FINM 331: MULTIVARIATE DATA ANALYSIS
FALL 2018
PROBLEM SET 2

The required files for all problems can be found in:

`http://www.stat.uchicago.edu/~lekheng/courses/331/hw2/`

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. (*Basic description of multivariate data*) The data set `p1.txt` contains national track records for women, with measurements for 100m, 200m, and 400m races in seconds, and longer distance races in minutes. Variable names are not included. Compute the following for the data set and round your answers to two decimal places.

(a) Sample means. Is there any variable for which the mean is not very meaningful?
(b) Sample covariance matrix and correlation matrix.
(c) Correlation matrix of the logarithm of the data.

The following R commands (with `p1.txt` in your working directory) load the data and label the variables appropriately:

```
track = read.table("p1.txt")
colnames(track) = c("Country","100m","200m","400m","800m","1500m","3000m","Marathon")
```

2. (a) (*Effects of scales*) Let $X_1, \dots, X_p$ denote p jointly distributed positive random variables, and let $a_1, \dots, a_p \in \mathbb{R}$ be positive constants. Let $Y_i = \log(a_i X_i)$, $i = 1, \dots, p$. Show that the covariance matrix of Y does not depend on the a_i's.

(b) (*Positive semidefiniteness of covariance matrix*) Show that the covariance matrix $\Sigma = \operatorname{Cov}(\mathbf{X}) \in \mathbb{R}^{p \times p}$ of a random vector $\mathbf{X} = [X_1, \dots, X_p]^{\mathsf{T}}$ must have all eigenvalues nonnegative.

(c) (*Expectation of random matrix*) Let $A \in \mathbb{R}^{m \times p}$ and $B \in \mathbb{R}^{p \times n}$ be constant matrices. Let $\mathbf{C} = A\mathbf{X}B$ where $\mathbf{X}$ is a *random matrix*, i.e., a $p \times p$ matrix whose (i, j)th entries are random variables X_{ij}, $i, j = 1, \dots, p$. Write down the (i, j)th entry of the matrix $\mathbf{C}$. Show that

$$E(\mathbf{C}) = AE(\mathbf{X})B \in \mathbb{R}^{m \times n}.$$

3. (*Population* PCA) The trivariate random vector $\mathbf{X} = [X_1, X_2, X_3]^{\mathsf{T}}$ has covariance matrix

$$\operatorname{Cov}(\mathbf{X}) = \Sigma = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 5 & 0 \\ 1 & 0 & 3 \end{bmatrix}.$$

(a) What are the eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$ of Σ? The following R commands define the matrix and compute the eigenvalues/vectors:

```
sigma = matrix(c(2,-1,1,-1,5,0,1,0,3),3,3)
eigen(sigma)
```

(b) Write down the population principal components Y_i in terms of X_1, X_2, X_3 for $i = 1, 2, 3$.
(c) Derive the value of $\operatorname{Var}(Y_1)$ using the properties we saw in the lectures:

$$\operatorname{Var}\left(\sum\nolimits_{i=1}^{n} X_i\right) = \sum\nolimits_{i=1}^{n} \operatorname{Var}(X_i) + 2\sum\nolimits_{i<j} \operatorname{Cov}(X_i, X_j) \quad \text{and} \quad \operatorname{Var}(aX_i) = a^2 \operatorname{Var}(X_i)$$

Date: October 22, 2018 (Version 1.0); due: November 5, 2018.

for any $a \in \mathbb{R}$. Then compare $\mathrm{Var}(Y_1)$ to the λ_i's.

4. (*Sample* PCA) Use the same data `p1.txt` from Problem **1**. Exclude the nominal variable "`Country`" in numerical calculations.

(a) Obtain the eigenvalues and eigenvectors for the sample correlation matrix R. What is the sum of all eigenvalues? Compare it to the dimensions of the data.

(b) Determine the first two principal components for the standardized (variance = 1) variables.

(i) Compare the two principal components PC1 and PC1 with the eigenvectors in (a). Comment.

(ii) What are the percentages of total (standardized) sample variation explained by the first and second principal components?

(c) Every observation has its coordinates in the space of principle components (PC1, ..., PC7).

(i) Construct a two-dimensional scatterplot of the 54 observations in the (PC1, PC2) plane.

(ii) Rank the nations based on their scores on the first principal component. List the top six and the last three countries. In your opinion, does the ranking correspond with athletic excellence for the counties?

5. (*Scaling effects in sample* PCA) `p5.txt` is the air quality data set we discussed in Slides 2. We will perform sample PCA using both the original data as well as the standardized data (i.e., variable variance = 1). The measurements are on air pollution variables recorded at noon in 42 different days.

(a) In each of the two cases, how many principal components are needed to effectively summarize at least 90% of the variability in the data?

(b) Plot two scree plots, one from PCA based on the original data, one based on the standardized data.

(c) Compare and comment, based on your results in (a) and (b).

The following R commands (with `p5.txt` in your working directory) load the data and label the variables appropriately:

```
air = read.table("p5.txt")
colnames(air) = c("Wind", "Radiation", "CO", "NO", "NO2", "O3", "HC")
```

6. This is the demographics data that we saw in the first lecture. The data matrix in `p6-data.txt` is a 49×7 matrix where each row is indexed by a country and each column is indexed by a demographic variable. So for example, if we denote the matrix by $A = [a_{ij}] \in \mathbb{R}^{49\times 7}$, then $a_{23} = 84$ is Austria's population per square kilometers (row index 2 = Austria, column index 3 = population per square kilometers). The row and column labels are reproduced in separate files for your convenience: `p6-row.txt`, `p6-column.txt` (you may not need these, depending on the program you use).

(a) (*Standardizing*) Write a program to (i) mean center the data matrix and then (ii) scale it by standard deviation . Denote the standardized data matrix by $\widehat{A}$.

(b) (*Singular value decomposition*) Find the first two right singular vectors of $\widehat{A}$, $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^7$. Project the data onto the two-dimensional space spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$. Plot this in a PCA scatter plot where the x- and y-axes correspond to $\mathbf{v}_1$ and $\mathbf{v}_2$ respectively and where the points correspond to the countries — label each point by the country it corresponds to. Identify the two obvious outliers.

(c) (*Singular value decomposition*) Now do the same with the two left singular vectors of $\widehat{A}$, $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^{49}$, i.e., project the data onto the two-dimensional space spanned by $\mathbf{u}_1$ and $\mathbf{u}_2$ and plot this in a graph as before. Note that in this case, the points correspond to the demographic variables — label them accordingly.

(d) (*Biplot*) Overlay the two graphs in (a) and (b) in a biplot. Identify the two demographic variables near the two outlier countries — these explain why the two countries are outliers.

(e) (*Outlier removal*) Remove the two outlier countries and redo (a) with this 47×7 matrix. This allows you to see features that were earlier obscured by the outliers. Which two European countries are most alike Japan?

(f) (*Biplot*) Note that the reason we didn't need to adjust the scale of the axes using the singular values of $\widehat{A}$ in our biplot because the standardization has taken care of the scaling. Suppose we only mean center but did not scale our matrix by standard deviation, show how we would perform the biplot.

(g) (*Effect of mean centering*) Let's suppose now that we neither mean center nor scale A by standard deviation, i.e., we use the raw data A instead of $\widehat{A}$. Repeat what we did in (b), (c), (d). Discuss your results.

7. (*Proofs behind* PCA) Let $A \in \mathbb{R}^{p\times p}$ be a symmetric matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and corresponding eigenvectors $Q = [\mathbf{q}_1, \mathbf{q}_2, \dots \mathbf{q}_p] \in \mathbb{R}^{p\times p}$.

(a) By considering the EVD of A, show that for any unit vector $\mathbf{x} = [x_1, \dots, x_p]^\mathsf{T} \in \mathbb{R}^p$,

$$\mathbf{x}^\mathsf{T} A \mathbf{x} = \lambda_1 y_1^2 + \cdots + \lambda_p y_p^2$$

for some unit vector $\mathbf{y} = [y_1, \dots, y_p]^\mathsf{T} \in \mathbb{R}^p$. Here unit vector means that $\|\mathbf{x}\|_2 = 1$.

(b) Using (a), show that

$$\max\{\mathbf{x}^\mathsf{T} A \mathbf{x} : \|\mathbf{x}\|_2 = 1\} = \lambda_1,$$
$$\min\{\mathbf{x}^\mathsf{T} A \mathbf{x} : \|\mathbf{x}\|_2 = 1\} = \lambda_p,$$

and that

$$\operatorname{argmax}\{\mathbf{x}^\mathsf{T} A \mathbf{x} : \|\mathbf{x}\|_2 = 1\} = \mathbf{q}_1,$$
$$\operatorname{argmin}\{\mathbf{x}^\mathsf{T} A \mathbf{x} : \|\mathbf{x}\|_2 = 1\} = \mathbf{q}_p.$$

(c) Suppose $\mathbf{x} = [x_1, \dots, x_p]^\mathsf{T} \in \mathbb{R}^p$ is a unit vector orthogonal to $\mathbf{q}_1 \in \mathbb{R}^p$, i.e., $\mathbf{x}^\mathsf{T}\mathbf{q}_1 = 0$, show that

$$\mathbf{x} = a_2 \mathbf{q}_2 + \cdots + a_p \mathbf{q}_p$$

and that $a_2^2 + \cdots + a_p^2 = 1$. Hence deduce that

$$\mathbf{x}^\mathsf{T} A \mathbf{x} = \lambda_2 y_2^2 + \cdots + \lambda_p y_p^2$$

for some unit vector $\mathbf{y} = [y_1, \dots, y_p]^\mathsf{T} \in \mathbb{R}^p$.

(d) Using (c), show that

$$\max\{\mathbf{x}^\mathsf{T} A \mathbf{x} : \|\mathbf{x}\|_2 = 1,\ \mathbf{x}^\mathsf{T}\mathbf{q}_1 = 0\} = \lambda_2,$$
$$\operatorname{argmax}\{\mathbf{x}^\mathsf{T} A \mathbf{x} : \|\mathbf{x}\|_2 = 1,\ \mathbf{x}^\mathsf{T}\mathbf{q}_1 = 0\} = \mathbf{q}_2.$$

(e) Generalize (d) and show that

$$\max\{\mathbf{x}^\mathsf{T} A \mathbf{x} : \|\mathbf{x}\|_2 = 1,\ \mathbf{x}^\mathsf{T}\mathbf{q}_1 = \cdots = \mathbf{x}^\mathsf{T}\mathbf{q}_{k-1} = 0\} = \lambda_k,$$
$$\operatorname{argmax}\{\mathbf{x}^\mathsf{T} A \mathbf{x} : \|\mathbf{x}\|_2 = 1,\ \mathbf{x}^\mathsf{T}\mathbf{q}_1 = \cdots = \mathbf{x}^\mathsf{T}\mathbf{q}_{k-1} = 0\} = \mathbf{q}_k$$

for $k = 2, \dots, p$.

(f) Using Problem **2**(b) and the previous parts, show that for a random vector $\mathbf{X} = [X_1, \dots, X_p]^\mathsf{T}$,

$$\mathbf{q}_k = \operatorname{argmax}\{\operatorname{Var}(\mathbf{a}^\mathsf{T}\mathbf{X}) : \|\mathbf{a}\|_2 = 1,\ \operatorname{Cov}(\mathbf{a}^\mathsf{T}\mathbf{X}, \mathbf{q}_1^\mathsf{T}\mathbf{X}) = \cdots = \operatorname{Cov}(\mathbf{a}^\mathsf{T}\mathbf{X}, \mathbf{q}_{k-1}^\mathsf{T}\mathbf{X}) = 0\},$$

where $\mathbf{q}_k$ is the kth eigenvector of $\Sigma = \operatorname{Cov}(\mathbf{X})$.