= a_1^2 \cdot 1 + a_2^2 \cdot 1 + 2a_1a_2[0 + 0] \\
&= 1
\end{aligned}$$

(b) Solution:
Note that $E(X_1^2) = E(X_2^2) = 1$ is shown in the process of part (a).

$$\begin{aligned}
E[(a_1X_1 + a_2X_2)^4] &= E(a_1^4X_1^4 + a_2^4X_2^4 + 6a_1^2a_2^2X_1^2X_2^2 + 4a_1^3X_1^3a_2X_2 + 4a_1X_1a_2^3X_2^3) \\
&= a_1^4E(X_1^4) + a_2^4E(X_2^4) + 6a_1^2a_2^2E(X_1^2X_2^2) + 4a_1^3a_2E(X_1^3X_2) + 4a_1a_2^3E(X_1X_2^3) \\
&= a_1^4E(X_1^4) + a_2^4E(X_2^4) + 6a_1^2a_2^2[Cov(X_1^2, X_2^2) + E(X_1^2)E(X_2^2)] \\
&+ 4a_1^3a_2[Cov(X_1^3, X_2) + E(X_1^3)E(X_2)] + 4a_1a_2^3[Cov(X_1, X_2^3) + E(X_1)E(X_2^3)] \\
&= a_1^4E(X_1^4) + a_2^4E(X_2^4) + 6a_1^2a_2^2[0 + 1 \cdot 1] + 4a_1^3a_2[0 + 0] + 4a_1a_2^3[0 + 0] \\
&= a_1^4E(X_1^4) + 6a_1^2a_2^2 + a_2^4E(X_2^4)
\end{aligned}$$