Solutions to Assignment #2 STA437H1S/2005H1S

1. **Note:** This question was a bit of a "fishing expedition" with no really right or wrong answers.
(a) The second Andrews plot indicates in red all the curves whose values exceed 7 in absolute value – there are 11 of these in total. (Their indices are $6, 8, 17, 21, 23, 39, 65, 69, 73, 76, 96$.) However, there is really only one clear outlier (observation 39) obvious from this plot. The R code (and output) used here is:

```
> colour <- rep("black",100)
> r <- andrews(test)
> points <- NULL
> for (i in 1:100) {
   if (max(abs(r[,i]))>7) points <- c(points,i)
 }
> points
 [1]  6  8 17 21 23 39 65 69 73 76 96
> colour[points] <- "red"
> r <- andrews(test,colour=colour)
```

(b) The "true" answer is 7 — however, it is virtually impossible to identify the number of outliers. However, one possible approach is to use the 11 observations identified in part (a) and see where they fall in the pairwise scatterplots for the original variables and the PC scores (again these observations are marked in red). The PC plots are particularly useful as these points tend to lie on the edges of the point cloud, particularly for the first few PCs.

2. (a) Write $g_i(t) = \sum_{k=1}^{p} x_{ik}\phi_k(t)$ where $\phi_1(t) = 1/\sqrt{2}$ and $\phi_2, \cdots, \phi_p$ are the sines and cosines. Note that

$$\int_0^1 \phi_k^2(t)\,dt = \frac{1}{2} \quad \text{and} \quad \int_0^1 \phi_k(t) \int_0^1 \phi_\ell(t)\,dt = 0$$

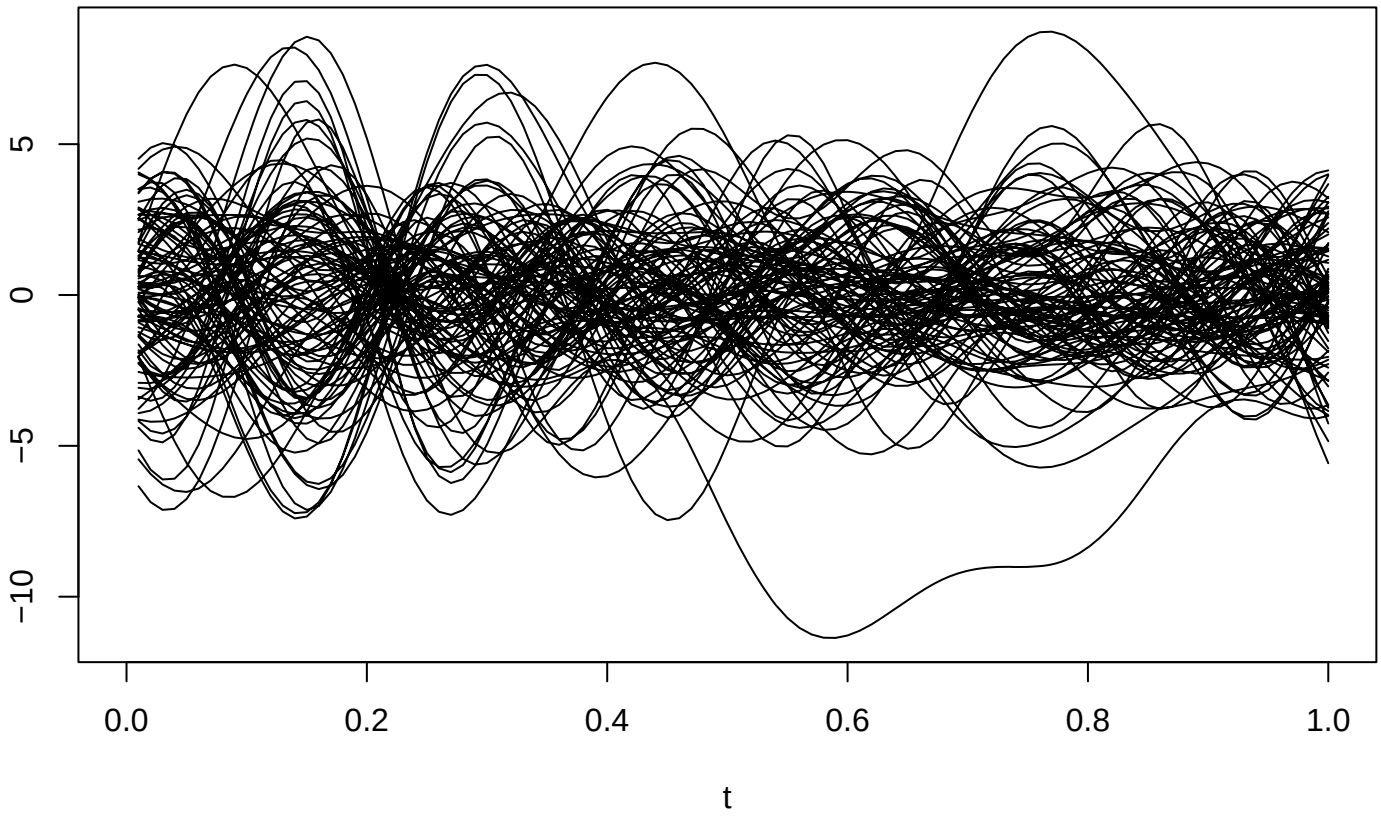
for each k ℓ. Thus

$$\begin{aligned}
\int_0^1 [g_i(t) - g_j(t)]^2\,dt &= \int_0^1 \left\{ \sum_{k=1}^{p}\sum_{\ell=1}^{p} (x_{ik} - x_{jk})(x_{i\ell} - x_{j\ell})\phi_k(t)\phi_\ell(t) \right\} dt \\
&= \sum_{k=1}^{p}\sum_{\ell=1}^{p} (x_{ik} - x_{jk})(x_{i\ell} - x_{j\ell}) \left\{ \int_0^1 \phi_k(t)\phi_\ell(t)\,dt \right\} \\
&= \frac{1}{2}\sum_{k=1}^{p} (x_{ik} - x_{jk})^2
\end{aligned}$$

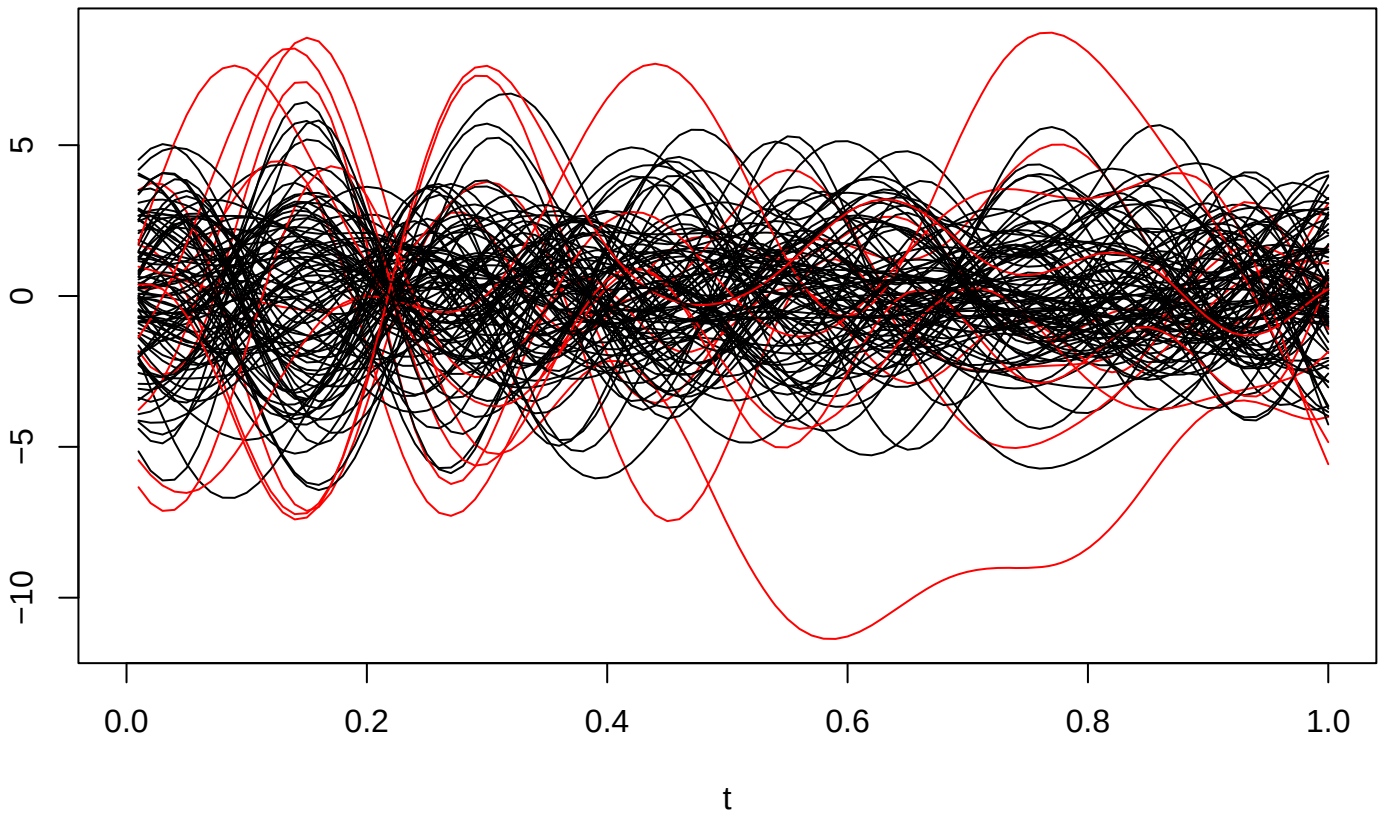
(b) $\bar{\boldsymbol{x}} = (\bar{x}_1, \cdots, \bar{x}_p)^T$ where

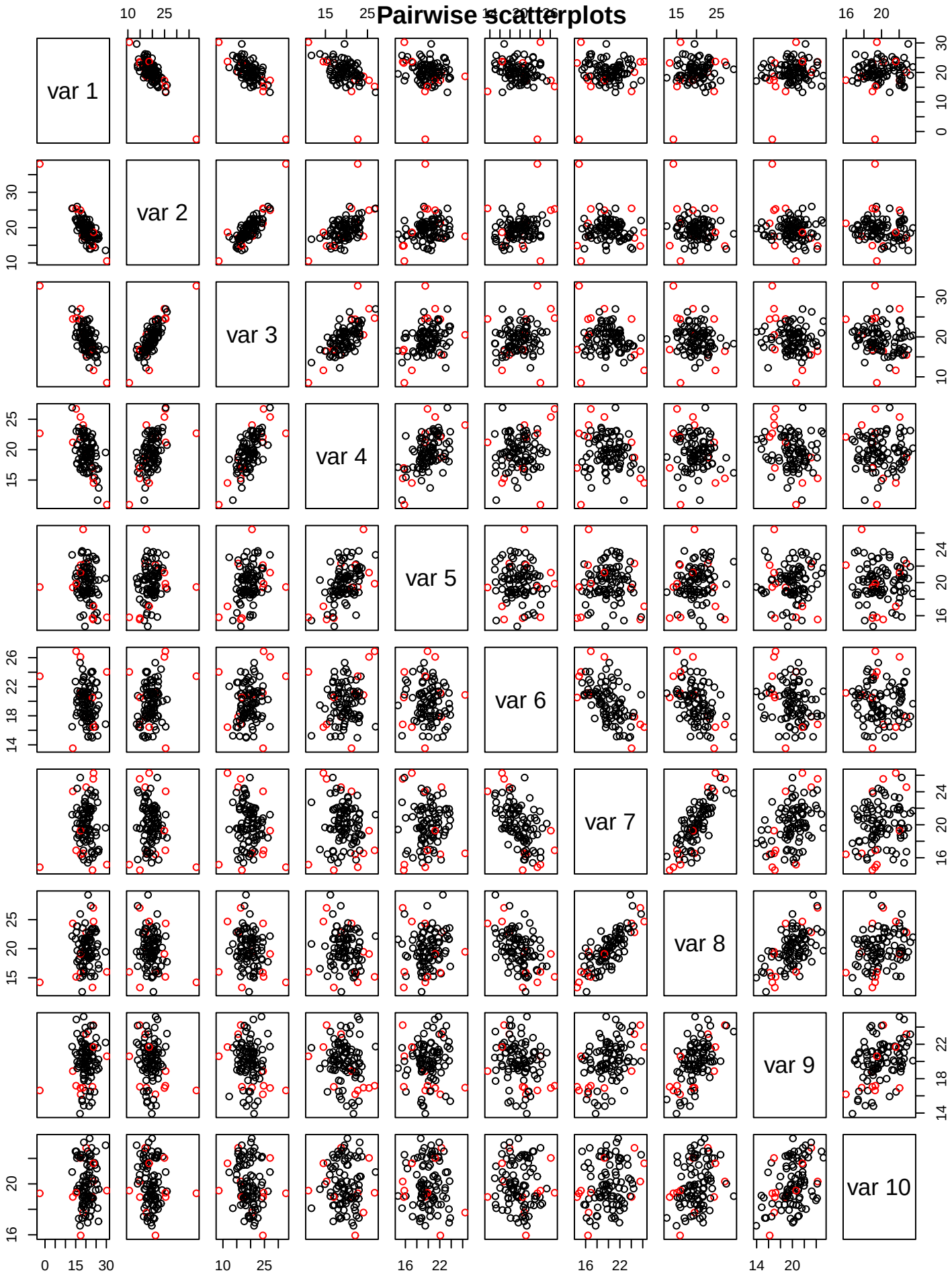
$$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}.$$

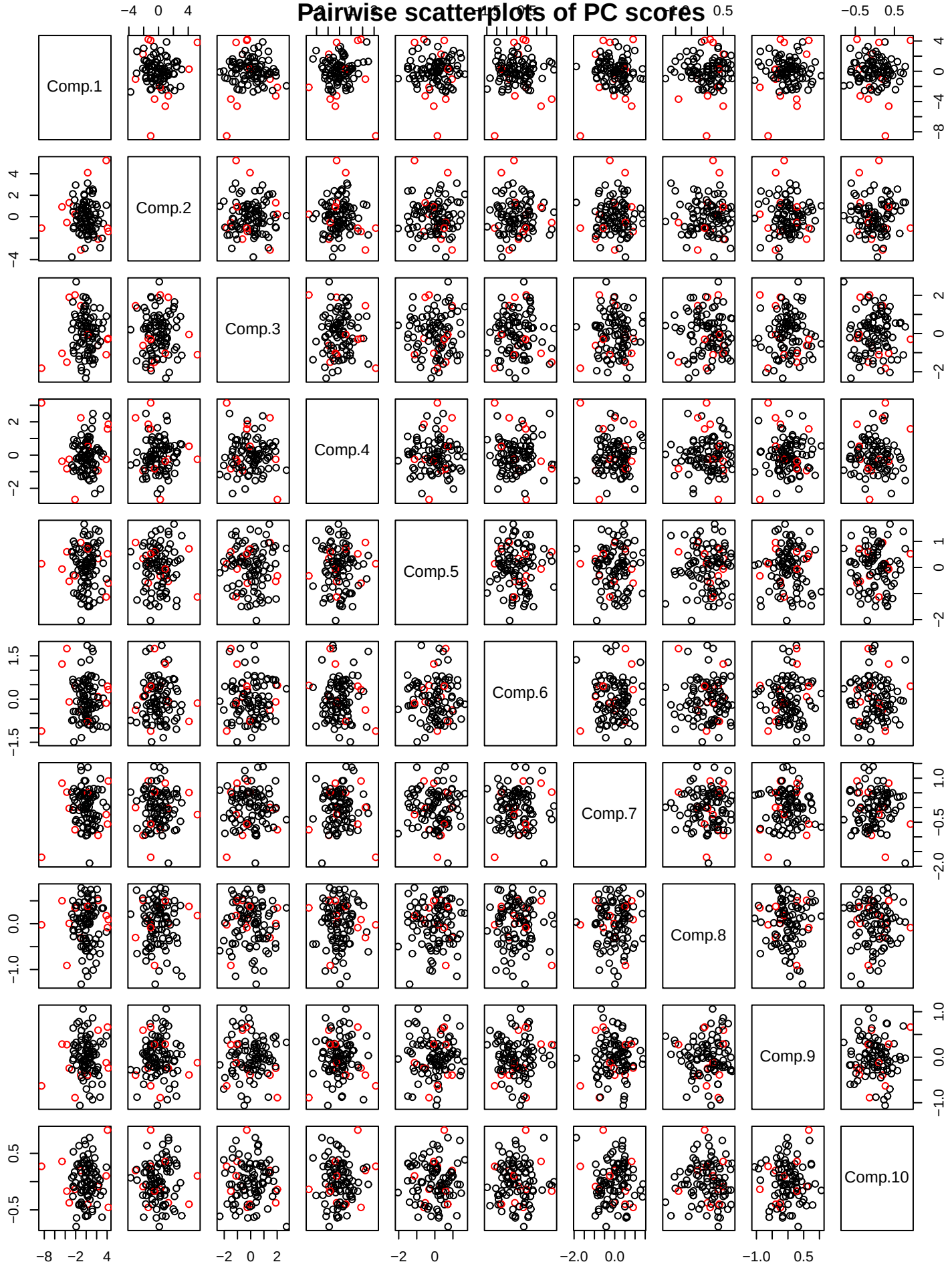
Andrews plot



Andrews plot with suspicious curves marked in red




Pairwise scatterplots
var 1
var 2
var 3
var 4
var 5
var 6
var 7
var 8
var 9
var 10


Pairwise scatterplots of PC scores
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6
Comp.7
Comp.8
Comp.9
Comp.10

Thus

$$\begin{aligned} g_{\bar{x}}(t) &= \frac{\bar{x}_1}{\sqrt{2}} + \bar{x}_2 \sin(2\pi t) + \bar{x}_3 \cos(2\pi t) + \bar{x}_4 \sin(4\pi t) + \bar{x}_5 \cos(4\pi t) + \cdots \\ &= \frac{1}{n} g_i(t) \end{aligned}$$

That is, the Andrews curve of the sample mean is simply the average of the Andrews curves.
(c) Using a similar argument to that in part (b), we have

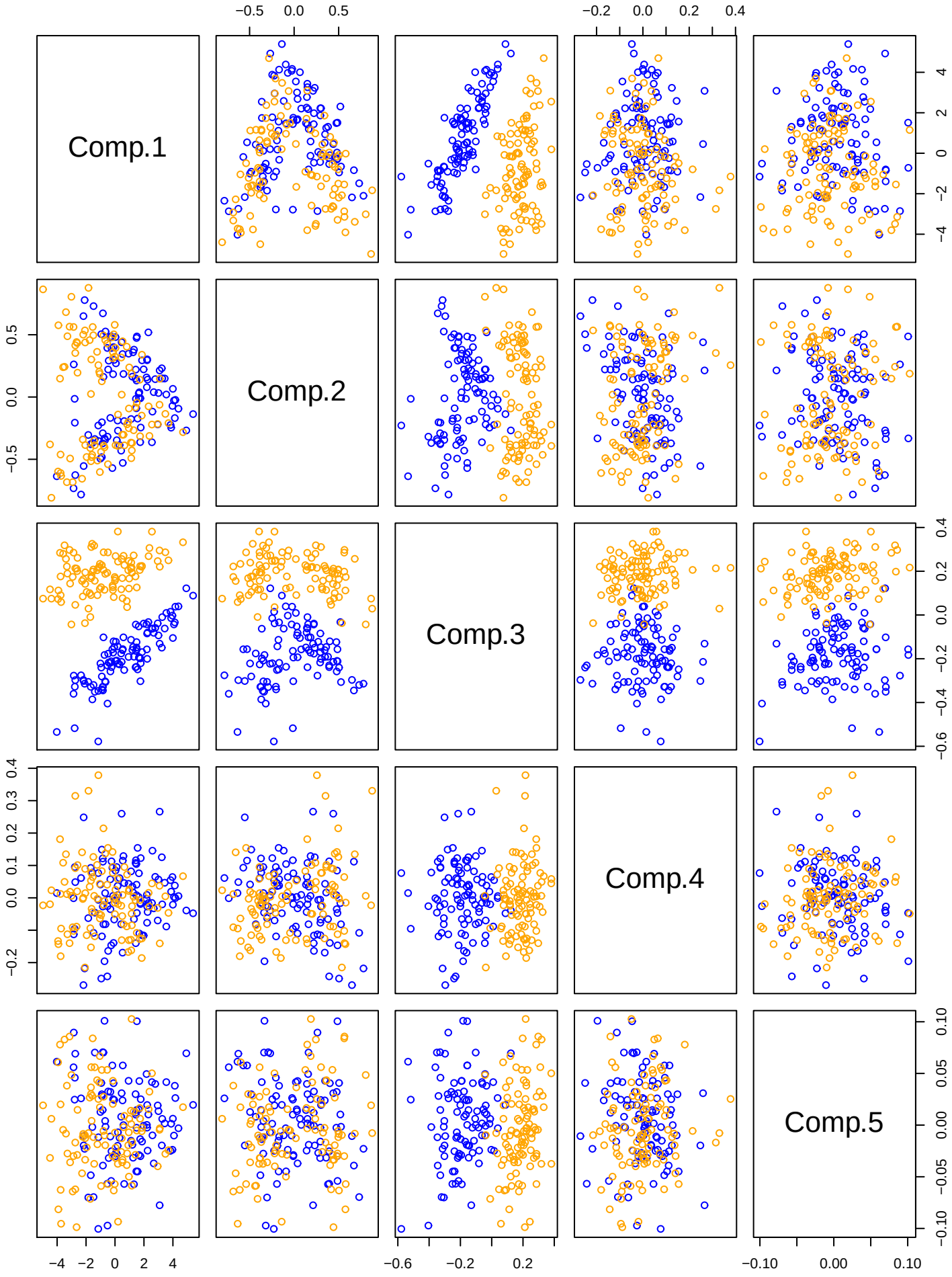
$$g_k(t) = \lambda g_i(t) + (1 - \lambda) g_j(t).$$

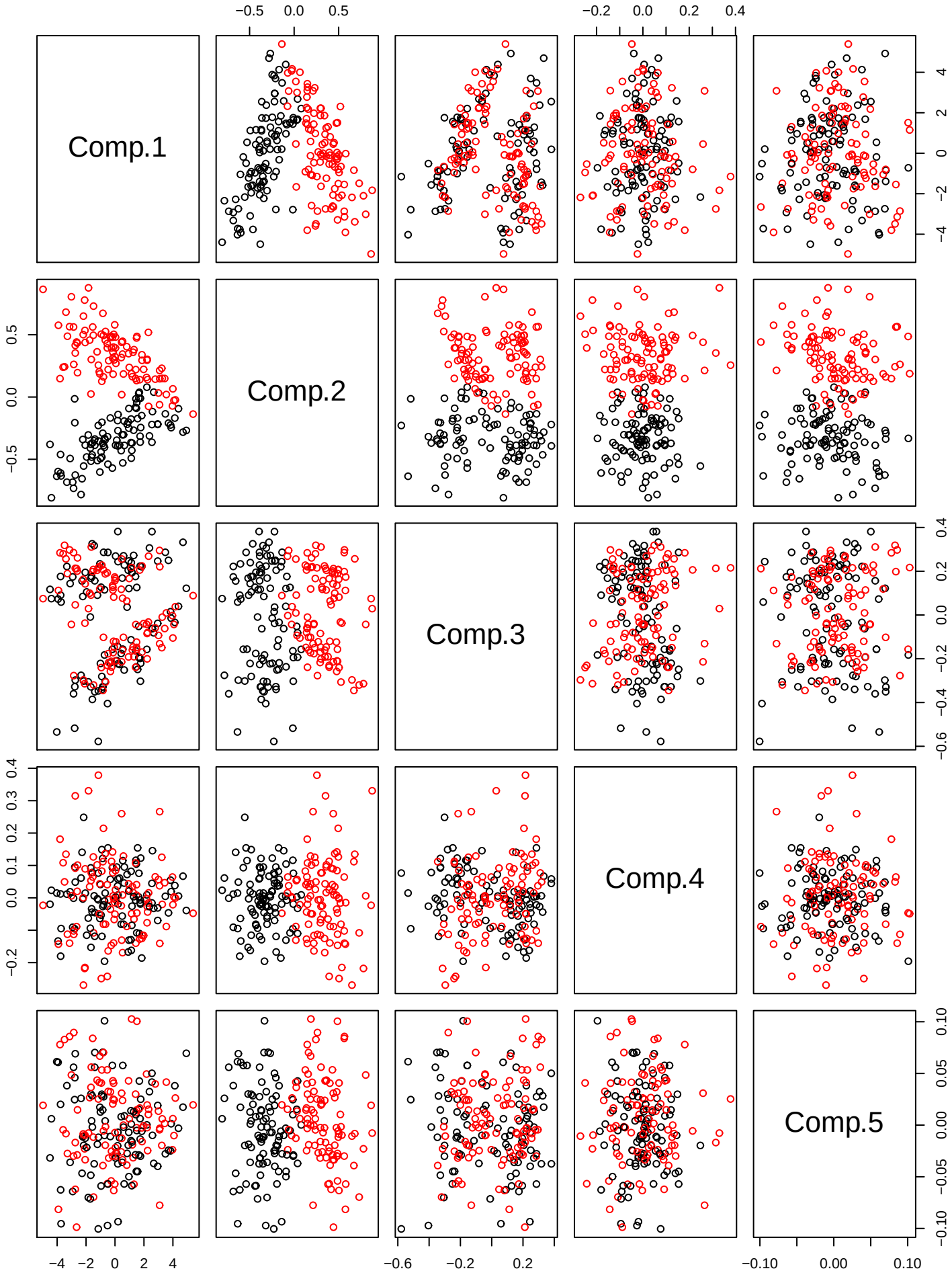
Essentially, $g_k(t)$ is "sandwiched" between $g_i(t)$ and $g_j(t)$ — it can never lie above both nor below both.

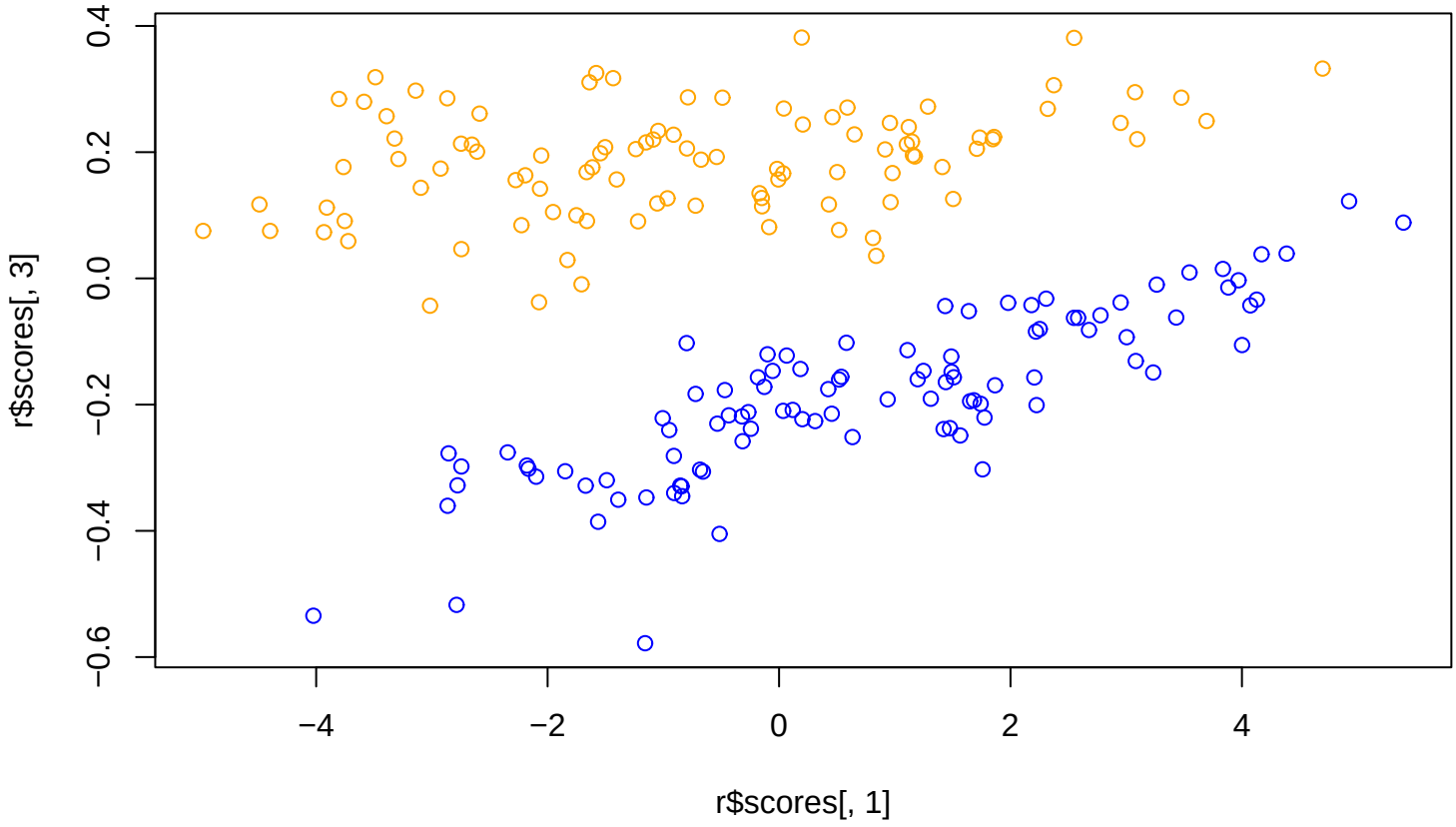
3. (a) The R output is as follows:

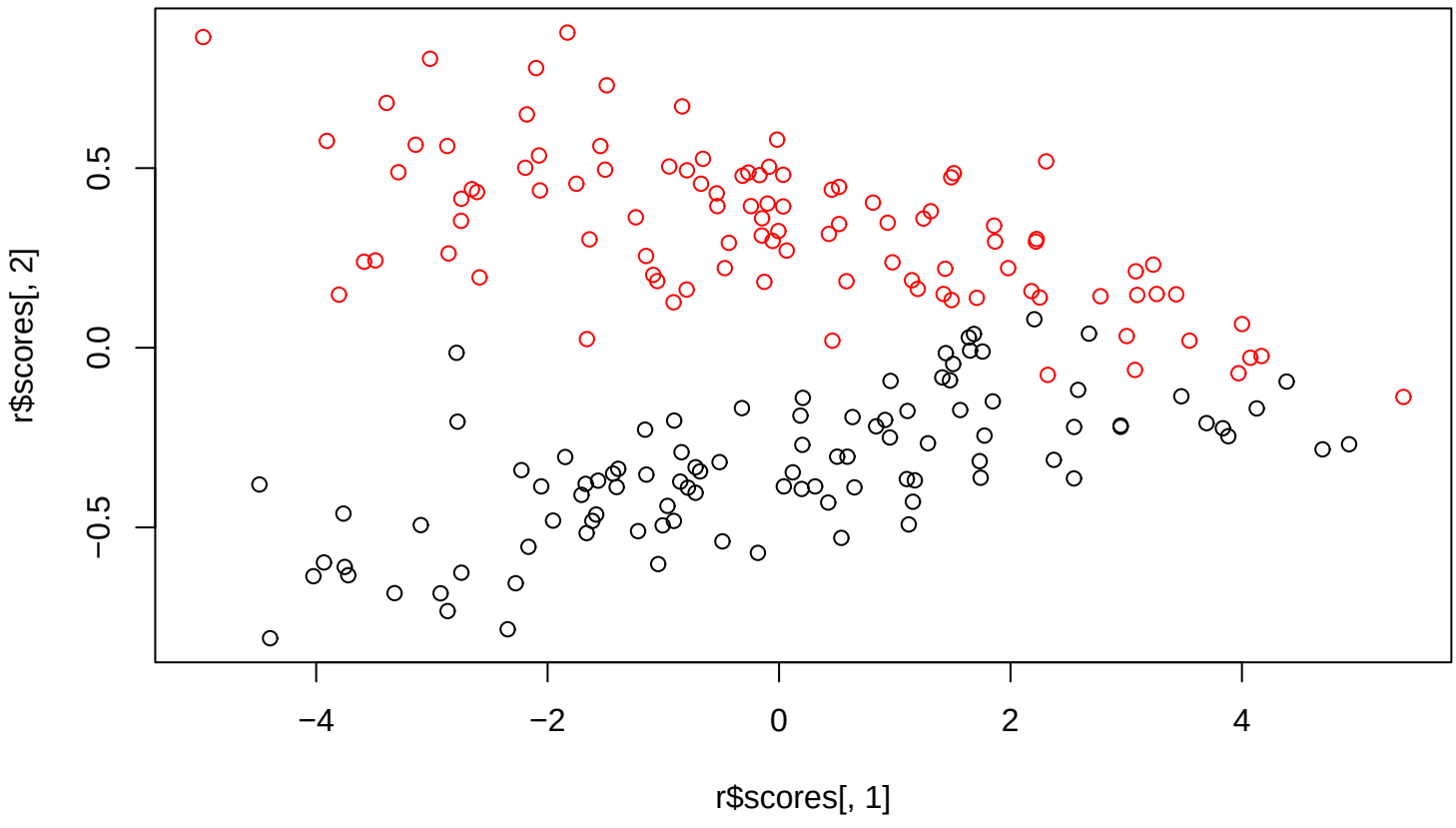
```
> r <- princomp(~FL+RW+CL+CW+BD,cor=T)
> summary(r,loadings=T)
Importance of components:
                          Comp.1     Comp.2      Comp.3      Comp.4       Comp.5
Standard deviation     2.188341 0.38946785 0.215946693 0.105524202 0.0413724263
Proportion of Variance 0.957767 0.03033704 0.009326595 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 loadings for the first principal component are are approximately the same – since the variables are different measures of the size of the crab, we can interpret the first PC as a measure of size of the crab. The second PC is more difficult – it essentially compare the variable RW to the other 4 variables.
(b) See the pairwise scatterplots – it appears that the species can be distinguished by looking at the first and third PCs.
(c) See the pairwise scatterplots – it appears that the sex can be distinguished by looking at the first and second PCs.
(d) We need to first normalize the variables and then compute the scores for the measurements. This can be done using the following R code:


Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
−0.5 0.0 0.5
−0.2 0.0 0.2 0.4
−4 −2 0 2 4
−0.5 0.0 0.5
−0.2 0.0 0.2 0.4
−4 −2 0 2 4
−0.5 0.0 0.5
−0.6 −0.4 −0.2 0.0 0.2 0.4
−0.2 0.0 0.1 0.2 0.3 0.4
−0.10 −0.05 0.00 0.05 0.10
−4 −2 0 2 4
−0.6 −0.2 0.2
−0.10 0.00 0.10


Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
−0.5 0.0 0.5
−0.2 0.0 0.2 0.4
−4 −2 0 2 4
0.5 0.0 −0.5
−4 −2 0 2 4
0.4 0.2 0.0 −0.2 −0.4 −0.6
0.4 0.3 0.2 0.1 0.0 −0.2
−4 −2 0 2 4
−0.6 −0.2 0.2
−0.10 0.00 0.10
0.10 0.05 0.00 −0.05 −0.10


r$scores[, 3]
0.4
0.2
0.0
−0.2
−0.4
−0.6
−4
−2
0
2
4
r$scores[, 1]


r$scores[, 2]
0.5
0.0
−0.5
−4
−2
0
2
4
r$scores[, 1]

```
> point <- c(18.7, 15.0, 35.0, 40.3, 16.6)
> means <- c(mean(FL),mean(RW),mean(CL),mean(CW),mean(BD))
> sds <- c(sd(FL),sd(RW),sd(CL),sd(CW),sd(BD))
> point.norm <- (point-means)/sds
> point.norm
[1] 0.8917625 0.8788190 0.4065890 0.4935877 0.7502689
> scores <- t(r$loadings)%*%point.norm
> as.vector(scores)
[1] -1.52471576  0.27008949  0.34726841  0.06119752  0.13176893
```

Comparing these scores to the plots, we conclude that the measurements come from an orange, female crab.

4. (a) This is straightforward: $E(a_1X_1 + a_2X_2) = a_1E(X_1) + a_2E(X_2) = 0$ and $\text{Var}(a_1X_1 + a_2X_2) = a_1^2\text{Var}(X_1) + a_2^2\text{Var}(X_2) = 1$.
(b) $(a_1X_1 + a_2X_2)^4 = a_1^4X_1^4 + 4a_1a_2^3X_1X_2^3 + 66a_1^2a_2^2X_1^2X_2^2 + 4a_1^3a_2X_1^3X_2 + a_2^4X_2^4$. Taking expected values and noting that $E(X_1X_2^3) = E(X_1^3X_2) = 0$, we have

$$E[(a_1X_1 + a_2X_2)^4] = a_1^4E(X_1^4) + 6a_1^2a_2^2 + a_2^4E(X_2^4).$$

(c) Note that

$$\begin{aligned} a_1^4\left\{E(X_1^4) - 3\right\} + a_2^4\left\{E(X_2^4) - 3\right\} &= a_1^4E(X_1^4) + a_2^4E(X_2^4) - 3(a_1^4 + a_2^4) \\ &= E[(a_1X_1 + a_2X_2)^4] - 6a_1^2a_2^2 - 3(a_1^4 + a_2^4) \\ &= E[(a_1X_1 + a_2X_2)^4] - 3(a_1^2 + a_2^2)^2 \\ &= E[(a_1X_1 + a_2X_2)^4] - 3 \end{aligned}$$

since $a_1^2 + a_2^2 = 1$. Now using the triangle inequality, we have

$$\left|E[(a_1X_1 + a_2X_2)^4] - 3\right| \leq a_1^4\left|E(X_1^4) - 3\right| + a_2^4\left|E(X_2^4) - 3\right|.$$

(d) If a_1 and a_2 are both non-zero then $0 < |a_1|, |a_2| < 1$ and so

$$\left|E[(a_1X_1 + a_2X_2)^4] - 3\right| < \max\left\{\left|E(X_1^4) - 3\right|, \left|E(X_2^4) - 3\right|\right\}$$

unless $E(X_1^4) = E(X_2^4) = 3$ (in which case $E[(a_1X_1 + a_2X_2)^4] = 3$).