High Dimensional Statistical Analysis

$\boxed{\text{Assignment 1}}$

Vector and Matrix Algebra, Multivariate Normal Distribution

Exercises

Problem 1 Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}$$

. Answer the following questions

1. Is $\mathbf{A}$ symmetric?
2. Perform the spectral decomposition of A.
3. One way of writing the spectral decomposition of $\mathbf{A}$ is

$$\lambda_1 \mathbf{e}_1 \mathbf{e}_1^T + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^T.$$

   Identify each matrix in the representation above.

4. Use the spectral decomposition of $\mathbf{A}$ given above and find $\sqrt{\mathbf{A}}$. Check that the matrix you found satisfies

$$\sqrt{\mathbf{A}}\sqrt{\mathbf{A}} = \mathbf{A}.$$

Problem 2 Consider the spectral decomposition of a positive definite matrix as given in Lecture 1:

$$\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T.$$

The columns of $\mathbf{P}$ are made of eigenvectors $\mathbf{e}_i$, $i = 1, \dots, n$ and they are orthonormalized, i.e. their lengths are one and they are orthogonal (peripendicular) one to another. The diagonal matrix $\mathbf{\Lambda}$ has the corresponding (positive) eigenvalues on the diagonal. Provide argument for the following

1. $\mathbf{P}^T = \mathbf{P}^{-1}$
2. Determinant of $\mathbf{\Lambda}$ is equal to the product of the terms on the diagonal.
3. Dederminant of $\mathbf{A}$ is the same as that of $\mathbf{\Lambda}$.
4. Find the inverse matrix to $\mathbf{\Lambda}$, i.e. $\mathbf{\Lambda}^{-1}$.

5. A simple way to determine the inverse of a matrix $\mathbf{A}$ from its spectral decomposition is through

$$\mathbf{A}^{-1} = \mathbf{P}\mathbf{\Lambda}^{-1}\mathbf{P}^T.$$

   Verify that the right hand side of the above indeed define the inverse of $\mathbf{A}$.

6. Check all these statements on the little example of Problem 1.

Problem 3 In a medical study, length L and weight W of newborn children is considered. It was assumed that (L, W) will be modeled through a bivariate normal distribution. The following information has been known: the mean weight is 3343[g], with the standard deviation of 528[g], while the mean length is 49.8[cm], with the standard deviation of 2.5[cm]. Additionally the correlation between the length and the weight has been established and equal to 0.75. The joint distribution of (W, L) is bivariate normal, i.e. $(W, L) \sim N(\boldsymbol{\mu}, \Sigma)$. Perform the following tasks and answer the questions:

1. Write explicitly the parameters $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$.
2. Write explicitly the density of the joint distribution.
3. Find eigenvalues and eigenvectors of the covariance matrix $\mathbf{\Sigma}$. Sketch few elipses corresponding to the constant density contours of the joint distributions. Mark on the plot the eigenvectors scaled by the square roots of the corresponding eigenvalues and comment.
4. How many parameters characterize a bivariate normal distribution? How many parameters characterize a p-dimensional normal distribution?
5. What is the distribution of L? Give its name and parameters.
6. Suppose that the hospital records of a new-born child was lost. Give a best guess for the value of his/her length. Provide with accuracy bounds of your 'educated' guess based on the 3-σ rule.

Problem 4 In the setup of the previous problem, assume that it was reported by the mother of the child that weight was 4025[g].

1. What is the distribution of L given this additional information? Give its name and parameters.
2. Improve your previous guess and provide with accuracy limits.
3. Compare the answers from this and previous problems and comment how additional information affected the prediction value and accuracy.

Problem 5 Let $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ be independent $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ random vectors of a dimension p.

1. Find the distribution of each of the following vectors:

$$\mathbf{V}_1 = \frac{1}{4}\mathbf{X}_1 - \frac{1}{2}\mathbf{X}_2 + \frac{1}{4}\mathbf{X}_3$$
$$\mathbf{V}_2 = \frac{1}{4}\mathbf{X}_1 - \frac{1}{2}\mathbf{X}_2 - \frac{1}{4}\mathbf{X}_3$$

2. Find the joint distribution of the above vectors.

Project 1: Weight and length of newborn children

Health services and health insurance companies are interested in determining what kind of medical examinations and diagnostic procedures should be administered to a newborn child. In one approach, there is a score system based on which it is determined when a child is healthy and does not require any special attention or when he/she is not in which case a series additional medical tests are performed.

Weight and length of a newborn child are most standard indicators of the health of a child. In order to decide on the score the following procedure is considered. If the weight and length fall outside 95% of the typical values for the population, the score of zero is given. If the measurements are falling in the category between 75% and 95% the score is one. In all other cases the score of two is assigned.

A random sample of records for 736 recently born children (singleton and not prematurely born) has been considered from hospital across a certain region. The records contain a large variety of information but extraction of weight and height data are given in the file *WeightHeight.txt*.

Part One

1. Using the date estimate the mean and the covariance for the length and the weight of children.

2. Verify graphically normal distribution of the data. Use a scatterplot and qq-plots for the marginal distributions.

3. Find the ellipsoids that would serve classification regions for scores as described above.

4. How many children would score zero, one, and two, respectively? Illustrate this classification on the graphs.

5. Find the spectral decomposition of the estimated covariance matrix.

6. Plot the data transformed according to $\mathbf{P}^T\mathbf{X}$, where $\mathbf{P}$ is the matrix made of the eigenvectors standing as the columns. Interpret the transformed data.

Part Two

Additionally to weight and length of a child, also the height of parents is included in the records. In order to tune the procedure of scoring the height of parents can be also used. The *ParentsWeightLength.txt* file contains this information.

1. Using the data estimate the mean and the covariance for all four variables .
2. Verify graphically the normal distribution of the data. Use scatterplots and qq-plots for the marginal distributions.
3. Identify the conditional distribution of the weight and length of a child given the heights of parents. Find an estimate of the covariance matrix of the conditional distribution and compare it with the original unconditional covariance.
4. How the ellipsoids based on the conditional distribution will look like?
5. How many children would score zero, one, and two, respectively? Illustrate this classification on the graph and compare with the one obtained without considering the heights of parents.
6. Suppose that the father of a child is 185[cm] tall and mother is 178[cm] tall. Plot the classification ellipsoids for their child.
7. Find spectral decomposition of the estimated covariance matrix for the complete set of the data.
8. Transforme the data the according to $\mathbf{P}^T\mathbf{X}$. Plot scatter plots of the transformed data.