At the beginning of the 20\textsuperscript{th} century, one researcher obtained measurements on seven physical characteristics for each of 3000 convicted male criminals. The characteristics he measured are:

- \(X_1\) Length of head from front to back (in cm.)
- \(X_2\) Head breadth (in cm.)
- \(X_3\) Face breadth (in cm.)
- \(X_4\) Length of left forefinger (in cm.)
- \(X_5\) Length of left forearm (in cm.)
- \(X_6\) Length of left foot (in cm.)
- \(X_7\) Height (in inches)

The sample correlation matrix is:

\[
\begin{matrix}
X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \\
X_1 & 1 & 0.402 & 0.395 & 0.301 & 0.305 & 0.399 & 0.340 \\
X_2 & 0.402 & 1 & 0.618 & 0.150 & 0.135 & 0.206 & 0.183 \\
X_3 & 0.395 & 0.618 & 1 & 0.321 & 0.289 & 0.363 & 0.345 \\
X_4 & 0.301 & 0.150 & 0.321 & 1 & 0.846 & 0.759 & 0.661 \\
X_5 & 0.305 & 0.135 & 0.289 & 0.846 & 1 & 0.797 & 0.800 \\
X_6 & 0.399 & 0.206 & 0.363 & 0.759 & 0.797 & 1 & 0.736 \\
X_7 & 0.340 & 0.183 & 0.345 & 0.661 & 0.800 & 0.736 & 1 \\
\end{matrix}
\]

The eigenvalues and eigenvectors for the sample correlation matrix are:

\[
\begin{array}{cccccccc}
\text{Eigenvalues} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
.285 & -.351 & .877 & -.088 & -.076 & .112 & -.023 \\
.211 & -.643 & -.246 & .686 & -.098 & -.010 & .020 \\
\text{Eigenvectors} & .294 & -.515 & -.387 & -.693 & -.112 & .029 & -.074 \\
 & .435 & .240 & -.113 & .126 & -.604 & .330 & .500 \\
 & .453 & .282 & -.079 & .127 & -.024 & .270 & -.787 \\
 & .453 & .167 & .028 & .023 & -.065 & -.873 & .024 \\
 & .434 & .182 & -.027 & -.090 & .776 & .208 & .352 \\
\end{array}
\]
(a) What is the percentage of the total standardized variation attributed to the second principal component?

(b) Estimate the correlation between the scores for the first and second principal components.

(c) Estimate the correlation between the scores for the first principal component and the standardized values of height ($X_7$).

(d) Estimate the variance of the scores for the first principal component.

(e) Give a formula for computing the scores of the first principal component. Give a brief interpretation of the first principal component.

(f) Give a brief interpretation of the second and third principal components.

(g) Describe how the first three principal components are related to patterns in the correlation matrix.
2. Maximum likelihood estimation was used to estimate the loadings on two factors for the data in Problem 1. The estimated loading matrix for a varimax rotation of the two factor model is shown below.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>0.270</td>
<td>0.483</td>
</tr>
<tr>
<td>X2</td>
<td>0.031</td>
<td>0.792</td>
</tr>
<tr>
<td>X3</td>
<td>0.209</td>
<td>0.764</td>
</tr>
<tr>
<td>X4</td>
<td>0.856</td>
<td>0.165</td>
</tr>
<tr>
<td>X5</td>
<td>0.957</td>
<td>0.122</td>
</tr>
<tr>
<td>X6</td>
<td>0.812</td>
<td>0.256</td>
</tr>
<tr>
<td>X7</td>
<td>0.799</td>
<td>0.221</td>
</tr>
</tbody>
</table>

(a) Compute the communality and the specific variance for the standardized height (X7) variable.

(b) What is the primary objective of a varimax rotation?

(c) Give a brief interpretation of the two rotated factors.

(d) Compute the value of the approximation to the correlation between the length of left forearm (X5) and the length of left foot (X6) provided by this two factor model.

(e) The value of the Tucker and Lewis reliability coefficient for this two factor model is 0.924. What does this coefficient measure? How would you interpret this value?
(f) Describe how the regression method would be used in this case to obtain estimates of scores for the two rotated factors. Just describe the basic features of this procedure, do not try to compute numerical values for the factor scores.

3. Consider the data from Problem 1. Indicate which of the following statistical methods have the specified property by circling the letter codes for the appropriate choices.

   A. One-way MANOVA F-test based on Wilk’s criterion
   B. Linear discriminant analysis
   C. CART (classification tree)
   D. Classification using logistic regression
   E. Principal components analysis of a covariance matrix
   F. Iterative principal factor analysis of a correlation matrix

(a) The results are **not** affected by recording the first trait \((X_1)\) in meters instead of centimeters and recording the last trait \((X_7)\) in feet instead of inches.

   A   B   C   D   E   F

(b) The results are **not** affected by applying a log-transformation to each trait.

   A   B   C   D   E   F

(c) The test or statistical method is based on an assumption of homogeneous covariance matrices across groups.

   A   B   C   D
4. In an experiment to assess possible side effects of a new drug, 25 rabbits were treated with the drug. Four blood samples were taken on each rabbit. The first blood sample was taken just before treatment started, and additional blood samples were taken at the end of each of the first 3 weeks of treatment. Measurements on the levels of each of 3 substances were made on each blood sample.

(a) Consider a $T^2$ test of the null hypothesis

$$H_0 : \left( \mu_{10}, \mu_{20}, \mu_{30} \right) = \left( \mu_{13}, \mu_{23}, \mu_{33} \right)$$

where $\mu_0 = \left( \mu_{10}, \mu_{20}, \mu_{30} \right)$ represents the mean levels of the 3 substances before treatment started and $\mu_3 = \left( \mu_{13}, \mu_{23}, \mu_{33} \right)$ represents the mean levels at the end of the third week of treatment. Give a formula for an appropriate $T^2$ test statistic. Identify any new notation that you introduce in your answer.

(b) Report degrees of freedom for the F-test corresponding to your $T^2$ test in Part A.

(c) Consider using all of the data to test the null hypothesis of no substance by time interaction. Report degrees of freedom for the F-test based on Wilk’s criterion.
A medical condition called VT occurs when defective heart tissue allows electrical impulses to deviate from the appropriate pathways. This can result in different heart chambers pumping against each other which leads to rapid heart rate and loss of blood pressure. There are other possible causes of such symptoms, however, that occur more often than VT. There is an accurate procedure for detecting VT, but it is lengthy, expensive, and involves some risk to the patient. Doctors would like to use safer and more readily available information to construct a classification rule for identifying which patients with rapid heart rate and low blood pressure symptoms are likely to have VT and should be subjected to the more accurate test for VT. The good news is that VT is easy to correct if it can be detected.

Data are available for 186 patients with a history of incidents of rapid heart beat and low blood pressure. These symptoms often resulted in dizziness and fainting spells. The accurate test for VT showed that 45 of these patients had VT and the other 141 patients did not have VT. The data include the following variables:

<table>
<thead>
<tr>
<th>AGE</th>
<th>Age of the patient (in years) at the time of the examination</th>
</tr>
</thead>
<tbody>
<tr>
<td>REASON</td>
<td>Reason for seeing a doctor, coded</td>
</tr>
<tr>
<td></td>
<td>1 passed out/dizziness</td>
</tr>
<tr>
<td></td>
<td>2 rapid heart beat</td>
</tr>
<tr>
<td></td>
<td>3 chest pains</td>
</tr>
<tr>
<td>EF</td>
<td>Ejection efficiency of the heart measured as the percentage of blood in the ventricle chamber that is ejected per heart beat. Lower values of this variable indicate more severe problems.</td>
</tr>
<tr>
<td>HV</td>
<td>Time for an electrical signal to travel between two electrodes applied to the patient. Longer times indicate more severe problems.</td>
</tr>
<tr>
<td>MI</td>
<td>Indicator of whether or not the patient experienced a myocardial infarction (a heart attack) coded 1 for yes and 0 for no.</td>
</tr>
<tr>
<td>GROUP</td>
<td>Coded 1 if VT was present</td>
</tr>
<tr>
<td></td>
<td>0 if VT was not present</td>
</tr>
</tbody>
</table>

(a) PROC DISCRIM in SAS produced the following set of coefficients for linear discriminants. Here the REASON variable was recorded as two binary variables

\[
R1 = \begin{cases} 
1 & \text{passed out / dizziness} \\
0 & \text{rapid heart beat or chest pain} 
\end{cases}
\]

\[
R2 = \begin{cases} 
1 & \text{rapid heart beat} \\
0 & \text{passed out / dizziness or chest pain} 
\end{cases}
\]
## Estimated Coefficients for Linear Discriminants

<table>
<thead>
<tr>
<th>Variable</th>
<th>VT is Present</th>
<th>VT is Not Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>-53.22</td>
<td>-54.24</td>
</tr>
<tr>
<td>AGE</td>
<td>0.18</td>
<td>0.19</td>
</tr>
<tr>
<td>R1</td>
<td>10.42</td>
<td>10.75</td>
</tr>
<tr>
<td>R2</td>
<td>10.63</td>
<td>11.65</td>
</tr>
<tr>
<td>EF</td>
<td>0.71</td>
<td>0.65</td>
</tr>
<tr>
<td>HV</td>
<td>0.64</td>
<td>0.63</td>
</tr>
<tr>
<td>MI</td>
<td>14.05</td>
<td>16.34</td>
</tr>
</tbody>
</table>

Use this information to classify a new patient that is 27 years old, with EV = 40 and HV = 60. This patient went to see a doctor after experiencing chest pain, but his patient did not experience a myocardial infarction.

(b) The estimated coefficients in part (a) were obtained using equal prior probabilities and equal costs of misclassification. Suppose the doctors want to set the prior probabilities to reflect the fact that only about 10 percent of the patients they would encounter would actually have the VT condition. Furthermore, the doctors want to adjust the misclassification costs. If a non-VT patient is incorrectly classified by the linear discriminant rule as a VT patient, that patient will be subjected to a more accurate test that involves some risk of infection and discomfort and offers no benefit to a non-VT patient. If a VT patient is classified as a non-VT patient, the VT patient will be assigned to treatments that will not be helpful. Summarizing the overall costs of these two mistakes to the future health and well-being of potential patients, the doctors decide that the cost of misclassifying a VT case as a non-VT case is twenty times greater than the cost of misclassifying a non-VT case as a VT case. Show how incorporating these prior probabilities and this ratio of misclassification costs would change the linear discriminants listed in part (a). Does the result for the patient from part (a) change?

(c) Describe conditions under which a linear discriminant rule minimizes the expected cost of misclassification. Do you think those conditions would be satisfied in this example?
(d) Describe how crossvalidation is used by PROC DISCRIM in SAS to estimate misclassification rates.

(e) Suppose the stepwise selection procedure provided by PROC STEPDISC in SAS selected EF for a linear discriminant model on the first step. Describe the test it would use to determine if including HV would make a “significant” improvement.

(f) The next two pages show output from using the tree function in S-PLUS for creating a classification tree. Use this information to determine if you should prune the classification tree. Give a brief explanation of how you decided to prune the tree. On the diagram of the tree shown on the next page, circle any branches you decide to prune off the tree.

(g) Use your pruned tree from Part (f) to classify a new patient that is 27 years old with EF = 40 and HV = 60. This patient went to the doctor because he experienced chest pain but he has never experienced a myocardial infarction (heart attack).

(h) What does the information on the next two pages tell you about the probability of misclassing a new patient who is not a member of the training samples?