The required files for all problems can be found in:
http://www.stat.uchicago.edu/~lekheng/courses/331/hw3/
The file name indicates which problem the file is for (p1*.txt for Problem 1, etc). You are welcomed to use any programming language or software packages you like.

1. (Factor Analysis) This is the same air quality data set we saw in the previous problem set but this time we will only take four variables $X_1, X_2, X_5$ and $X_6$ by leaving out CO, NO, and HC variables.
   (a) Obtain the principal component solution to the factor model $X = \mu + LF + \varepsilon$ with number of factors $m = 1$ and $m = 2$ using:
      (i) the sample covariance matrix;
      (ii) the sample correlation matrix.
   In other words, you should find the matrix factor loadings $L \in \mathbb{R}^{n \times m}$, the specific variances $\psi_1, \ldots, \psi_p \in \mathbb{R}$, and write down the proportions of variability (in percentages) due to the factors.
   (b) Find the angle between the first factor loading in (i) and that the first factor loading in (ii).
   (c) For the $m = 2$ case, compare the factor loadings obtained in (i) and that in (ii) using orthogonal Procrustes analysis.
   (d) Comment on your results.

2. (Population Canonical Correlation Analysis) The $2 \times 1$ random vectors $X$ and $Y$ have joint mean vector and joint covariance matrix
   \[
   \mu = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \in \mathbb{R}^{4 \times 4}, \quad \Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix} \in \mathbb{R}^{4 \times 4},
   \]
   where
   \[
   \mu_X = \begin{bmatrix} -3 \\ 2 \end{bmatrix}, \quad \mu_Y = \begin{bmatrix} 0 \\ 1 \end{bmatrix},
   \]
   and
   \[
   \Sigma_X = \begin{bmatrix} 8 & 2 \\ 2 & 5 \end{bmatrix}, \quad \Sigma_Y = \begin{bmatrix} 6 & -2 \\ -2 & 7 \end{bmatrix}, \quad \Sigma_{YX} = \Sigma_{XY} = \begin{bmatrix} 3 & 1 \\ -1 & 3 \end{bmatrix}.
   \]
   (a) Calculate the canonical correlation $\rho_1$ (the largest), $\rho_2$ (the second largest).
   (b) Find the canonical correlation variables $(U_1, V_1)$ and $(U_2, V_2)$ corresponding to $\rho_1$ and $\rho_2$.
   (c) Let $U = [U_1, U_2]^T$ and $V = [V_1, V_2]^T$. Evaluate
      \[
      E \left( \begin{bmatrix} U \\ V \end{bmatrix} \right) \quad \text{and} \quad \text{Cov} \left( \begin{bmatrix} U \\ V \end{bmatrix} \right) = \begin{bmatrix} \Sigma_U & \Sigma_{UV} \\ \Sigma_{VU} & \Sigma_V \end{bmatrix}.
      \]
   (d) Comment on the correlation structure between and within $U$ and $V$.

3. (Sample canonical correlation analysis) The data set for this problem is obtained by taking four different measures of stiffness, shock, vibrate, static1, static2, for each of $n = 30$ boards. The first measurement involves sending a shock wave down the board, the second measurement...
is determined while vibrating the board, and the last two measurements are obtained from static
tests. The squared distances \( d^2_j = (x_j - \bar{x})^T S^{-1} (x_j - \bar{x}) \) are also included as the last column in
the data matrix.
Let \( X = [X_1, X_2]^T \) be the random vector representing the dynamic measures of stiffness, and let
\( Y = [Y_1, Y_2]^T \) be the random vector representing the static measures of stiffness. Load the data
matrix \( p3.txt \) (R command: `stiff = read.table("p3.txt")`)
(a) Perform a canonical correlation analysis of these data by computing the singular value de-
composition of an appropriate matrix formed from the sample covariance matrices. You may
compare your result with that obtained by your software (if you use R, it is `cancor(X1,X2)`).
(b) Write the first canonical correlation variables \( U_1 \) and \( V_1 \) as linear combinations of `shock`,
`vibrate, static1, static2`.
(c) Produce two scatterplots of the data: one in the coordinate plane of the first canonical
correlation vectors, one in the plane of the second canonical correlation vectors.
(d) Based on the two plots and the values of the canonical correlations \( \{\rho_1, \rho_2\} \), comment on
the correlation structure captured by each canonical pair.
(e) Repeat (a) with sample correlation matrices in place of sample covariance matrices and
verify that the pairs of canonical vectors obtained are related via scaling by the sample
standard deviation matrix.

4. (Canonical correlation analysis for angular measurements) Some observations are in the form
of angles. Here we will see how to deal with such data.
(a) Consider a bivariate random vector \( X = [X_1, X_2]^T \) with a uniform distribution inside a
circle of radius 1 centered at some unknown point \( \mu = [\mu_1, \mu_2]^T \in \mathbb{R}^2 \).
Then \( E(X) = \mu \). A sample of \( n = 4 \) is taken. The observed values are
\[
\begin{bmatrix}
0.9 \\
0
\end{bmatrix}, \quad \begin{bmatrix}
0.6 \\
0.6
\end{bmatrix}, \quad \begin{bmatrix}
0.6 \\
-0.6
\end{bmatrix}, \quad \begin{bmatrix}
-0.9 \\
0
\end{bmatrix}.
\]
Compute sample mean \( \overline{x} \) and sample covariance matrix. Is \( \overline{x} \) a good estimator of \( \mu \)? Why
or why not?
(b) We consider an angular valued random variable \( \theta \), note that this can always be represented
as a random vector \( Y = [\cos \theta, \sin \theta]^T \) that takes value on the circle. Show that
\[
b^T Y = \sqrt{b_1^2 + b_2^2} \cos(\theta - \beta)
\]
where \( b_1/\sqrt{b_1^2 + b_2^2} = \cos \beta \) and \( b_2/\sqrt{b_1^2 + b_2^2} = \sin \beta \). Here \( b = [b_1, b_2]^T \in \mathbb{R}^2 \) is a constant
vector.
(c) Let \( X = X \) be a random vector with a single component, i.e., just a random variable. Here
\( X \) is not angular valued. Show that the population canonical correlation is
\[
\rho_1 = \max_{\beta} \text{Corr}(X, \cos(\theta - \beta))
\]
and that selecting the population canonical correlation variable \( V_1 \) amounts to selecting a
new ‘origin’ or ‘baseline’ \( \beta \) for the angle \( \theta \).
(d) Let \( X \) is a random variable representing ozone (\( O_3 \)) levels and \( \theta \) is a angular random variable
representing wind direction measured from the north. We make 19 observations to obtain
the sample correlation matrix

\[
R = \begin{bmatrix}
R_X & R_{X\theta} \\
R_{\theta X} & R_{\theta}
\end{bmatrix} = \begin{bmatrix}
O_3 & \cos \theta & \sin \theta \\
\cos \theta & 1.000 & 0.166 & 0.694 \\
\sin \theta & 0.166 & 1.000 & -0.051 \\
0.694 & -0.051 & 1.000 & \sin \theta
\end{bmatrix}.
\]

Find the sample canonical correlation \(\hat{\rho}_1\) and the sample canonical correlation variable \(\hat{V}_1\) representing the new origin \(\beta\).

(c) Show that if we fix \(\lambda\) and deduce that from (a) and (b) that

\[
a^\top X = \sqrt{a_1^2 + a_2^2} \cos(\phi - \alpha).
\]

Now show that

\[\rho_1 = \max_{\alpha, \beta} \text{Corr}(\cos(\phi - \alpha), \cos(\theta - \beta)).\]

(f) Let \(\phi\) and \(\theta\) be two angular random variables representing wind directions at 6:00 A.M. and at 12:00 P.M. We make 21 measurements of \(X\) and \(Y\) (related to \(\phi\) and \(\theta\) as in (b) and (d)). We obtain the sample correlation matrix

\[
R = \begin{bmatrix}
R_X & R_{XY} \\
R_{YX} & R_Y
\end{bmatrix} = \begin{bmatrix}
\cos \phi & \sin \phi & \cos \theta & \sin \theta \\
\cos \phi & 1.000 & -0.291 & 0.440 & 0.372 \\
\sin \phi & -0.291 & 1.000 & -0.205 & 0.243 \\
\cos \theta & 0.440 & -0.205 & 1.000 & 0.181 \\
\sin \theta & 0.372 & 0.243 & 0.181 & 1.000
\end{bmatrix}.
\]

Find the sample canonical correlation \(\hat{\rho}_1\) and sample canonical correlation variables \(\hat{U}_1\) and \(\hat{V}_1\).

5. (Proofs behind CCA) Let \(A \in \mathbb{R}^{p \times p}\) and \(B \in \mathbb{R}^{q \times q}\) be symmetric positive definite matrices and \(C \in \mathbb{R}^{p \times q}\). Let

\[
G = A^{-1/2}CB^{-1/2} \in \mathbb{R}^{p \times q}.
\]

We shall write \(\lambda_{\max}(M)\) for the largest eigenvalue of a matrix \(M\).

(a) Suppose \(p = q\). Show that eigenvalues of \(B^{-1}A\), \(B^{-1/2}AB^{-1/2}\), and \(AB^{-1}\) are all equal.

What are the relations between the eigenvectors?

(b) Suppose \(p = q\). Show that

\[
\max_{x \in \mathbb{R}^p} \{x^\top Ax : x^\top Bx = 1\} = \max_{y \in \mathbb{R}^q} \{y^\top B^{-1/2}AB^{-1/2}y : y^\top y = 1\}.
\]

By using Problem 7 in Homework 2, deduce that

\[
\max_{x \in \mathbb{R}^p} \{x^\top Ax : x^\top Bx = 1\} = \lambda_{\max}(B^{-1/2}AB^{-1/2}),
\]

\[
\text{argmax}_{x \in \mathbb{R}^p} \{x^\top Ax : x^\top Bx = 1\} = q_{\text{max}},
\]

where \(q_{\text{max}} \in \mathbb{R}^p\) is the eigenvector of \(B^{-1}A\) corresponding to the largest eigenvalue.

(c) Show that if we fix \(x \in \mathbb{R}^p\) and just maximize over all \(y \in \mathbb{R}^q\), then

\[
\max_{y \in \mathbb{R}^q} \{(x^\top Cy)^2 : y^\top By = 1\} = \max_{y \in \mathbb{R}^q} \{y^\top [C^\top xx^\top]C : y^\top By = 1\}
\]

and deduce that from (a) and (b) that

\[
\max_{y \in \mathbb{R}^q} \{(x^\top Cy)^2 : y^\top By = 1\} = \lambda_{\max}(B^{-1}C^\top xx^\top C).
\]
Show that the largest eigenvalue of a rank-1 matrix $ab^T$ is $b^Ta$ and deduce that
\[
\max_{y \in \mathbb{R}^q} \{ (x^TCy)^2 : y^TBy = 1 \} = x^TCB^{-1}C^Tx.
\]

(d) Using (a), (c), and Problem 7 in Homework 2, show that
\[
\max_{x \in \mathbb{R}^p, y \in \mathbb{R}^q} \{ (x^TCy)^2 : x^TAx = 1, y^TBy = 1 \} = \lambda_{\text{max}}(GG^T).
\]

(e) Let $\sigma_1, \ldots, \sigma_p \in \mathbb{R}$, $u_1, \ldots, u_p \in \mathbb{R}^p$, $v_1, \ldots, v_p \in \mathbb{R}^q$ be the singular values and left/right singular vectors of $G$. By Problem 7 in Homework 2, show that
\[
\max_{x \in \mathbb{R}^p} \{ x^TG^Tx : x^T1 = 1, u_i^Tx = 0, i = 1, \ldots, k-1 \} = \sigma_k^2,
\]
\[
\arg\max_{x \in \mathbb{R}^p} \{ x^TG^Tx : x^T1 = 1, u_i^Tx = 0, i = 1, \ldots, k-1 \} = u_k,
\]
for $k = 1, \ldots, p$. Hence deduce that
\[
\max_{x \in \mathbb{R}^p, y \in \mathbb{R}^q} \{ x^TCy : x^T1 = 1, y^TBy = 1, u_i^TAx = 0, i = 1, \ldots, k-1 \} = \sigma_k,
\]
\[
\arg\max_{x \in \mathbb{R}^p, y \in \mathbb{R}^q} \{ x^TCy : x^T1 = 1, y^TBy = 1, u_i^TAx = 0, i = 1, \ldots, k-1 \} = (A^{-1/2}u_k, B^{-1/2}v_k),
\]
for $k = 1, \ldots, p$. Finally show that
\[
\max_{x \in \mathbb{R}^p, y \in \mathbb{R}^q} \{ x^TCy : x^T1 = 1, y^TBy = 1, v_i^TAx = 0, i = 1, \ldots, k-1 \} = \sigma_k,
\]
\[
\arg\max_{x \in \mathbb{R}^p, y \in \mathbb{R}^q} \{ x^TCy : x^T1 = 1, y^TBy = 1, v_i^TAx = 0, i = 1, \ldots, k-1 \} = (A^{-1/2}u_k, B^{-1/2}v_k),
\]
for $k = 1, \ldots, p$.

6. *(Linear discriminant analysis)* The admissions committee of a business school used GPA and GMAT scores to make admission decisions. The values for the variable admit = 1, 2, 3 correspond to admission decisions of yes, no, waitlist. Label the data set p6.txt (R commands: gsbdata = read.table("p6.txt"); colnames(gsbdata)=c("GPA", "GMAT","admit")).

(a) Calculate $\mathbf{x}_i$, $\mathbf{x}$ and $S_{\text{pool}}$.

(b) Calculate the sample within groups matrix $W$, its inverse $W^{-1}$, and the sample between groups matrix $B$. Find the eigenvalues and eigenvectors of $W^{-1}B$. (R command for $A^{-1}$ is solve(A)).

(c) Use the linear discriminants derived from these eigenvectors to classify the two new observations $\mathbf{x} = [3.21, 497]^T$ and $\mathbf{x} = [3.22, 497]^T$.

(d) Scatterplot the original data set on the plane of the first two discriminants, labeled by admission decisions. Comment on the results in (c). Is this a good admission policy?