1. Read the data contained in homeless_smalldata.csv in your statistical software. We wish to perform a comparison of non-homeless and homeless subjects in the dataset, comparing individual observations on the PCS (physical component score) and MCS (mental component scores) and the CESD measure of depressive symptoms. The homeless variable in the dataset is 0 if the subject is non-homeless, and is 1 if homeless. Assume that for each sub-population of the non-homeless and the homeless, the three variables PCS, MCS and CESD form the trivariate normal distribution $N_3(\mu_1, \Sigma)$ and $N_3(\mu_2, \Sigma)$, respectively.

(a) State the null and alternative hypotheses relevant to the problem.

(b) Refer to the case IV, page 5, Lecture note 3, determine the dimension $p$ and the sample sizes $n_1$, $n_2$.

(c) Conduct the test. Report your (observed) test statistic, the theoretical null distribution, and the p-value.

(d) Plot the data and argue that the assumptions of the test are reasonable.

2. Let $X_i$ be i.i.d. $N_2(\mu, \Sigma)$. Given observations $x_i$, $i = 1, \ldots, 42$, we have sufficient statistics $\bar{x} = (.564, .603)'$ (sample mean) and $S = \begin{bmatrix} .0144 & .0117 \\ .0117 & .0146 \end{bmatrix}$, (sample variance). Now consider constructing a confidence region of size $\alpha$ for $\mu = (\mu_1, \mu_2)'$.

(a) What is the MLE for $\Sigma$?

(b) Evaluate the expression for 95% elliptical confidence region for $\mu$. Denote this region as $R_1$.

(c) Is $\mu_1 = (.60, .58)'$ in $R_1$?

(d) Evaluate the simultaneous confidence intervals for $\mu_1$ and $\mu_2$ (cf. Proposition 2, lecture 3). Denote this region as $R_2$. This notation is used in problem 3 below.

(e) Conduct a hypothesis test for

$$H_0 : \mu = (.60, .58) \; \text{vs} \; H_1 : \text{not } H_0.$$ 

Report your (observed) test statistic, the theoretical null distribution, and the p-value.

3. Extra credit question for undergraduate students. Graduate students are required to do this problem.

Bonferroni method for multiple comparison. An alternative to the simultaneous confidence interval is the Bonferroni method for multiple comparison. In problem 1, since there are only two parameters ($\mu_1, \mu_2$) involved, we will see that Bonferroni method is advantageous compared to the simultaneous confidence intervals.

In general, for $m$ linear combinations $a_1^t \mu, \ldots, a_m^t \mu$, denote $C_i$ be the confidence interval for the $i$th linear combination, $a_i^t \mu$ with confidence level $\alpha_i$. We have

$$P(a_i^t \mu \in C_i, \; \text{for all } i) = 1 - P( \text{at least for one } i, a_i^t \mu \notin C_i)$$

$$\geq 1 - \sum_{i=1}^m P(a_i^t \mu \notin C_i)$$

$$= 1 - (\alpha_1 + \cdots + \alpha_m). \quad (1)$$

(a) Show that the interval $C_i(\alpha_i) = a_i^t \bar{x} \pm t_{n-1}(\alpha_i/2)\sqrt{a_i^t Sa a_i}/n$ is a confidence interval of $a_i^t \mu$ with confidence level $\alpha_i$.

(b) Verify the inequality (1), for $m = 3$. [Hint: You may use Venn Diagrams for related events.]
(c) Verify the following:

\[ P(\mu_i^* \in C_i(\alpha/m), \text{ for all } i) \geq 1 - \alpha. \]

This gives a Bonferroni type simultaneous confidence intervals \( C_i(\alpha/m) \) for level \( \alpha \).

(d) Using the data in problem 2, evaluate the simultaneous confidence intervals for \( \mu_1 \) and \( \mu_2 \) using the Bonferroni method. Denote this region as \( R_3 \).

(e) Sketch a graph for the regions \( R_1, R_2, R_3 \) in \( x-y \) coordinates. Which one do you prefer and why?

4. Given multivariate data \( \mathbf{X} = [x_1, \ldots, x_n] \), denote the centered observations \( \tilde{x}_i = x_i - \bar{x} \). Show that the criterion for the first sample principal component direction \( \mathbf{u} \), maximization of the sample variance of \( \mathbf{u}^\top \tilde{x}_i \), is equivalent to minimization of the residual sum of squares. We define the \( i \)th residual \( r_i \) as

\[ ||\tilde{x}_i - \frac{\mathbf{u}^\top \tilde{x}_i}{\mathbf{u}^\top \mathbf{u}}\mathbf{u}||. \]

5. PCA for \textit{pendigit} data. The dataset is from UC Irvine Machine Learning Repository at http://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits. Read the data set description at the webpage linked above and also Section 7.2.10 of Izenman. We take a subset of data contained in \textit{pendigit3.txt} (download this file at the course webpage). The data set contains \( n = 1055 \) observations of \( p = 16 \) variables, with last column of the dataset, containing '3's, represents that the observation is a writing of a digit 3. The 16 variables are equispaced locations of pen at 8 timepoints, and are arranged as \((x_1, y_1), (x_2, y_2), \ldots, (x_8, y_8)\).

(a) Visualize the first observation by plotting \( x \) against \( y \). You should see the digit 3.

(b) Perform PCA for this dataset and report the result by 1) scatterplot matrix of principal components and 2) scree plot.

(c) Do the data look as if they are from a MVN distribution?

(d) How many components will you keep? Give a justification.

(e) Walk along each of the first four principal components. Explain the modes of variation captured by each component.

(f) Now open the data in \textit{pendigit8.txt}. Combine two datasets into one, forming a data matrix with \( N = 2110 \) observations. Perform PCA on this dataset and report the result by scatterplot matrix of principal components. Do you see a nice separation of observations corresponding to the digit 3 and digit 8?