Homework 1

Due Feb 4 at 6 pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages).

1. (Three random variables) If a random variable $\tilde{w}$ is positively correlated with another random variable $\tilde{y}$, and $\tilde{y}$ is positively correlated with a third random variable $\tilde{z}$, can $\tilde{w}$ and $\tilde{z}$ be negatively correlated? If no, prove it. If yes,

   (a) provide an example of three such random variables;

   (b) provide a small sample of data with three variables such that the three sample correlations have this property. Compute and provide the data and the sample correlation matrix of your data.

2. (Averaging noisy data) We want to approximate a signal represented by a zero-mean random variable $\tilde{x}$ with unit variance. We have access to n measurements $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n$, where $\tilde{y}_i := \tilde{x} + \tilde{z}_i$ for $1 \leq i \leq n$. Each $\tilde{z}_i$ is a zero-mean random variable with variance σ^2. The random variables $\tilde{x}, \tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n$ are all mutually independent. We decide to approximate $\tilde{x}$ by scaling the sum of all measurements: the estimator is $\alpha \sum_{i=1}^{n} \tilde{y}_i$ for some $\alpha \in \mathbb{R}$.

   (a) What value of α minimizes the mean squared error?

   (b) What does the estimator tend to as $\sigma^2 \to 0$ and $\sigma^2 \to \infty$?

   (c) What is the mean squared error of the estimator? How does it scale with n?

3. (Interference) We model a signal of interest as a random variable $\tilde{a}$ with mean μ and variance σ^2, which is known to be nonnegative. The signal cannot be observed directly. The available measurement is modeled as a random variable $\tilde{y}$ which equals $\tilde{w}\tilde{a}$, where $\tilde{w}$ is an interfering signal that is equal to -1 with probability 1/2 and 1 with probability 1/2. We assume that $\tilde{w}$ and $\tilde{a}$ are independent.

   (a) What is the linear MMSE estimator of $\tilde{a}$ given $\tilde{y} = y$?

   (b) What is the MSE of the linear MMSE estimator?

   (c) Propose a nonlinear estimator that has a better MSE than the linear MMSE.

4. Using the formulas for regression coefficients in simple linear regression, show that the least squares line always passes through the point $(\bar{X}, \bar{Y})$.

5. (Temperature) The table in `tempature.csv` includes the average temperatures of each month at different locations (longitude and latitude).

   (a) Create a scatterplot of latitude and longitude of the 2000 locations (using a natural choice of which variable to assign to the x-axis.) Calculate the minimum and maximum latitudes in the dataset and make a plot of the marginal distribution of latitude. Does the distribution of locations appear to be a random sample of earth's spherical surface?

   (b) Make scatterplots of the latitude and temperature data and compute the correlation separately for January and July. Is there a natural justification to choose one variable for the x-axis and the other for the y-axis? Why?

   (c) Describe what you find both within and between each scatterplot and compare the two correlations. Hypothesize on what could explain any noteworthy differences found.

   (d) Compute the correlation between latitude and temperature in July for the locations in the southern hemisphere (latitude < 0). Plot scatterplots and the linear MMSE estimator and corresponding residuals.

   (e) Repeat the experiment for locations in the northern hemisphere (latitude > 0) in January.

   (f) Explain your findings in the previous part. What do you notice? Hypothesize some explanations.