Homework 6

Due October 25 at 11 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.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. 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). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting code.

1. (Correlation coefficient.)

   For any 2×2 symmetric matrix

   $$M := \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \tag{1}$$

   such that $b \neq 0$, the eigenvalues of M equal

   $$\lambda_1 = \frac{a + c + d}{2}, \tag{2}$$

   $$\lambda_2 = \frac{a + c - d}{2}, \tag{3}$$

   where $d := \sqrt{a^2 + 4b^2 + c^2 - 2ac}$. The eigenvectors equal

   $$u_1 = \begin{bmatrix} \frac{a-c+d}{2b} \\ 1 \end{bmatrix}, \quad u_2 = \begin{bmatrix} \frac{a-c-d}{2b} \\ 1 \end{bmatrix}. \tag{4}$$

   Note that the eigenvectors are not normalized for simplicity. Use this to prove that for any zero-mean random variables $\tilde{a}$ and $\tilde{b}$ if $\rho_{\tilde{a},\tilde{b}} = 1$ then

   $$\mathrm{P}\left(\tilde{b} = \frac{\sigma_{\tilde{b}}}{\sigma_{\tilde{a}}}\tilde{a}\right) = 1. \tag{5}$$

2. (Financial data) In this exercise you will use the code in the findata folder. For the data loading code to work properly, make sure you have the pandas Python package installed on your system.

   Throughout, we will be using the data obtained by calling `load_data()` in `findata_tools.py`. This will give you the names, and closing prices for a set of 18 stocks over a period of 433 days ordered chronologically. For a fixed stock (such as msft), let $P_1, \ldots, P_{433}$ denote its

sequence of closing prices ordered in time. For that stock, define the daily returns series $R_i := P_{i+1} - P_i$ for $i = 1, \ldots, 432$. Throughout we think of the daily stock returns as features, and each day (but the last) as a separate datapoint in $\mathbb{R}^{18}$. That is, we have 432 datapoints each having 18 features.

(a) Looking at the first two principal directions of the centered data, give the two stocks with the largest coefficients (in absolute value) in each direction. Give a hypothesis why these two stocks have the largest coefficients, and confirm your hypothesis using the data. The file `findata_tools.py` has `pretty_print()` functions that can help you output your results. You are not required to include the principal directions in your submission.

(b) Standardize the centered data so that each stock (feature) has variance 1 and compute the first 2 principal directions. This is equivalent to computing the principal directions of the correlation matrix (the previous part used the covariance matrix). Using the information in the comments of *generate_findata.py* as a guide to the stocks, give an English interpretation of the first 2 principal directions computed here. You are not required to include the principal directions in your submission.

(c) Assume the stock returns each day are drawn independently from a multivariate distribution $\tilde{x}$ where $\tilde{x}[i]$ corresponds to the ith stock. Assume further that you hold a portfolio with 200 shares of each of appl, amzn, msft, and goog, and 100 shares of each of the remaining 14 stocks in the dataset. Using the sample covariance matrix as an estimator for the true covariance of $\tilde{x}$, approximate the standard deviation of your 1 day portfolio returns $\tilde{y}$ (this is a measure of the risk of your portfolio). Here $\tilde{y}$ is given by

$$\tilde{y} := \sum_{i=1}^{18} \alpha[i]\tilde{x}[i],$$

where $\alpha[i]$ is the number of shares you hold of stock i.

(d) Assume further that $\tilde{x}$ from the previous part has a multivariate Gaussian distribution. Compute the probability of losing 1000 or more dollars in a single day. That is, compute

$$\Pr(\tilde{y} \leq -1000).$$

Note: The assumptions made in the previous parts are often invalid and can lead to inaccurate risk calculations in real financial situations.

3. (Streaks) In this problem we consider the problem of testing whether a randomly generated sequence is truly random. A certain computer program is supposed to generate Bernoulli iid sequences with parameter 0.5. When you try it out, you are surprised that it contains long streaks of 1s. In particular, you generate a sequence of length 200, which turns out to contain a sequence of 8 ones in a row.

(a) Let $\tilde{s}$ be equal to the longest streak of 1s in an iid Bernoulli sequence of length 5. Compute the pmf of $\tilde{s}$ exactly.

(b) Complete the script *streaks.py* to estimate the pmf of $\tilde{s}$ using Monte Carlo simulation. Compare it to your answer in the previous question. The script will also apply your code to estimate the pmf of $\tilde{s}$ when the Bernoulli iid sequence has length 200. Include your code in the answer as well as the figures generated by the script.

(c) Approximate the probability that the longest streak of ones in a Bernoulli iid sequence of length 200 has length 8 or more. Is the sequence of 8 ones evidence that the program may not be generating truly random sequences?

4. (Radioactive sample) Consider the following experiment. We have a radioactive sample situated at unit distance from a line of sensors. Each time a sensor detects a particle emitted from the sample we obtain a reading of the position of the sensor in the x axis (we assume that we have so many sensors that you can model this position as a continuous random variable). We model the measurements as an i.i.d. sequence distributed as a random variable $\tilde{m} = c + \tilde{x}$ where the pdf of $\tilde{x}$ is symmetric around the origin, that is $f_{\tilde{x}}(x) = f_{\tilde{x}}(-x)$ for all real numbers x. Your task is to estimate the position of the sample c from these data.

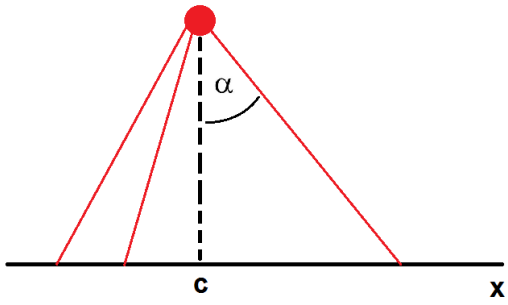

Figure 1: Diagram of the experiment.

(a) The file *radioactive_sample_1.txt* contains a vector of measurements $m_1, m_2, \ldots$. Plot a moving average of the measurements $\frac{1}{n}\sum_{i=1}^{n} m_i$ for $n = 1, 2, 3, \ldots$ (and submit the plot). Use the plot to give an estimate for the value of c. (Hint: What is the expected value of $\tilde{m}$?)

Under what assumptions on $\tilde{x}$ can you prove that the estimation method you propose work?

(b) The file *radioactive_sample_2.txt* contains a vector of measurements corresponding to a different radioactive sample. Does the estimation method described above work in this case? Submit the plot of the new moving average.

(c) A colleague suggests that the angle α between the trajectory of the particles emitted by the new sample and the vertical axis (illustrated in Figure 1) might be well modeled by a random variable $\tilde{a}$ that is uniformly distributed between $-\pi/2$ and $\pi/2$. Compute the pdf and mean of $\tilde{x}$ under this assumption.
(Hint: remember the trigonometric function tan and its inverse arctan.)
Would such model explain your observations in (b)?

(d) The sample mean can be affected by extreme values and outliers, whereas the sample median is more robust. The sample median converges to the median of an iid sequence of random variables even when the mean is not well defined. Use the sample median from *radioactive_sample_2.txt* to estimate c.