ST 557: Applied Multivariate Analysis
Fall 2019

Homework 1

Due: Friday October 11

1. Suppose variables X and Y are measured on a sample of 20 units, producing the following data (data file 'HW1-1.csv' available on Canvas):

	X	Y
1	-1.54	0.46
2	-4.25	-1.23
3	-0.85	0.34
4	-2.90	-0.94
5	-1.09	-0.84
6	-5.92	-0.70
7	3.51	2.92
8	0.09	0.76
9	-2.08	-0.99
10	5.01	1.58
11	3.22	-0.43
12	3.67	1.29
13	-1.61	-1.60
14	2.23	1.90
15	0.20	-0.06
16	4.39	1.50
17	1.15	-1.81
18	3.63	0.99
19	-4.32	-0.72
20	1.42	0.82

   (a) Read this data into R using `read.csv()`. Create a 2-dimensional scatter plot of the 20 observations (use `plot()` function in R).

   (b) Find the sample mean vector, and add this point to the plot using `points()`. You can make this point a different color using the `col=` argument, or you can make it a different plotting character using the `pch=` argument. For example:

   ```
   > points(sampMean[1], sampMean[2], pch=16, col=2)
   ```

   (c) Find the sample covariance matrix.

   (d) Find the eigendecomposition (spectral decomposition) of the sample covariance matrix using `eigen()`.

   (e) Add the eigenvector corresponding to the largest eigenvalue to the plot as a vector from the sample mean using `lines()`. Be careful here: if the first eigenvector is (v_1, v_2) and the sample mean vector is $(\bar{x}_1, \bar{x}_2)$, you want a line from $(\bar{x}_1, \bar{x}_2)$ to $(\bar{x}_1 + v_1, \bar{x}_2 + v_2)$.
   Describe how the direction of this eigenvector relates to the cloud of data points.

2. (Same as previous problem; different data.) Suppose variables X and Y are measured on a sample of 20 units, producing the following data (data file 'HW1-2.csv' available on Canvas):

	X	Y
1	-1.54	1.96
2	-4.25	4.17
3	-0.85	1.13
4	-2.90	2.79
5	-1.09	0.80
6	-5.92	6.31
7	3.51	-2.49
8	0.09	0.28
9	-2.08	1.85
10	5.01	-4.85
11	3.22	-3.84
12	3.67	-3.48
13	-1.61	1.01
14	2.23	-1.56
15	0.20	-0.26
16	4.39	-4.19
17	1.15	-2.20
18	3.63	-3.59
19	-4.32	4.50
20	1.42	-1.19

   (a) Read this data into R using `read.csv()`. Create a 2-dimensional scatter plot of the 20 observations (use `plot()` function in R).

   (b) Find the sample mean vector, and add this point to the plot using `points()`.

   (c) Find the sample covariance matrix.

   (d) Find the eigendecomposition (spectral decomposition) of the sample covariance matrix.

   (e) Add the eigenvector corresponding to the largest eigenvalue to the plot as a vector from the sample mean using `lines()`. Describe how the direction of this eigenvector relates to the cloud of data points.

3. Let $\mathbf{A}$ be the following matrix:

$$\mathbf{A} = \begin{bmatrix} 5.125 & 3.875 & 2.125 & -1.125 & 0.000 \\ 3.875 & 5.125 & -1.125 & 2.125 & 0.000 \\ 2.125 & -1.125 & 5.125 & 3.875 & 0.000 \\ -1.125 & 2.125 & 3.875 & 5.125 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & -3.000 \end{bmatrix}$$

   (a) Find the eigendecomposition (spectral decomposition) of $\mathbf{A}$ (use the `eigen()` function in R).

   (b) Is $\mathbf{A}$ positive definite? If not, give a vector $\mathbf{x}$ for which $\mathbf{x}^T\mathbf{A}\mathbf{x} < 0$. (Hint: convince yourself that for any eigenvector $\mathbf{v}$ of $\mathbf{A}$ with corresponding eigenvalue λ, we have

$$\mathbf{v}^T\mathbf{A}\mathbf{v} = \lambda\mathbf{v}^T\mathbf{v} = \lambda\|\mathbf{v}\|.$$

   Note that the Euclidean norm $\|\mathbf{v}\|$ is always positive—why?)

   (c) Let a vector $\mathbf{x}$ be given by

$$\mathbf{x} = 4\mathbf{v}_1 + 2\mathbf{v}_5$$

where $\mathbf{v}_1$ is the eigenvector corresponding to the largest eigenvalue λ_1 of $\mathbf{A}$, and $\mathbf{v}_5$ is the eigenvector corresponding to the smallest eigenvalue λ_5 of $\mathbf{A}$. In terms of $\mathbf{v}_1, \mathbf{v}_5, \lambda_1$, and λ_5, what is $\mathbf{Ax}$? (You do not need to give a numerical answer here, just a symbolic answer.)

4. The Iris Flower Data Set is available as the file 'IrisData.csv' on Canvas. Download the file and read it into R using `read.csv()`.

   (a) Find the sample mean vector for this dataset.

   (b) Find the sample mean vector for each species (Species 1, 2, and 3) separately.

   (c) Find the sample correlation matrix for this dataset, using the `cor()` function in R. Which two variables are most highly correlated?

   (d) Find the individual sample correlation matrices for each species (Species 1, 2, and 3) separately. Are the same two variables most highly correlated for all three species?

   (e) The function `pairs()` in R produces all scatter plots between every combination of two variables in a dataset. Produce the pairs scatter plot of this entire dataset, using the `pairs()` function. Color the points to correspond to the species using the `col=` option in the `pairs()` function. Describe what you see: are there distinct groups of flowers? How well do you think you could predict the species based on knowing the Petal Width? If you could only use two of these four variables to distinguish between the species, which two would you want to use, and why?

5. Let $\mathbf{A}$ be any $(n \times p)$ matrix, for arbitrary dimensions n and p, and let $\mathbf{B}$ be the product matrix $\mathbf{B} = \mathbf{A}^T\mathbf{A}$.

   (a) Show that $\mathbf{B}$ is a symmetric matrix.

   (b) Show that $\mathbf{B}$ is a positive semi-definite matrix: $\mathbf{x}^T\mathbf{B}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^p$ (that is, for all p-dimensional vectors $\mathbf{x}$.)

   (c) Let $\mathbf{X}$ be any $(n \times p)$ matrix. In matrix notation, the sample covariance matrix $\mathbf{S}$ is given by

$$\mathbf{S} = \frac{1}{n-1}\left(\mathbf{X} - \bar{\mathbf{X}}\right)^T\left(\mathbf{X} - \bar{\mathbf{X}}\right)$$

   where $\bar{\mathbf{X}}$ is an $(n \times p)$ matrix with n identical rows equal to the sample mean vector $(\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2, \ldots, \bar{\mathbf{X}}_p)$. Use this fact to argue that the sample covariance matrix must be positive semi-definite.