**Carapace Measurements for Female Turtles**

- Data on three dimensions of female turtle carapaces (shells):
  - $X_1 = \log(\text{carapace length})$
  - $X_2 = \log(\text{carapace width})$
  - $X_3 = \log(\text{carapace height})$

- Since the measurements are all on the same scale, we extracted the PCs from the sample covariance matrix $S$

- See SAS program and output.
/* This program is posted as turtlef.sas. It performs a principal component analysis of carapace measurements for female turtles. */

data females;
  input id $ length width height;
  length=log(length);
  width=log(width);
  height=log(height);
datalines;
...
/* Compute estimates of principal components from the sample covariance matrix and plot the scores */
proc princomp data=females cov n=3
  out=scores prefix=prin;
  var length width height;
run;
proc print data=scores; run;

proc plot data=scores;
  plot prin1*prin2 / hpos=56;
  plot prin1*(length width height)=id / hpos=56;
run;

proc corr data=scores;
  var prin1 prin2 prin3 length width height;
run;
/* Compute estimates of principal components from the sample correlation matrix and plot the scores */

proc princomp data=females out=pcorr prefix=pcorr;
  var length width height;
  run;

proc plot data=pcorr;
  plot pcorr1*pcorr2=id / hpos=56;
  run;
### Eigenvalues of the Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.06622582</td>
<td>0.06545996</td>
<td>0.9806</td>
<td>0.9806</td>
</tr>
<tr>
<td>2</td>
<td>0.00076585</td>
<td>0.00022165</td>
<td>0.0113</td>
<td>0.9919</td>
</tr>
<tr>
<td>3</td>
<td>0.00054420</td>
<td></td>
<td>0.0081</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

### Eigenvectors

<table>
<thead>
<tr>
<th></th>
<th>prin1</th>
<th>prin2</th>
<th>prin3</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>0.626665</td>
<td>0.552570</td>
<td>-0.549506</td>
</tr>
<tr>
<td>width</td>
<td>0.487816</td>
<td>0.271745</td>
<td>0.829572</td>
</tr>
<tr>
<td>height</td>
<td>0.607723</td>
<td>-0.787922</td>
<td>-0.099260</td>
</tr>
</tbody>
</table>
Female Painted Turtles: Using R – I

# use principal components on the logarithms of the female dataset
 turtlesf.pc <- prcomp(log(turtlesf))
 turtlesf.pc

Standard deviations:
[1] 0.25734377 0.02767404 0.02332817

Rotation:

<table>
<thead>
<tr>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
</tr>
</thead>
<tbody>
<tr>
<td>carapace.length</td>
<td>0.6266648</td>
<td>-0.5525704</td>
</tr>
<tr>
<td>carapace.width</td>
<td>0.4878158</td>
<td>-0.2717450</td>
</tr>
<tr>
<td>carapace.height</td>
<td>0.6077228</td>
<td>0.7879217</td>
</tr>
</tbody>
</table>

# obtain the variances of these principal components
 cumsum(turtlesf.pc$sdev^2)/sum(turtlesf.pc$sdev^2)

# [1] 0.980602 0.991942 1.000000
Female Painted Turtles: Using R – II

# with the correlation matrix (use scale = T)
turtlesf.pcr <- prcomp(log(turtlesf), scale = T)
turtlesf.pcr

Standard deviations:
[1] 1.7147071 0.1829964 0.1621475

Rotation:

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
</tr>
</thead>
<tbody>
<tr>
<td>carapace.length</td>
<td>0.5780354</td>
<td>-0.1130261</td>
<td>-0.8081461</td>
</tr>
<tr>
<td>carapace.width</td>
<td>0.5771777</td>
<td>-0.6434538</td>
<td>0.5028252</td>
</tr>
<tr>
<td>carapace.height</td>
<td>0.5768371</td>
<td>0.7570946</td>
<td>0.3067029</td>
</tr>
</tbody>
</table>

cumsum(turtlesf.pcr$sdev^2)/sum(turtlesf.pcr$sdev^2)
[1] 0.9800735 0.9912361 1.0000000
Five Socioeconomic Variables

- Data on socioeconomic variables: population (in thousands), median school years, total employment (in thousands), health services employment (in hundreds) and median home value (in tens of thousands) were obtained for $n = 14$ census tracks in Madison, Wisconsin.

- We extracted the PCs using both the covariance matrix $S$ and the correlation matrix $R$ (for illustration).

- See SAS program and output.
/* This program is posted as madison.sas. It illustrates how to apply principal component analysis to a covariance or correlation matrix when the covariance matrix is entered instead of the raw data. The variables are:
   X1 = total population (thousands)
   X2 = professional degrees (percent)
   X3 = employed age over 16 (percent)
   X4 = government employment (percent)
   X5 = median home value ($100,000) */
/* Enter the data as a TYPE=COV data set */

DATA madison(TYPE=COV);
   INPUT _TYPE_ $ _NAME_ $ x1-x5;
datalines;
   cov X1  3.397 -1.102  4.306  -2.078  0.027
   cov X2 -1.102  9.673 -1.513  10.953  1.203
   cov X3  4.306 -1.513 55.626 -28.937 -0.044
   cov X4 -2.078 10.953 -28.937  89.067  0.957
   cov X5  0.027  1.203 -0.044  0.957  0.319
run;

PROC PRINT data=madison;
   title "Principal Component Analysis";
   title2 "Madison: Five Socioeconomic Variables";
   run;
/* Compute principal components from the covariance matrix*/

PROC factor DATA=madison(TYPE=CORR) SCREE
   SCORE METHOD=PRIN N=5 COV EV REORDER;
run;

/* Compute principal components from the correlation matrix*/

PROC factor DATA=madison(TYPE=CORR) SCREE
   SCORE METHOD=PRIN N=5 EV REORDER;
run;
The FACTOR Procedure

Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Covariance Matrix:
Total = 158.082 Average = 31.6164

<table>
<thead>
<tr>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 107.015101</td>
<td>67.342335</td>
<td>0.6770</td>
<td>0.6770</td>
</tr>
<tr>
<td>2 39.672766</td>
<td>31.301634</td>
<td>0.2510</td>
<td>0.9279</td>
</tr>
<tr>
<td>3 8.371131</td>
<td>5.503229</td>
<td>0.0530</td>
<td>0.9809</td>
</tr>
<tr>
<td>4 2.867903</td>
<td>2.712804</td>
<td>0.0181</td>
<td>0.9990</td>
</tr>
<tr>
<td>5 0.155098</td>
<td></td>
<td>0.0010</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

5 factors will be retained by the NFACTOR criterion.
The FACTOR Procedure

Initial Factor Method: Principal Components

Eigenvectors

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>-0.03889</td>
<td>0.07116</td>
<td>-0.18788</td>
<td>0.97714</td>
<td>-0.05764</td>
</tr>
<tr>
<td>x2</td>
<td>0.10532</td>
<td>0.12975</td>
<td>0.96100</td>
<td>0.17134</td>
<td>-0.13855</td>
</tr>
<tr>
<td>x3</td>
<td>-0.49236</td>
<td>0.86439</td>
<td>-0.04580</td>
<td>-0.09105</td>
<td>0.00497</td>
</tr>
<tr>
<td>x4</td>
<td>0.86307</td>
<td>0.48033</td>
<td>-0.15318</td>
<td>-0.02969</td>
<td>0.00670</td>
</tr>
<tr>
<td>x5</td>
<td>0.00912</td>
<td>0.01473</td>
<td>0.12499</td>
<td>0.08164</td>
<td>0.98864</td>
</tr>
</tbody>
</table>
### Factor Pattern

<table>
<thead>
<tr>
<th></th>
<th>Factor1</th>
<th>Factor2</th>
<th>Factor3</th>
<th>Factor4</th>
<th>Factor5</th>
</tr>
</thead>
<tbody>
<tr>
<td>x4</td>
<td>0.94604</td>
<td>0.32057</td>
<td>-0.04696</td>
<td>-0.00533</td>
<td>0.00028</td>
</tr>
<tr>
<td>x3</td>
<td>-0.68291</td>
<td>0.72999</td>
<td>-0.01777</td>
<td>-0.02067</td>
<td>0.00026</td>
</tr>
<tr>
<td>x2</td>
<td>0.35031</td>
<td>0.26277</td>
<td>0.89399</td>
<td>0.09330</td>
<td>-0.01754</td>
</tr>
<tr>
<td>x1</td>
<td>-0.21826</td>
<td>0.24318</td>
<td>-0.29493</td>
<td>0.89782</td>
<td>-0.01232</td>
</tr>
<tr>
<td>x5</td>
<td>0.16708</td>
<td>0.16426</td>
<td>0.64028</td>
<td>0.24480</td>
<td>0.68936</td>
</tr>
</tbody>
</table>

### Variance Explained by Each Factor

<table>
<thead>
<tr>
<th>Factor</th>
<th>Weighted</th>
<th>Unweighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor1</td>
<td>107.015101</td>
<td>1.55963970</td>
</tr>
<tr>
<td>Factor2</td>
<td>39.672766</td>
<td>0.79081848</td>
</tr>
<tr>
<td>Factor3</td>
<td>8.371131</td>
<td>1.29868638</td>
</tr>
<tr>
<td>Factor4</td>
<td>2.867903</td>
<td>0.87517608</td>
</tr>
<tr>
<td>Factor5</td>
<td>0.155098</td>
<td>0.47567937</td>
</tr>
</tbody>
</table>
The FACTOR Procedure
Initial Factor Method: Principal Components

Scoring Coefficients Estimated by Regression

Standardized Scoring Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Factor1</th>
<th>Factor2</th>
<th>Factor3</th>
<th>Factor4</th>
<th>Factor5</th>
</tr>
</thead>
<tbody>
<tr>
<td>x4</td>
<td>0.78737786</td>
<td>0.71969778</td>
<td>-0.499655</td>
<td>-0.1654548</td>
<td>0.16051423</td>
</tr>
<tr>
<td>x3</td>
<td>-0.354976</td>
<td>1.02353352</td>
<td>-0.1180527</td>
<td>-0.4010121</td>
<td>0.0941982</td>
</tr>
<tr>
<td>x2</td>
<td>0.03166402</td>
<td>0.06406882</td>
<td>1.03302566</td>
<td>0.31467917</td>
<td>-1.0941605</td>
</tr>
<tr>
<td>x1</td>
<td>-0.0069283</td>
<td>0.02082245</td>
<td>-0.1196833</td>
<td>1.06346316</td>
<td>-0.2697694</td>
</tr>
<tr>
<td>x5</td>
<td>0.00049803</td>
<td>0.00132078</td>
<td>0.02439921</td>
<td>0.02722947</td>
<td>1.41785012</td>
</tr>
</tbody>
</table>
The FACTOR Procedure
Initial Factor Method: Principal Components

Eigenvalues of the Correlation Matrix: Total = 5  Average = 1

<table>
<thead>
<tr>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.99164330</td>
<td>0.3983</td>
<td>0.3983</td>
</tr>
<tr>
<td>2</td>
<td>1.36728852</td>
<td>0.2735</td>
<td>0.6718</td>
</tr>
<tr>
<td>3</td>
<td>0.86417925</td>
<td>0.1728</td>
<td>0.8446</td>
</tr>
<tr>
<td>4</td>
<td>0.53519675</td>
<td>0.1070</td>
<td>0.9517</td>
</tr>
<tr>
<td>5</td>
<td>0.24169218</td>
<td>0.0483</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

5 factors will be retained by the NFACTOR criterion.
Principal Component Analysis
Madison: Five Socioeconomic Variables

The FACTOR Procedure
Initial Factor Method: Principal Components

Eigenvectors

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>-0.26281</td>
<td>0.46281</td>
<td>0.78391</td>
<td>-0.21701</td>
<td>-0.23481</td>
</tr>
<tr>
<td>x2</td>
<td>0.59323</td>
<td>0.32591</td>
<td>-0.16384</td>
<td>0.14523</td>
<td>-0.70280</td>
</tr>
<tr>
<td>x3</td>
<td>-0.32593</td>
<td>0.60506</td>
<td>-0.22466</td>
<td>0.66277</td>
<td>0.19481</td>
</tr>
<tr>
<td>x4</td>
<td>0.47928</td>
<td>-0.25233</td>
<td>0.55083</td>
<td>0.57129</td>
<td>0.27719</td>
</tr>
<tr>
<td>x5</td>
<td>0.49303</td>
<td>0.49983</td>
<td>-0.06899</td>
<td>-0.40765</td>
<td>0.57979</td>
</tr>
</tbody>
</table>
### Factor Pattern

<table>
<thead>
<tr>
<th></th>
<th>Factor1</th>
<th>Factor2</th>
<th>Factor3</th>
<th>Factor4</th>
<th>Factor5</th>
</tr>
</thead>
<tbody>
<tr>
<td>x2</td>
<td>0.83720</td>
<td>0.38109</td>
<td>-0.15231</td>
<td>0.10625</td>
<td>-0.34551</td>
</tr>
<tr>
<td>x5</td>
<td>0.69579</td>
<td>0.58445</td>
<td>-0.06413</td>
<td>-0.29823</td>
<td>0.28504</td>
</tr>
<tr>
<td>x4</td>
<td>0.67639</td>
<td>-0.29505</td>
<td>0.51206</td>
<td>0.41794</td>
<td>0.13627</td>
</tr>
<tr>
<td>x3</td>
<td>-0.45997</td>
<td>0.70750</td>
<td>-0.20885</td>
<td>0.48486</td>
<td>0.09577</td>
</tr>
<tr>
<td>x1</td>
<td>-0.37089</td>
<td>0.54117</td>
<td>0.72873</td>
<td>-0.15876</td>
<td>-0.11544</td>
</tr>
</tbody>
</table>

### Variance Explained by Each Factor

<table>
<thead>
<tr>
<th></th>
<th>Factor1</th>
<th>Factor2</th>
<th>Factor3</th>
<th>Factor4</th>
<th>Factor5</th>
</tr>
</thead>
<tbody>
<tr>
<td>x2</td>
<td>1.9916433</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x5</td>
<td></td>
<td>1.3672885</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x4</td>
<td></td>
<td></td>
<td>0.8641793</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x3</td>
<td></td>
<td></td>
<td></td>
<td>0.5351967</td>
<td></td>
</tr>
<tr>
<td>x1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2416922</td>
</tr>
</tbody>
</table>
The FACTOR Procedure

Initial Factor Method: Principal Components

Scoring Coefficients Estimated by Regression

<table>
<thead>
<tr>
<th></th>
<th>Factor1</th>
<th>Factor2</th>
<th>Factor3</th>
<th>Factor4</th>
<th>Factor5</th>
</tr>
</thead>
<tbody>
<tr>
<td>x2</td>
<td>0.42035653</td>
<td>0.27872241</td>
<td>-0.1762451</td>
<td>0.19851896</td>
<td>-1.429546</td>
</tr>
<tr>
<td>x5</td>
<td>0.34935247</td>
<td>0.42745266</td>
<td>-0.0742123</td>
<td>-0.5572272</td>
<td>1.17934947</td>
</tr>
<tr>
<td>x4</td>
<td>0.33961153</td>
<td>-0.2157921</td>
<td>0.59253715</td>
<td>0.78091145</td>
<td>0.56382512</td>
</tr>
<tr>
<td>x3</td>
<td>-0.2309476</td>
<td>0.51744578</td>
<td>-0.2416705</td>
<td>0.90594901</td>
<td>0.39625006</td>
</tr>
<tr>
<td>x1</td>
<td>-0.1862214</td>
<td>0.39579816</td>
<td>0.84326307</td>
<td>-0.296637</td>
<td>-0.4776133</td>
</tr>
</tbody>
</table>
Five Socioeconomic Variables

• How many PCs to keep? When using $S$, we note that we can explain over 93% of the variability with the first two PCs.

• Thus, reducing the dataset from five variables to two PCs appears reasonable.

• The number of PCs retained will depend on the relative sizes of the eigenvalues of the covariance, or correlation, matrix, which depend on relative sizes of variances of the original traits and correlation patterns.

• Scree plots are sometimes useful.

• Interpretation is important.
Scree Plot for Analysis of Five Socioeconomic Variables using $S$
Scree Plot for Analysis of Five Socioeconomic Variables using \textit{R}
Interpretation of Principal Components

- Interpretation can be difficult as well. In this example, when using $S$:

  1. First PC is a contrast between percentage of the population employed in government jobs ($x_4$) and the percentage of adults who are employed ($X_3$). Scores for this component are large for tracts with relatively high government employment and relatively low adult employment rate.

  2. Second PC is weighted sum of all five variables, with the greatest weight on the adult employment percentage. This component has large scores for tracts with relatively high adult employment rates and relative high values for most of the other four variables.
Interpretation of PCs

- Typically, we will look at both the coefficients (the $\widehat{e}_{ik}$) and the contributions of each variable (the $r_{\widehat{y}_i,x_k}$) in order to interpret the meaning of the PC.

- The PC scores for each observation constitute the 'new' dataset.

- We can explore them just like we explore variables before moving on to further analyses.
Exploring Properties of PCs

- If the observation vectors $X$ are samples from a population with a normal distribution, then the scores for the $i$-th PC have a normal distribution with mean 0 and variance $\lambda_i$.

- Also, scores for different PCs are uncorrelated and
  
  $X \sim N_p(\mu, \Sigma) \rightarrow Y \sim N_p(0, \Lambda),$

  where $\Lambda$ is the diagonal matrix with $\lambda_i$ along the main diagonal.

- We can investigate whether PC scores are approximately normal by looking at Q-Q plots of the scores and applying tests.

- We should also check for outliers using the PC scores.
Q-Q Plot PC1 Scores

Q–Q plot for PC1 in Madison example
Q-Q Plot PC2 Scores

Q–Q plot for PC2 in Madison example
Large Sample Inferences

- The eigenvalue-eigenvector pairs of $S$ (or of $R$) form the basis of principal component analysis.
  - The eigenvectors determine the directions of maximum variability.
  - The eigenvalues are the variances of the linear combinations.

- Because estimates $(\hat{\lambda}, \hat{e})$ are subject to sampling variability, it would be useful to know their sampling distributions so the accuracy of the estimates can be assessed.

- It might be of interest to test whether the $q$ smallest eigenvalues are equal in order to decide if the PC’s associated with the $q$ smallest eigenvalues only represent random variation with no pattern.
Sampling distributions of \((\hat{\lambda}, \hat{\varepsilon})\)

- The sampling distributions of \((\hat{\lambda}, \hat{\varepsilon})\) are difficult to derive so we present results without proof.

- For large \(n\), and for \(\hat{\lambda} = [\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_p]'\) and for \(\lambda\) the corresponding \(p\)-dimensional vector of unobservable population eigenvalues of \(\Sigma\),

\[
\sqrt{n}(\hat{\lambda} - \lambda) \rightarrow \mathcal{N}_p(0, 2\Lambda^2),
\]

where \(\Lambda\) is diagonal \(\{\lambda_p\}\).
Sampling distributions of \((\hat{\lambda}, \hat{e})\)

- This result implies that the eigenvalues are independently distributed and that the \(i\)th sample eigenvalue has an approximate distribution that is \(N(\lambda_i, 2\lambda_i^2/n)\).

- Then, a large sample \(100(1 - \alpha)\%\) CI for \(\lambda_i\) is given by

\[
\frac{\hat{\lambda}_i}{1 + z(\alpha/2)\sqrt{2/n}} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{1 - z(\alpha/2)\sqrt{2/n}}.
\]

- An alternative approximation is \(\ln(\hat{\lambda}_i) \sim N(\ln(\lambda_i), 2/n)\) which yields an approximate confidence interval

\[
\hat{\lambda}_i \exp\left(-z_{\alpha/2}\sqrt{2/n}\right) \leq \lambda_i \leq \hat{\lambda}_i \exp\left(z_{\alpha/2}\sqrt{2/n}\right)
\]
Sampling distributions of $(\hat{\lambda}, \hat{e})$

- For large samples, $\sqrt{n}(\hat{e}_i - e_i)$ is approximately distributed as $N_p(0, E_i)$, where

$$E_i = \lambda_i \sum_{k=1, k \neq i}^{p} \frac{\lambda_k}{\lambda_k - \lambda_i} 2 e_k e'_k,$$

- This implies that the $\hat{e}_i$ are normally distributed around the 'true' $e_i$ for large $n$. The elements of $\hat{e}_i$ are correlated, and the size of the correlations depend on the separation of the true $\lambda_i$s.
Sampling distributions of \((\hat{\lambda}, \hat{e})\)

- Estimated variances for \(\hat{e}_{ik}\) are given by the diagonal elements of \(\hat{E}_i/n\), where \(\hat{E}_i\) is obtained by substituting estimates \(\hat{\lambda}_i\) and \(\hat{e}_i\) for the true values.

- For large samples, each \(\hat{\lambda}_i\) is approximately distributed independently of the elements of \(\hat{e}_i\).

- The results listed above assume that the p-dimensional vectors of responses are a random sample from a p-dimensional normal distribution.
Test for whether first \( q \) eigenvalues contain all the variation

- Test \( H_0 : \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq \eta \) against \( H_a : \frac{\sum_{i=1}^{q} \lambda_1}{\sum_{i=1}^{p} \lambda_i} < \text{eta} \)

- Large-sample test rejects \( H_0 \) if

\[
\frac{\sum_{i=1}^{q} \hat{\lambda}_i}{\sum_{i=1}^{p} \hat{\lambda}_i} - \eta < -z_{\alpha} \frac{\sqrt{2(\sum_{i=q+1}^{p} \hat{\lambda}_i)^2 \sum_{i=1}^{q} \hat{\lambda}_i^2 + 2(\sum_{i=1}^{q} \hat{\lambda}_i)^2 \sum_{i=q+1}^{p} \hat{\lambda}_i^2}}{\sqrt{n(\sum_{i=1}^{p} \hat{\lambda}_i)^2}}
\]

- follows from the first-Order Multivariate Delta Theorem, and fact that under \( H_0 \), \( \frac{\sum_{i=1}^{q} \lambda_1}{\sum_{i=1}^{p} \lambda_i} = 1 \).

If \( \hat{\Theta} \sim N_p(\Theta, \Gamma) \), then \( g(\hat{\Theta}) \sim N_p(g(\Theta), \nabla g' \Gamma \nabla g) \) where \( \nabla g \) denote the gradient vector of the real-valued differentiable function \( g \).
**R code for testing for adequacy of PCs**

```r
PCs.proportion.variation.enuff <- function(lambda, q = 1, propn, nobs) {
  den <- sum(lambda) # sum of all the eigenvalues
  num <- sum(lambda[1:q]) # sum of the first q eigenvalues
  if (num/den >= propn) return(1)
  else {
    se <- sqrt(2 * sum(lambda[-(1:q)])^2 * sum(lambda[1:q]^2) +
              2 * sum(lambda[1:q])^2 * sum(lambda[-(1:q)]^2)) /
             (nobs * den^2)
    # asymptotic sd of the test statistic
    test.stat <- (num/den - propn)/se
    return(pnorm(test.stat))
  }
}

PCs.proportion.variation.enuff(turtlesf.pc$sdev^2, q = 1, propn = 0.9825, nobs = 24)
[1] 0.08440965
```
Test for equal eigenvalues

• Test the null hypothesis $H_0: \lambda_{q+1} = \lambda_{q+2} = \cdots = \lambda_{q+r}$ that the $r$ smallest population eigenvalues are equal. ($p=q+r$)

• Large sample chi-square test rejects the null hypothesis if

$$X^2 = (v)(r)ln \left[ \frac{1}{r} \sum_{i=q+1}^{q+r} \tilde{\lambda}_i \right] - v \sum_{i=q+1}^{q+r} ln(\tilde{\lambda}_i) > \chi^2_{r(r+1)/2-1}$$

where $v = (df \ for \ S) - (2p+5)/6$

• This test is derived from the check on the equality of the ratio of the geometric mean of the roots to the arithmetic mean of the roots with one.
**Test for equal eigenvalues**

- The above follows from the second-order **Multivariate Delta Theorem** (since $\nabla \equiv 0$):
  If $\hat{\Theta} \sim N_p(\Theta, \Gamma)$, then $g(\hat{\Theta}) - g(\Theta)$ is asymptotically distributed as $-\frac{1}{2} Y' \mathcal{H}_g Y$ where $Y \sim N_p(0, \Gamma)$ and $\mathcal{H}_g$ is the Hessian of $g$.

- This is useful in determining if $\Sigma$ has some structure to the matrix: *i.e.*, we may be interested in testing whether $\Sigma = \Phi + \sigma^2 I$, where $\Phi$ is a rank-deficient matrix of rank $q(<p)$. Alternatively, whether $X = Y + \sigma Z + \mu$, where $Y$ is independent of $Z \sim N_p(0, I)$, and $Y$ is normally distributed with zero mean and variance $\Phi$, *i.e.*, it lies in some $q$-dimensional space.
R code for testing for structure in $\Sigma$

- see file: last.PCs.equal.variances.R

```r
last.PCs.equal.variances <- function(lambda, r = 2, df) {
  # df = df of the sample variance-covariance matrix
  p <- length(lambda)
  v <- df - (2 * p + 5)/6
  Xi <- v*(r*log(mean(lambda[-(1:(p-r))]))-sum(log(lambda[-(1:(p-r))])))
  return(pchisq(Xi, df = r * (r + 1)/2 -1, lower.tail=F))
}

last.PCs.equal.variances(lambda = turtlesf.pc$sdev^2, p = 2, df = 23)
# [1] 0.7353765
```
Female Turtles Data

- Pairwise scatter plots of original variables and PCs.
All Turtles Data

- Plots of male (blue) and female (green) turtles.
Using Pooled or Partial Covariance Matrices

- The data posted as turtles.dat has data on both male and female turtles.
- We can perform PC’s on the pooled variance-covariance matrix.
- Note that the code is not that direct....
- Plots of PCs of the male (blue) and female (green) turtles.
# This code creates scatter plot matrices and
# principal components for the 100k road race
# data (Everitt 1994). This code is posted as
# race100k.R. The data are posted as race100k.dat
#
# There is one line of data for each of 80
# racers with eleven numbers on each line.
# The first ten columns give the times (minutes)
# to complete successive 10k segments of the race.
# The last column has the racer’s age (in years).

race.mat <- matrix(scan("c:/stat501/data/race100k.dat"),
                  ncol=11,byrow=T)
# First compute the number of columns in the matrix

```r
p1 <- dim(race.mat)[2]
```

# Compute sample size and the number of section times

```r
n <- dim(race.mat)[1]
p <- p1 - 1
```

# Use the pairs function to create a scatter plot matrix. Note that the columns to be included in the plot are put into the "choose" list. The panel.smooth function uses locally weighted regression to pass a smooth curve through each plot. The abline function uses least squares to fit a straight line to each plot. This helps you to
# see if most of the marginal association between two
# variables on can be described by a straight line. Recall
# that principal components are computed from variances
# and covariances (or correlations), which can only account
# for straight line relationships.

```r
par(pch=5,fin=c(5,5))
choose<-c(1,2,6,10,11)
pairs(race.mat[,choose],labels=c("0-10k time", "10-20k time","50-60k time","90-100k time","age"),
panel=function(x,y){panel.smooth(x,y)
abline(lsfit(x,y),lty=2) })
```
# Compute principal components from the covariance matrix.  
# This function creates a list with the following components  
#   sdev: standard deviations of the component scores (  
#          square roots of eigenvalues of the covariance  
#          matrix)  
#   rotation: The coefficients needed to compute the scores  
#          (elements of eigenvectors)  
#   x: a nxp matrix of scores

race.pc <- prcomp(race.mat[, -p1])

# Print the results

race.pc$sdev

637
<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1288</td>
<td>-0.2106</td>
<td>0.3615</td>
<td>-0.0335</td>
<td>0.1473</td>
<td>-0.2058</td>
</tr>
<tr>
<td>2</td>
<td>0.1520</td>
<td>-0.2491</td>
<td>0.4168</td>
<td>-0.0708</td>
<td>0.2238</td>
<td>-0.1309</td>
</tr>
<tr>
<td>3</td>
<td>0.1992</td>
<td>-0.3143</td>
<td>0.3411</td>
<td>-0.0539</td>
<td>0.2470</td>
<td>0.0526</td>
</tr>
<tr>
<td>4</td>
<td>0.2397</td>
<td>-0.3300</td>
<td>0.2027</td>
<td>-0.0066</td>
<td>0.0047</td>
<td>0.1439</td>
</tr>
<tr>
<td>5</td>
<td>0.3144</td>
<td>-0.3021</td>
<td>-0.1351</td>
<td>0.1107</td>
<td>-0.3564</td>
<td>0.2846</td>
</tr>
<tr>
<td>6</td>
<td>0.4223</td>
<td>-0.2147</td>
<td>-0.2223</td>
<td>-0.0868</td>
<td>-0.3730</td>
<td>0.2916</td>
</tr>
<tr>
<td>7</td>
<td>0.3359</td>
<td>0.0496</td>
<td>-0.1936</td>
<td>-0.6016</td>
<td>-0.1897</td>
<td>-0.6436</td>
</tr>
<tr>
<td>8</td>
<td>0.4067</td>
<td>0.0086</td>
<td>-0.5380</td>
<td>0.1290</td>
<td>0.7198</td>
<td>0.0348</td>
</tr>
<tr>
<td>9</td>
<td>0.3990</td>
<td>0.2675</td>
<td>0.1492</td>
<td>0.7175</td>
<td>-0.2098</td>
<td>-0.4142</td>
</tr>
<tr>
<td>10</td>
<td>0.3854</td>
<td>0.6888</td>
<td>0.3482</td>
<td>-0.2789</td>
<td>0.0545</td>
<td>0.4051</td>
</tr>
<tr>
<td></td>
<td>PC7</td>
<td>PC8</td>
<td>PC9</td>
<td>PC10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[1,]</td>
<td>0.4324</td>
<td>-0.2802</td>
<td>0.0389</td>
<td>0.6900</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[2,]</td>
<td>0.3256</td>
<td>-0.2294</td>
<td>0.0463</td>
<td>-0.7128</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[3,]</td>
<td>-0.3435</td>
<td>0.4576</td>
<td>-0.5868</td>
<td>0.0829</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[4,]</td>
<td>-0.4479</td>
<td>0.1045</td>
<td>0.7451</td>
<td>0.0708</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[5,]</td>
<td>-0.2450</td>
<td>-0.6462</td>
<td>-0.3061</td>
<td>-0.0054</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[6,]</td>
<td>0.5390</td>
<td>0.4494</td>
<td>0.0379</td>
<td>-0.0229</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[7,]</td>
<td>-0.1844</td>
<td>-0.0219</td>
<td>-0.0193</td>
<td>-0.0190</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[8,]</td>
<td>0.0293</td>
<td>-0.0817</td>
<td>0.0366</td>
<td>0.0180</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[9,]</td>
<td>-0.0461</td>
<td>0.1132</td>
<td>-0.0027</td>
<td>-0.0398</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[10,]</td>
<td>-0.0300</td>
<td>-0.0945</td>
<td>-0.0025</td>
<td>0.0320</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# compute proportion of total variance explained by each component

```r
s <- var(race.pc$x)
pvar<-round(diag(s)/sum(diag(s)), digits=6)
cat("proportion of variance: ", pvar, fill=T)
```

```
proportion of variance:  0.747716 0.100095 0.054130 0.037855 
  0.026458 0.017520 0.008165 0.004317
  0.002433 0.001311
```
cpvar <- round(cumsum(diag(s))/sum(diag(s)), digits=6)
cat("cumulative proportion of variance: ", cpvar, fill=T)

cumulative proportion of variance:  0.747716  0.847812  0.901941
                                 0.939796  0.966254  0.983774
                                 0.991939  0.996256  0.998689
                                 1

# plot component scores

par(pch=5, fin=c(5,5))
pairs(race.pc$x[,c(1,2,3)],labels=c("PC1","PC2","PC3"))
# To compute principal components from a correlation matrix, you must first standardize the data

# race.s <- scale(race.mat, center=T, scale=T)

# Plot standardized data

choose<-c(1,2,5,10,11)

pairs(race.s[,choose],labels=c("0-10k time", "10-20k time", "50-60k time", "90-100k time", "age"),
panel=function(x,y){panel.smooth(x,y)
    abline(lsfit(x,y),lty=2) })
# Compute principal components from the correlation matrix

```r
race.cor <- var(race.s)
cat("correlation matrix for 10k splits:", fill=T)
```

```
[1,] 1.000000 0.951060 0.844587 0.785856 0.620534 0.617892
[2,] 0.951060 1.000000 0.890311 0.826125 0.641443 0.632765
[3,] 0.844587 0.890311 1.000000 0.921086 0.755946 0.725099
[4,] 0.785856 0.826125 0.921086 1.000000 0.886909 0.841856
[5,] 0.620534 0.641443 0.755946 0.886909 1.000000 0.936414
[6,] 0.617892 0.632765 0.725099 0.841856 0.936414 1.000000
[7,] 0.531396 0.540931 0.605026 0.690654 0.754197 0.839576
[8,] 0.477372 0.505452 0.619982 0.698215 0.785781 0.840322
[9,] 0.542343 0.533807 0.583576 0.667353 0.741349 0.772573
[10,] 0.414260 0.438128 0.467253 0.508577 0.541742 0.655918
[11,] 0.149172 0.127104 0.012183 0.046802 -0.016074 -0.042419
```

645
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>0.53139648</td>
<td>0.4773723</td>
<td>0.5423438</td>
<td>0.4142609</td>
<td>0.14917250</td>
</tr>
<tr>
<td>[2,]</td>
<td>0.54093190</td>
<td>0.5054520</td>
<td>0.5338073</td>
<td>0.4381283</td>
<td>0.12710409</td>
</tr>
<tr>
<td>[3,]</td>
<td>0.60502621</td>
<td>0.6199821</td>
<td>0.5835765</td>
<td>0.4672533</td>
<td>0.01218286</td>
</tr>
<tr>
<td>[4,]</td>
<td>0.69065419</td>
<td>0.6982152</td>
<td>0.6673533</td>
<td>0.5085772</td>
<td>0.04680206</td>
</tr>
<tr>
<td>[5,]</td>
<td>0.75419742</td>
<td>0.7857815</td>
<td>0.7413497</td>
<td>0.5417422</td>
<td>-0.01607529</td>
</tr>
<tr>
<td>[6,]</td>
<td>0.83957633</td>
<td>0.8403225</td>
<td>0.7725735</td>
<td>0.6559189</td>
<td>-0.04241971</td>
</tr>
<tr>
<td>[7,]</td>
<td>1.00000000</td>
<td>0.7796014</td>
<td>0.6972448</td>
<td>0.7191956</td>
<td>-0.04059097</td>
</tr>
<tr>
<td>[8,]</td>
<td>0.77960144</td>
<td>1.0000000</td>
<td>0.7637562</td>
<td>0.6634709</td>
<td>-0.20674428</td>
</tr>
<tr>
<td>[9,]</td>
<td>0.69724482</td>
<td>0.7637562</td>
<td>1.0000000</td>
<td>0.7797619</td>
<td>-0.12320048</td>
</tr>
<tr>
<td>[10,]</td>
<td>0.71919560</td>
<td>0.6634709</td>
<td>0.7797619</td>
<td>1.0000000</td>
<td>-0.11289354</td>
</tr>
<tr>
<td>[11,]</td>
<td>-0.04059097</td>
<td>-0.2067443</td>
<td>-0.1232005</td>
<td>-0.1128935</td>
<td>1.00000000</td>
</tr>
</tbody>
</table>
races.pc <- prcomp(race.s[, -p1])

cat("standard deviations of component scores:", fill=T)

standard deviations of component scores:
races.pc$sdev
[1] 2.6912189 1.1331038 0.7439637 0.5451001 0.4536530
[6] 0.4279130 0.3300239 0.2204875 0.1984028 0.1923427

cat("component coefficients", fill=T)
component coefficients
races.pc$rotation

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>0.2965</td>
<td>-0.4495</td>
<td>-0.2863</td>
<td>0.0574</td>
<td>-0.0860</td>
<td>0.4499</td>
</tr>
<tr>
<td>[2,]</td>
<td>0.3043</td>
<td>-0.4511</td>
<td>-0.2518</td>
<td>0.0785</td>
<td>-0.1180</td>
<td>0.1020</td>
</tr>
<tr>
<td>[3,]</td>
<td>0.3254</td>
<td>-0.3416</td>
<td>0.0276</td>
<td>-0.0205</td>
<td>-0.0860</td>
<td>-0.5216</td>
</tr>
<tr>
<td>[4,]</td>
<td>0.3444</td>
<td>-0.2008</td>
<td>0.2291</td>
<td>-0.0971</td>
<td>0.1984</td>
<td>-0.3341</td>
</tr>
<tr>
<td>[5,]</td>
<td>0.3378</td>
<td>0.0618</td>
<td>0.4470</td>
<td>-0.1872</td>
<td>0.3353</td>
<td>0.0648</td>
</tr>
<tr>
<td>[6,]</td>
<td>0.3454</td>
<td>0.1593</td>
<td>0.2933</td>
<td>0.0469</td>
<td>0.2016</td>
<td>0.1351</td>
</tr>
<tr>
<td>[7,]</td>
<td>0.3127</td>
<td>0.2711</td>
<td>0.0290</td>
<td>0.7346</td>
<td>0.1097</td>
<td>0.2510</td>
</tr>
<tr>
<td>[8,]</td>
<td>0.3122</td>
<td>0.3018</td>
<td>0.2043</td>
<td>-0.0377</td>
<td>-0.8533</td>
<td>-0.0610</td>
</tr>
<tr>
<td>[9,]</td>
<td>0.3083</td>
<td>0.2897</td>
<td>-0.2672</td>
<td>-0.6288</td>
<td>0.0684</td>
<td>0.3525</td>
</tr>
<tr>
<td>[10,]</td>
<td>0.2668</td>
<td>0.3998</td>
<td>-0.6337</td>
<td>0.0831</td>
<td>0.1845</td>
<td>-0.4359</td>
</tr>
<tr>
<td></td>
<td>PC7</td>
<td>PC8</td>
<td>PC9</td>
<td>PC10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[1,]</td>
<td>-0.1923123</td>
<td>0.01485076</td>
<td>-0.612826069</td>
<td>-0.047078428</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[2,]</td>
<td>-0.1635240</td>
<td>-0.18379664</td>
<td>0.741251560</td>
<td>0.002483905</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[3,]</td>
<td>0.3603175</td>
<td>0.54150034</td>
<td>-0.057189353</td>
<td>-0.265471742</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[4,]</td>
<td>0.1820329</td>
<td>-0.51310747</td>
<td>-0.186770540</td>
<td>0.544828894</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[5,]</td>
<td>-0.2252704</td>
<td>-0.24525627</td>
<td>0.006760302</td>
<td>-0.647837829</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[6,]</td>
<td>-0.4346583</td>
<td>0.55907502</td>
<td>0.104542699</td>
<td>0.442476362</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[7,]</td>
<td>0.4530905</td>
<td>-0.05705153</td>
<td>0.037490586</td>
<td>-0.057160164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[8,]</td>
<td>-0.1212751</td>
<td>-0.13448246</td>
<td>-0.058637963</td>
<td>-0.011755653</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[9,]</td>
<td>0.4491343</td>
<td>0.08420582</td>
<td>0.107575891</td>
<td>0.069455481</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[10,]</td>
<td>-0.3345390</td>
<td>-0.09137728</td>
<td>-0.097185763</td>
<td>-0.082188533</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
s <- var(races.pc$x)
pvar<-round(diag(s)/sum(diag(s)), digits=6)
cat("proportion of variance: ", pvar, fill=T)

proportion of variance:  0.724266  0.128392  0.055348  0.029713  0.02058  0.018311  0.010892  0.004861  0.003936  0.0037

cpvar <- round(cumsum(diag(s))/sum(diag(s)), digits=6)
cat("cumulative proportion of variance: ", cpvar, fill=T)

cumulative proportion of variance:  0.72427  0.85266  0.9080  0.93772  0.9583  0.9766  0.987503  0.99236  0.9963  1
# Use the principal component scores from the raw data
to look for differences among mature (age < 40) and
senior (age > 40) runners. Mature runners will be
indicated by "M" and senior runners will be indicated
by "S".

```r
race.type <- rep("M", n)
race.type[race.mat[, p1] >= 40] <- "S"
```

# Plot component scores

```r
par(fin=c(5,5))
plot(race.pc$x[,1], race.pc$x[,2],
     xlab= "P1: Overall speed",
     ylab= "PC2: Change in speed ", type= "n")
text(race.pc$x[,1], race.pc$x[,2], labels= race.type)
```
P1: Overall speed
PC2: Change in speed

M
M
SM
M
S
M
S
MS
S
M MS
M...M
MS
M
S
M
SS
M
S
M
S
MM
MM