STAT505 Assessment #9

1. J&W Exercise 8.13. The data are given in "radiotherapy.dat".

(a) The covariance matrix is

	symptoms	activity	sleep	eat	appetite	skin
symptoms	4.67517	0.91288	0.54421	0.27933	1.03660	0.16016
activity	0.91288	0.62374	0.10128	0.12422	0.38558	-0.01936
sleep	0.54421	0.10128	0.51291	0.07232	0.30676	0.07762
eat	0.27933	0.12422	0.07232	0.11014	0.21674	0.00740
appetite	1.03660	0.38558	0.30676	0.21674	0.84669	-0.03151
skin	0.16016	-0.01936	0.07762	0.00740	-0.03151	0.86489

The correlation matrix is

	symptoms	activity	sleep	eat	appetite	skin
symptoms	1.0000	0.5346	0.3514	0.3893	0.5210	0.0797
activity	0.5346	1.0000	0.1791	0.4739	0.5306	-0.0264
sleep	0.3514	0.1791	1.0000	0.3043	0.4655	0.1165
eat	0.3893	0.4739	0.3043	1.0000	0.7097	0.0240
appetite	0.5210	0.5306	0.4655	0.7097	1.0000	-0.0368
skin	0.0797	-0.0264	0.1165	0.0240	-0.0368	1.0000

(b) Since the variables appear to be on quite different scales, we will work with the correlation matrix. The eigenvectors are listed below from left to right in the order of decreasing eigenvalue.

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
symptoms	0.444503	0.067926	0.238098	-0.608552	-0.594830	0.130208
activity	0.437105	-0.210714	0.489278	-0.183126	0.698670	0.058910
sleep	0.341981	0.360805	-0.733236	-0.267827	0.321219	0.200802
eat	0.469050	-0.090012	0.002962	0.648415	-0.199778	0.558146
appetite	0.522938	-0.089193	-0.156486	0.252166	-0.112102	-0.786090
skin	0.030527	0.897077	0.376544	0.200918	0.048765	-0.098939

The eigenvalues, along with their proportions of variability explained, are

Eigenvalue	Proportion
2.82524587	0.4709
1.06035484	0.1767
0.80465519	0.1341
0.65858239	0.1098
0.40471023	0.0675
0.24645148	1.0000

(c) Since the first principal component explains only 47% of the variability, we would need more than one here. Either three or four would be a better choice since those choices would account for a cumulative 78.17% or 89.15% variability, respectively.

(d) The correlations between the variables and the principal components are given below. It's difficult to give precise interpretations for these, but based on their strongest correlations, it appears essentially that Prin1 is an average of all variables except sleep and skin; Prin2 is skin alone; Prin3 is sleep; and Prin4 is the difference between eat and symptoms.

	Prin1	Prin2	Prin3	Prin4
symptoms	0.74714	0.06995	0.21358	-0.49386
activity	0.73471	-0.21698	0.43889	-0.14861
sleep	0.57482	0.37153	-0.65773	-0.21735
eat	0.78840	-0.09269	0.00266	0.52621
appetite	0.87898	-0.09185	-0.14037	0.20464
skin	0.05131	0.92375	0.33777	0.16305

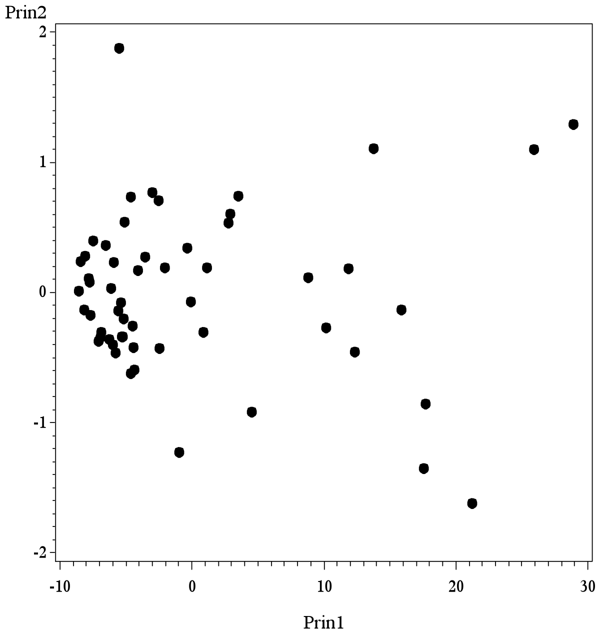
SAS code:

```
data radio;
  infile "v:\radiotherapy.dat";
  input symptoms activity sleep eat appetite skin;
  run;
proc princomp data=radio  out=a1r;
  var symptoms activity sleep eat appetite skin;
  run;
proc corr data=a1r noprob;
  var  prin1 prin2 prin3 prin4 symptoms activity sleep eat appetite skin;
  run;
```

2. The data in “track.dat” give the men’s national track records for 55 countries in 1984 for the following distances: 100m, 200m, 400m, 800m, 1500m, 5000m, 10000m, and marathon. The first three variables are measured in seconds, while the remaining variables are measured in minutes. Express the values of the first three variables in minutes by dividing each number by 60. This can be accomplished using the following data step:

```
data track;
   infile "v:\track.dat";
   input d100 d200 d400 d800 d1500 d5000 d10000 marathon country $;
   d100=d100/60;
   d200=d200/60;
   d400=d400/60;
   run;
```

   (a) Perform a principal component analysis using the covariance matrix; that is, using the raw data expressed in minutes. Include a scatter plot of the first two principal components.



i. How many principal components are required to explain 90% of the total variation for this data?

   Using the covariance matrix, only one principal component is required to explain 90% of the variation of times; it explains over 99% of the variation.

ii. For the number of components in part i, give the formula for each component and a brief interpretation.

   Note that the coefficients SAS provides assumes the mean is subtracted from each variable first. So, for example, `d_100` below actually represents an individual's 100-meter time minus the mean 100-meter time.

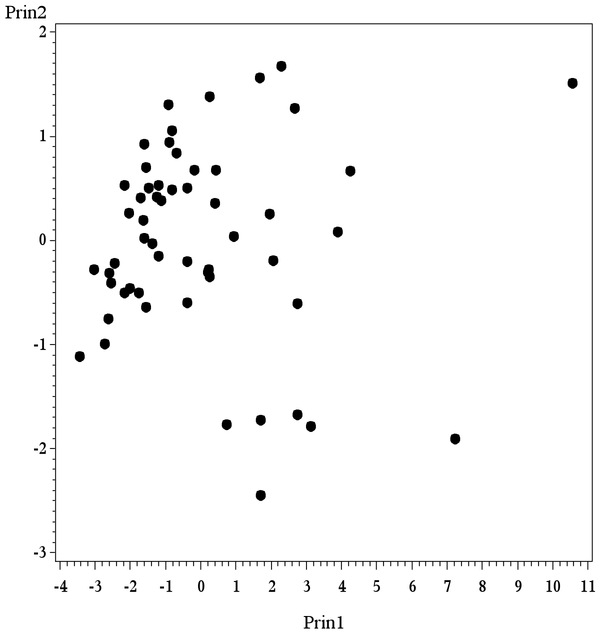
$$\begin{aligned}\widehat{Y}_1 = {}&0.000327d_{100} + 0.000687d_{200} + 0.00183d_{400}\\&+ 0.00549d_{800} + 0.0144d_{1500} + 0.0797d_{5000}\\&+ 0.182d_{10000} + 0.980d_{\text{marathon}}.\end{aligned}$$

   The first component is essentially represented by marathon.

iii. What countries have the highest and lowest values for each principal component (only include the number of components specified in part i). For each of those countries, give the principal component scores.

   After inspecting the output data set, the USA has the lowest score for the first principal component, and the country of Cookis has the highest value. SAS computes these from the formula above but subtracts the mean from each variable beforehand. It reports the values of -8.5735 and 28.9126 for the USA and Cookis, respectively.

(b) Perform a principal component analysis using the correlation matrix. Include a scatter plot of the first two principal components.



i. How many principal components are required to explain 90% of the total variation for this data?

Using the correlation matrix, two principal components are required to explain 90% of the variation. Together, they account for 93.75% of the variation.

ii. For the number of components in part i, give the formula for each component and a brief interpretation.

The formulas for these two components are below. If the events are grouped as "short" distance and "long" distance, then the first principal component is essentially the sum of both groups, and the second principal component is essentially the difference between the two (short minus long).

$$\begin{aligned}\widehat{Y}_1 =& 0.318d_{100} + 0.337d_{200} + 0.356d_{400}\\ &+ 0.369d_{800} + 0.372d_{1500} + 0.364d_{5000}\\ &+ 0.367d_{10000} + 0.342\text{marathon}\\ \widehat{Y}_2 =& 0.567d_{100} + 0.462d_{200} + 0.248d_{400}\\ &+ 0.0124d_{800} - 0.138d_{1500} - 0.312d_{5000}\\ &- 0.307d_{10000} - 0.439\text{marathon}.\end{aligned}$$

iii. What countries have the highest and lowest values for each principal component (only include the number of components specified in part i). For each of those countries, give the principal component scores.

The extreme principal component values again belong to the USA and Cookis. For the USE, they are −3.4305 and −1.11, respectively, and for Cookis, they are 10.5556 and 1.509. To find these values, SAS first subtracts the mean from each variable *and* divides by the standard deviation before applying the formulas above.

(c) Compare the results from parts (a) and (b). Which gives the best interpretation of the data?

Not surprisingly when using the covariance matrix, the marathon variable alone almost entirely dominates the variation, which is not desirable. When using the correlation matrix, we find two principal components that are essentially uncorrelated versions of measures of short distance and long distance. This provides a more satisfying reduction of the data than that provided by marathon alone.

SAS code:

```
proc princomp data=track cov out=a2;
  var d100 d200 d400 d800 d1500 d5000 d10000 marathon;
  run;
proc gplot data=a2;
  axis1 length=4 in; axis2 length=4 in;
  plot prin2*prin1 / vaxis=axis1 haxis=axis2;
  symbol v=J f=special h=2 i=none color=black;
  run;
proc sort data=a2;
  by prin1;
  run;
proc print;
  run;
proc princomp out=b2;
  var d100 d200 d400 d800 d1500 d5000 d10000 marathon;
  run;
proc gplot data=b2;
  axis1 length=4 in; axis2 length=4 in;
  plot prin2*prin1 / vaxis=axis1 haxis=axis2;
  symbol v=J f=special h=2 i=none color=black;
  run;
proc sort data=b2;
  by prin1;
  run;
proc print;
  run;
```