1. This problem uses data from the 2004 Summer Olympics Mens Decathlon. The dataset can be loaded from R package, FactoMineR by using the following R commands:

```r
> library(FactoMineR)
> data(decathlon)
> mydata <- decathlon[~(1:13),1:10]
```

The data contains results for 28 decathletes in 10 different events: timed events in seconds and distance events in meters. Values for some results that were originally missing have been imputed in this dataset.

(a) Perform a principal components analysis of this dataset on both the covariance and correlation matrices, reporting both estimated variances and loadings. Why are the results different? Why does PCA on the correlation matrix make more sense for this data? The remaining parts assume that you use PCA on the correlation matrix.

(b) Using your results from applying PCA on the correlation matrix, explain/interpret the loadings for PC1, and explain sign pattern (positive/negative) of the loadings.

(c) Make a principal components biplot using the first 2 principal components. Label the observations according to the country of the decathlete.

(d) What does the origin (0, 0) correspond to in terms of the 10 results?

(e) Explain/interpret the clustering of the arrows in the PC biplot.

(f) The arrow corresponding to the 1500m event appears separated from the other two clusters of arrows. Why might that be? Hint: You may have to search Wikipedia or Google to find out what the 1500m event is and how it is different from the others.

(g) Which region of the PC biplot corresponds to above average performance-better results, i.e. if you were a decathlete, where would you want to be in this biplot? Make a plot and circle this region. Hint: Think carefully about distance, timing results and their relationship with performing well. Identify/circle the point corresponding to the gold medalist.

2. (Izenman, Exercise 7.3) The data set includes three variables, length, width, and height, of the carapaces of 48 painted turtles, 24 female and 24 male. Take logarithms of all three variables. Estimate the mean vector and covariance matrix of the male turtles and of the female turtles separately. Find the eigenvalues and eigenvectors of each estimated covariance matrix and carry out a PCA of each data set. Find an expression for the volume of a turtle carapace for males and for females. (Hint: use the fact that the variables are logarithms of the original measurements.) Compare volumes of male and female carapaces.
3. Let \( X \sim \text{MVN}_p(\mu, \Sigma) \). Explain the relationship between the principal components of \( X \) and the contours of its density function

\[
f(x \mid \mu, \Sigma) = \frac{1}{\sqrt{\text{det}(2\pi\Sigma)}} \exp\left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}.
\]

In particular, explain the geometry of the contours of \( f \) and its relationship to the principal components and their variances.

4. Let \( Y \) and \( X \) be data matrices of dimension \( n \times p \) and \( n \times q \), respectively, and let \( \hat{B} \) denote the least squares estimator \( B \) in the model

\[
Y = 1_n \mu^T + XB + \epsilon,
\]

where \( \epsilon \) has i.i.d mean 0 entries with variance \( \sigma^2 \).

Note: Recall \( B_* = \Sigma_{XX}^{-1} \Sigma_{XY} \) from HW3 Q1. Thus, its LSE is \( \hat{B} = S_{XX}^{-1} S_{XY} \) where \( S_{XX} = X^T C X, \Sigma_{XY} = X^T C Y \) and \( C = \frac{1}{n} 1_n 1_n^T \).

(a) Express \( \hat{B} \) in terms of \( Y \) and the sample principal components of \( X \) and the eigenvalues and eigenvectors of \( S_{XX} \).

(b) Suppose we estimate \( B \) by projecting \( X \) onto the first \( d \) principal component directions, and then regressing \( Y \) on \( Z = \tilde{X} \tilde{V}_{1:d} \tilde{V}_{1:d}^T \), where \( \tilde{X} \) is the centered version of \( X \) and \( \tilde{V}_{1:d} \) is the pd matrix whose columns are the first \( d \) eigenvectors of \( S_{XX} \). Express the resulting estimate \( \hat{B}_{pc} \) in a similar way as in the previous part.