Stats 519: HW 4 - PCA (CH 4)

Due on March 2, 2009

Dr. Stephen Lee 1:30

Monte J. Shaffer

Problem 1

The RECORDS dataset may be found on the website. This includes athletic records for 55 countries in each of 8 track and field events. Read this data into R and use the prcomp command to answer question 4.8. The file RECORS contains data on the athletic recors for each of 55 countries in races (seconds and minutes). Analyze the data using principal components and answer the following questions:

```
> trace = function(square) {sum(diag(square));}
> setwd("C:/latex/statsMultiVariate/datasets");
> myData = read.csv('RECORDS.csv',header=FALSE);
>    colnames(myData)=c("Country","100m","200m","400m","800m","1500m","5000m","10000m","marathon");
> X  = (myData[,-1]);
> Xs = scale(myData[,-1]);
> rownames(X) = rownames(Xs) =myData[,1];
> n = length(myData[,2]);
> S = t(as.matrix(Xs))%*%as.matrix(Xs)/(n-1);  # this is R if X=Xs
> S.e = eigen(S);
     > Lambda  = S.e$values;
     > U       = S.e$vectors;
     > P = prcomp(Xs, cor=TRUE, retx=TRUE);
     > summary(P, loadings=TRUE);
          Importance of components:
                                  PC1  PC2    PC3    PC4    PC5     PC6     PC7     PC8
          Standard deviation     2.45 1.02 0.7426 0.3713 0.3297 0.27742 0.23901 0.15675
          Proportion of Variance 0.75 0.13 0.0689 0.0172 0.0136 0.00962 0.00714 0.00307
          Cumulative Proportion  0.75 0.88 0.9494 0.9666 0.9802 0.98979 0.99693 1.00000
     > zapsmall(cov(P$x));
                    PC1      PC2      PC3      PC4      PC5      PC6      PC7      PC8
          PC1 6.004257 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
          PC2 0.000000 1.039034 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
          PC3 0.000000 0.000000 0.551478 0.000000 0.000000 0.000000 0.000000 0.000000
          PC4 0.000000 0.000000 0.000000 0.137849 0.000000 0.000000 0.000000 0.000000
          PC5 0.000000 0.000000 0.000000 0.000000 0.108723 0.000000 0.000000 0.000000
          PC6 0.000000 0.000000 0.000000 0.000000 0.000000 0.076962 0.000000 0.000000
          PC7 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.057124 0.000000
          PC8 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.024572

> ## Bartlett Test of Sphericity
> p  = 8; lnS  = log(det(S)); test = -1*((n-1)-(2*p+5)/6)*lnS;   crit = qchisq(.95, df=(p^2-p)/2);
> test > crit;
> print("If TRUE, Reject Null of Sphericity => continue with dimension reduction");

     > round(Lambda,digits=3);
     > VAF=100*round(Lambda/trace(S),digits=3);
     > Z  = as.matrix(Xs)%*%as.matrix(U);
     > F=round(cor(Xs,Z),digits=3);     # loadings
     > A=round(rbind(F,Lambda,VAF),digits=2);
     > rownames(A)=c("100m","200m","400m","800m","1500m","5000m","10000m","marathon","EIGEN","%VAF");
```

```
             [,1]  [,2] [,3]  [,4]  [,5]  [,6] [,7]  [,8]
100m        -0.79 -0.39 -0.41 -0.18  0.14  0.01 -0.02 -0.02
200m        -0.36 -0.85  0.38  0.03 -0.03 -0.02  0.00  0.01
400m        -0.90 -0.09 -0.32  0.02 -0.26 -0.01  0.04  0.02
800m        -0.95  0.03 -0.12  0.24  0.10 -0.14 -0.06  0.02
1500m       -0.96  0.07  0.05  0.13  0.07  0.14  0.14 -0.04
5000m       -0.95  0.18  0.17 -0.02 -0.06  0.09 -0.16 -0.07
10000m      -0.95  0.19  0.16 -0.09  0.03  0.07 -0.02  0.12
marathon    -0.90  0.28  0.26 -0.15  0.00 -0.16  0.08 -0.04
-----------------------------------------------------------
EIGEN        6.00  1.04  0.55  0.14  0.11  0.08  0.06  0.02
%VAF        75.10 13.00  6.90  1.70  1.40  1.00  0.70  0.30
> ## nFactors
> library(nFactors);
    > nResults = nScree(eig = S.e$values,aparallel = parallel(subject = n, var = p)$eigen$qevpea);
> plotuScree(S.e$values);
> plotnScree(nResults, main="Component Retention Analysis");
```

a) How much variation is accounted for by the first two principal components? **88% of the variance is accounted for by the first two principal components.**

b) How many principal components are necessary to account for 80 percent of the variation in the original data? **Two components account for at least 80% of the variation of the scaled data.**

c) How many components would you extract? How would you interpret the components? **I would only extract one, a common "G" factor related to the events [Please see Advanced Factor Retention Analysis on the Scree Plot]. Using the simple Kaiser rule, the variance of the second component accounts for slightly more than one accounting for only 13% of the variance. The shorter events are lower correlations on the first factor. Too many cross-loadings to find meaning beyond the one factor. If choosing one, the general factor could represent overall speed for records: countries like the United States are at one end of the spectrum (generally faster countries), and countries like Cook Island are at the other end of the spectrum (generally slower countries).**

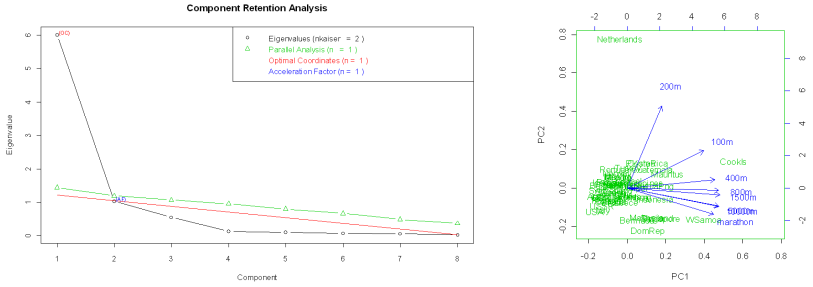

Figure 1: Scree Analysis and Biplot of PCA

Problem 2

Examine the first two principal components of the records data with a scatterplot. Do you observe any unusual structure?

```
> ## using princomp
> X.princomp = princomp(Xs); # just in case it doesn't scale
> ## princomp$loadings
>    rownames(X.princomp$loadings)=c("100m","200m","400m","800m","1500m","5000m","10000m","marathon");
> X.princomp$loadings;   ## not the same as F !!!
>    rownames(X.princomp$scores)=myData[,1];
> X.princomp$scores      ## same as Z except for sign differences
> summary(X.princomp);
    > plot(X.princomp$scores[,1],X.princomp$scores[,2]);
    > text(X.princomp$scores[,1],X.princomp$scores[,2],cex=0.7,lwd=2, labels=myData[,1]);
> library(rgl);
> plot3d(X.princomp$scores[,1:3], type="s", radius=.1, col=rainbow(300)[rank(Z[,1])]);
> text3d(X.princomp$scores[,1:3],text=myData[,1]);
```

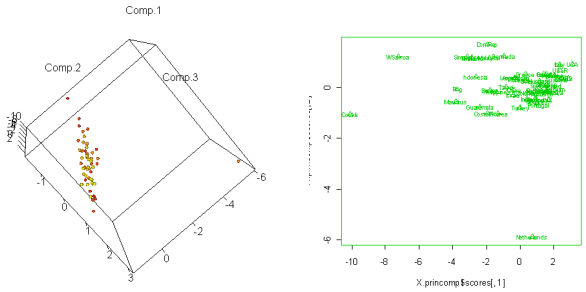

Figure 2: RGL and scatterplot

It appears that the Netherlands, Cook Islands, Luxemburg, and Western Samoa are clear outliers, because their scores are so much slower than the "elite" countries of the world. Netherlands is a bit different than some of the others because although it doesn't perform well in short distances, it is elite in runs over 200m.

```
> X[c(12,34,38,53,54,55),]
             100m  200m  400m 800m 1500m 5000m 10000m marathon
CookIs      12.18 23.20 52.94 2.02  4.24 16.70  35.38   164.70
Luxemburg   10.35 20.77 47.40 1.82  3.67 13.64  29.08   141.27
Netherlands 10.52 29.95 45.10 1.74  3.62 13.36  27.61   129.02
USA          9.93 19.75 43.86 1.73  3.53 13.20  27.43   128.22
USSR        10.07 20.00 44.60 1.75  3.59 13.20  27.53   130.55
WSamoa      10.82 21.86 49.00 2.02  4.24 16.28  34.71   161.83
------------------------------------------------------------------
MEAN        10.47 21.10 46.44 1.79  3.70 13.83  29.00   136.62
```

Problem 3

The raw data collected by Ofir and Simonson (2001) to measure the NFC (need for cognition) are available in the file COGNITION. The data file contains 19 variables: a respondent ID and scores for 18 NFC items..

```
> trace = function(square) {sum(diag(square));}
> reverseScore = function(myX,myList,myValue,howMany)
     {
     for(i in 1:length(myList))
          {
          myIndex = myList[i];
          for(j in 1:howMany)
               {
               myX[myIndex,j]=(myValue)-myX[myIndex,j];
               }
          }
     return(myX);
     }
> setwd("C:/latex/statsMultiVariate/datasets");
> myData = na.omit(read.table('COGNITION.txt',header=FALSE,na.strings='.'));
> X  = (myData[,-1]);
> Xs = scale(myData[,-1]);
# > X = reverseScore(myData[,-1],c(3,4,5,7,8,9,12,16,17),5,18);
     # reverse score appropriate ... necessary?
     # shouldn't be, or should correlate on a negative factor...
     # Actually, tried and doesn't help
# > Xs    = scale(X);
> rownames(X) = rownames(Xs) =myData[,1];
> n = length(myData[,2]);
> S = t(as.matrix(Xs))%*%as.matrix(Xs)/(n-1);  # this is R if X=Xs
> S.e = eigen(S);
     > Lambda  = S.e$values;
     > U       = S.e$vectors;
     > P = prcomp(Xs, cor=TRUE, retx=TRUE);
     > summary(P, loadings=TRUE);
          Importance of components:
                                 PC1    PC2    PC3    PC4    PC5   PC6    PC7    PC8
          Standard deviation     2.404 1.1797 1.1036 1.0089 0.9965 0.949 0.9401 0.8770 0.8306
          Proportion of Variance 0.321 0.0773 0.0677 0.0565 0.0552 0.050 0.0491 0.0427 0.0383
          Cumulative Proportion  0.321 0.3984 0.4661 0.5226 0.5778 0.628 0.6769 0.7196 0.7580


> ## Bartlett Test of Sphericity
> p  = 18;     lnS  = log(det(S)); test = -1*((n-1)-(2*p+5)/6)*lnS;   crit = qchisq(.95, df=(p^2-p)/2);
> test > crit;
> print("If TRUE, Reject Null of Sphericity => continue with dimension reduction");

     > round(Lambda,digits=3);
     > VAF=100*round(Lambda/trace(S),digits=3);
     > Z  = as.matrix(Xs)%*%as.matrix(U);
```

```
> F=round(cor(Xs,Z),digits=3);     # loadings
> A=round(rbind(F,Lambda,VAF),digits=2);
> rownames(A)=c("X1","X2","X3","X4","X5","X6","X7","X8","X9","X10","X11","X12","X13","X14","X15","X
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
X1	-0.60	-0.16	0.12	0.14	-0.15	-0.41	0.06	-0.02
X2	-0.74	-0.10	0.03	0.09	-0.01	0.09	0.19	-0.20
X3	0.61	-0.08	0.12	0.01	-0.05	-0.54	-0.17	-0.12
X4	0.66	-0.33	0.11	0.21	0.01	0.03	-0.12	-0.15
X5	0.69	-0.20	0.04	0.36	-0.08	-0.07	-0.19	0.12
X6	-0.44	-0.16	0.29	0.08	0.02	0.12	-0.66	-0.29
X7	0.55	-0.20	-0.16	0.18	-0.47	0.01	0.08	0.15
X8	0.49	-0.14	0.23	-0.42	-0.57	0.20	-0.01	0.01
X9	0.56	-0.24	0.35	-0.18	0.15	-0.05	0.06	-0.20
X10	-0.62	-0.15	-0.31	-0.01	-0.04	0.22	-0.35	0.24
X11	-0.68	-0.12	-0.04	-0.12	-0.36	0.06	-0.20	0.06
X12	0.62	0.08	-0.16	0.45	0.11	0.12	-0.13	0.27
X13	-0.56	-0.51	-0.13	-0.09	-0.01	-0.25	0.11	0.13
X14	-0.55	-0.46	0.17	-0.02	0.31	-0.12	-0.02	0.28
X15	-0.41	-0.37	-0.04	0.46	-0.24	0.13	0.30	-0.23
X16	0.39	-0.44	-0.37	-0.10	0.26	0.32	0.07	-0.34
X17	0.55	-0.47	0.03	-0.29	0.15	0.08	0.05	0.31
X18	-0.21	0.03	0.76	0.19	0.05	0.33	0.17	0.21
EIGEN	5.78	1.39	1.22	1.02	0.99	0.90	0.88	0.77
%VAF	32.10	7.70	6.80	5.70	5.50	5.00	4.90	4.30

```
> ## nFactors
> library(nFactors);
    > nResults = nScree(eig = S.e$values,aparallel = parallel(subject = n, var = p)$eigen$qevpea);
> plotuScree(S.e$values);
> plotnScree(nResults, main="Component Retention Analysis");
```

a) Analyze NFC data using principal components. Discuss. **The first component only accounts for 32% of the variance which would suggest that one factor does not do an adequate job capturing the latent construct NFC. Only about one-third of the variance is captured in the instrument. Further analysis verifies that one to four factors could be included, yet the total variance is still under 55% using four factors.**

b) Discuss addition of items. **Adding items to form the construct is a convention performed by most marketers; however, it does not properly weight the items in the scale to the orthogonal construct(s). If the orthogonal projection of lambdas can be understood as linear combinations, teh best result would be a weighted average of items, where some items have more importance in the construct development; e.g., there are several methods to determine these weights: for example, Regression Method, Bartlett Scores, Anderson-Rupin Method, etc. NOTE: Generally, the first principal component scores are of the same sign; the reverse sign signifies the item could be reverse scored; however, the correlations would just be opposite. The PCA suggests that $X2$ should have a higher weight (0.74) and $X18$ should have a lower weight (0.21) in the developement of orthogonal factors that hopefully would match with a theoretical latent construct. If we just add**

them, we in doing an average weight of one for each question. I would use either method depending on the situation. If I were only using NFC and the results match previous results, I may follow the convention; the statistician in me challenges this convention, but some articles have demonstrated that either technique is equally robust in "most situations" because the summary statistic (sum, mean, etc.) is meaningful to represent a single orthogonal construct.

Problem 4

There doesn't appear to be clear vectors of data known as hyperplanes. The data seems "noisy" which would explain the low VAF (variance accounted for) in the data. As such, I would not conclude a "G" factor exists and would, in this case, not favor a simple sum of the data. NOTE: Bartlett's test of sphericity fails, so data reduction may not be wise.

```
> ## using princomp
> X.princomp = princomp(Xs); # just in case it doesn't scale
> ## princomp$loadings

> X.princomp$loadings;   ## not the same as F !!!
>    rownames(X.princomp$scores)=myData[,1];
> X.princomp$scores      ## same as Z except for sign differences
> summary(X.princomp);
    > plot(X.princomp$scores[,1],X.princomp$scores[,2]);
    > text(X.princomp$scores[,1],X.princomp$scores[,2],cex=0.7,lwd=2, labels=myData[,1]);
> library(rgl);
> plot3d(X.princomp$scores[,1:3], type="s", radius=.1, col=rainbow(300)[rank(Z[,1])]);
> text3d(X.princomp$scores[,1:3],text=myData[,1]);
```

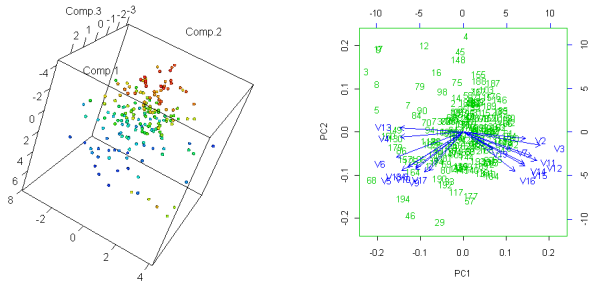

Figure 3: RGL and scatterplot

Problem 5

```
> read.tri = function(file)
     {
     x = scan(file);
     lx = length(x);
     d = (sqrt(8*lx+1)-1)/2;
     m = matrix(0, d, d);
     m[upper.tri(m, T)] = x;
     m = m + t(m) - diag(diag(m));
     return(m);
     }
> setwd("C:/latex/statsMultiVariate/datasets");
> R=read.tri("VOCATIONS.txt");
> myNames = c("public speaking","law/politics","business management","sales","merchandising","office pra
> rownames(R) = myNames;
> R.e = eigen(R);
     > Lambda  = R.e$values;
     > U       = R.e$vectors;

> ## Bartlett Test of Sphericity
> p  = 22;     lnR  = log(det(R)); test = -1*((n-1)-(2*p+5)/6)*lnR;   crit = qchisq(.95, df=(p^2-p)/2);
> test > crit;
> print("If TRUE, Reject Null of Sphericity => continue with dimension reduction");

     > round(Lambda,digits=3);
     > VAF=100*round(Lambda/trace(R),digits=3);
     > D_half  = diag(sqrt(Lambda));
     > #F=round(cor(Xs,Z),digits=3);    # loadings
     > F  = as.matrix(U)%*%D_half; #equivalent to cor(X,Z);
     > A=round(rbind(F,Lambda,VAF),digits=2);
     > rownames(A)=c(myNames,"EIGEN","%VAF");

> ## nFactors
> library(nFactors);
     > nResults = nScree(eig = R.e$values,aparallel = parallel(subject = n, var = p)$eigen$qevpea);
> plotuScree(R.e$values);
> plotnScree(nResults, main="Component Retention Analysis");

> A[,1:4]
                       [,1]  [,2]  [,3]  [,4]
public speaking       -0.72  0.09  0.39  0.07
law/politics          -0.61 -0.10  0.37  0.09
business management   -0.81 -0.42  0.08 -0.16
sales                 -0.78 -0.27  0.14 -0.12
merchandising         -0.82 -0.36  0.12 -0.16
office practice       -0.68 -0.44 -0.04 -0.25
military activities   -0.40 -0.22 -0.32  0.33
technical supervision -0.69 -0.43 -0.14 -0.13
```

Problem 5 continued on next page...

```
mathematics           -0.07 -0.13 -0.58 -0.41
science               -0.14  0.29 -0.75 -0.28
mechanical            -0.36  0.00 -0.76 -0.18
nature                -0.36  0.56 -0.41  0.33
agriculture           -0.27  0.23 -0.34  0.65
adventure             -0.26 -0.05 -0.28  0.42
recreation leadership -0.34 -0.22 -0.12  0.61
medical service       -0.35  0.33 -0.25 -0.02
social service        -0.43  0.42  0.31  0.17
religious             -0.48  0.37  0.01  0.07
teaching              -0.58  0.40  0.07 -0.08
music                 -0.26  0.74  0.08 -0.39
art                   -0.23  0.78 -0.01 -0.22
writing               -0.35  0.65  0.38  0.01
EIGEN                  5.60  3.48  2.62  1.88
%VAF                  25.40 15.80 11.90  8.60
```

```
    > plot(F[,1],F[,2]);
    > text(F[,1],F[,2],cex=0.7,lwd=2, labels=myData[,1]);
> library(rgl);
> plot3d(F[,1:3], type="s", radius=.1, col=rainbow(300)[rank(Z[,1])]);
```

There appears to be more than one underlying dimension to describe male vocational interests. The first factor accounts for 25% of the variance. It would represent generally outgoing people (assuming the negative doesn't mean opposite for the first factor). The second factor seems to positively correlate with lack of artisan skills: music, art, and writing. The third factor seems to correlate with hard skills: math, science, and mechanical. The fourth factor seems to those who like math and are not favorable toward recreational leadership (this assumes the negative is misrepresented, so negative means liking).

Problem 6

```
> R=read.tri("DRUG_USE.txt");
> myNames = c("cigarettes","beer","wine","liquor","cocaine","tranquilizers","drug store medication","her
> rownames(R) = myNames;
> R.e = eigen(R);
     > Lambda  = R.e$values;
     > U       = R.e$vectors;

> ## Bartlett Test of Sphericity
> p  = 13;     lnR  = log(det(R)); test = -1*((n-1)-(2*p+5)/6)*lnR;   crit = qchisq(.95, df=(p^2-p)/2);
> test > crit;
> print("If TRUE, Reject Null of Sphericity => continue with dimension reduction");

     > round(Lambda,digits=3);
     > VAF=100*round(Lambda/trace(R),digits=3);
     > D_half  = diag(sqrt(Lambda));
     > #F=round(cor(Xs,Z),digits=3);    # loadings
     > F  = as.matrix(U)%*%D_half; #equivalent to cor(X,Z);
     > A=round(rbind(F,Lambda,VAF),digits=2);
     > rownames(A)=c(myNames,"EIGEN","%VAF");

> ## nFactors
> library(nFactors);
     > nResults = nScree(eig = R.e$values,aparallel = parallel(subject = n, var = p)$eigen$qevpea);
> plotuScree(R.e$values);
> plotnScree(nResults, main="Component Retention Analysis");

> A[,1:3]

                      [,1]  [,2]  [,3]
cigarettes           -0.58 -0.40 -0.06
beer                 -0.60 -0.57  0.13
wine                 -0.55 -0.56  0.21
liquor               -0.67 -0.46  0.05
cocaine              -0.44  0.41  0.05
tranquilizers        -0.61  0.37 -0.17
drug store medication -0.37  0.27  0.71
heroin               -0.42  0.45  0.14
marijane             -0.71 -0.23 -0.23
hashish              -0.69  0.07 -0.35
inhalants            -0.58  0.24  0.31
hallucinogenics      -0.52  0.47 -0.11
stimulants           -0.69  0.33 -0.23
EIGEN                 4.38  2.05  0.95
%VAF                 33.70 15.70  7.30


     > plot(F[,1],F[,2]);
```

Problem 6 continued on next page...

```
    > text(F[,1],F[,2],cex=0.7,lwd=2, labels=myData[,1]);
> library(rgl);
> plot3d(F[,1:3], type="s", radius=.1, col=rainbow(300)[rank(Z[,1])]);
```

There appears to be more than one underlying dimension to describe drug usage. Although I would hypothesize that it would represent soft (cigs, beer, wine) and hard (cocaine, etc.), two factors emerge. Positive, negative is again confusing.

Problem 7

Exercise 3.1 of Everitt text, pg 62:

Suppose that $x' = [x_1\ x_2]$ is such that $x_2 = 1 - x_1$ and $x_1 = 1$ with probability p and $x_1 = 0$ with probability $p - 1$. Find the covariance of x and its eigenvalues and eigenvectors.

Proof.

$$COV(x) = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

Given: $x_1 \sim BIN(1,p)$;$x_2 \sim BIN(1,1-p)$. Known distributions have known means and variances, so $\sigma_{ii} = p(1-p)$. The covariance term $\sigma_{12} = \sigma_{21}$ can also be determined using the expectation definition of covariance: $cov(x_1, x_2) = E[x_1 \cdot x_2] - E[x_1] \cdot E[x_2]$:

$$\begin{aligned} E[x_1 \cdot x_2] - E[x_1] \cdot E[x_2] &= E[x_1 \cdot (1 - x_1)] - \mu_1 \cdot \mu_2 \\ &= E[x_1 - x_1^2] - \mu_1 \cdot \mu_2 \\ &= E[x_1] - E[x_1^2] - p \cdot (1-p) \\ &= p - p - p \cdot (1-p) \\ &= -p \cdot (1-p) \end{aligned} \tag{1}$$

$$COR(x) = \frac{COV(x)}{\sqrt{var(x_1) \cdot var(x_2)}} = \frac{}{\sqrt{p(1-p) \cdot p(1-p)}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

(perfectly correlated in opposite directions, which makes sense.)
To find eigenvalues, solve $|R - \lambda I| = 0$. Results $\lambda = 2, 0$. Therefore the eigenvectors are:

$$\begin{bmatrix} -.0707 & -.0707 \\ .0707 & -.0707 \end{bmatrix}$$

□

Problem 8

Exercise 3.3 of Everitt text, pg 62:

The proof follows from the concept of matrix multiplication.

Proof. Since we are given the λ_i's.

$VAR(PC_i) = a_i' \cdot R \cdot a_i = a_i' \cdot \lambda_i \cdot a_i$

Therefore, if $a_i' \cdot a_i = 1$, then $VAR(PC_i) = \lambda_i$. □