Homework 9

Due December 5 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. (Noisy measurement) We are interested in measuring a certain quantity modeled by a random variable $\tilde{x}$ with zero mean and unit variance. Our available measurements $\tilde{y} := \tilde{x} + \tilde{z}$ are corrupted by additive random noise, modeled as a random variable $\tilde{z}$ with zero mean and variance σ^2, which is independent from $\tilde{x}$.

   (a) What is the best linear estimate of $\tilde{x}$ given $\tilde{y} = y$?

   (b) What is the corresponding mean squared error?

   (c) What happens to the estimate and the error when $\sigma \to 0$? Explain why this makes sense.

   (d) What happens to the estimate and the error when $\sigma \to \infty$? Explain why this makes sense.

2. (Rufus) Nora's dog Rufus lives in her garden, which is the shaded area in Figure 1. After observing Rufus for a while she decides that his position within the garden is uniformly distributed (i.e. the probability density of his position is the same at every point of the garden). Let $\tilde{x}$ be the position of Rufus on the x axis in Figure 1 and $\tilde{y}$ his position on the y axis.

   (a) Compute the mean of $\tilde{x}$.

   (b) Compute the pdf of $\tilde{y}$. Sketch it.

   (c) What is the pdf of $\tilde{x}$ conditioned on $\tilde{y}$? Sketch it.

   (d) Are $\tilde{x}$ and $\tilde{y}$ independent? Justify your answer.

   (e) Are $\tilde{x}$ and $\tilde{y}$ uncorrelated? Justify your answer.

3. (Random vector) A random vector $\tilde{x}$ with zero mean has a covariance matrix $\Sigma_{\tilde{x}}$ with the following eigendecomposition

$$\Sigma_{\tilde{x}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}. \tag{1}$$

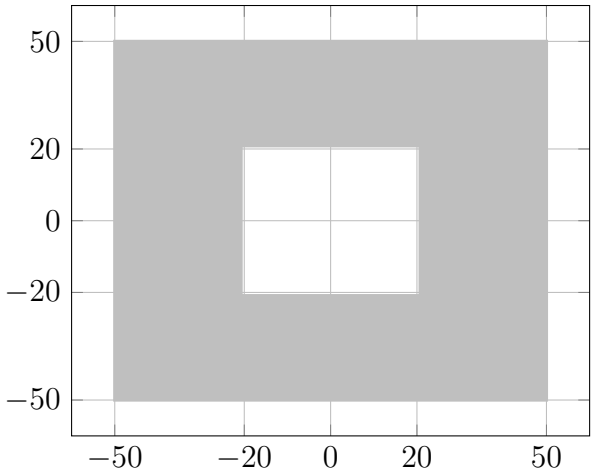

Figure 1: Nora's garden (in gray).

(a) What is the variance of each of the entries of the random vector $\tilde{x}[1]$, $\tilde{x}[2]$ and $\tilde{x}[3]$?

(b) Is it possible to find a unit-norm vector u such that the inner product between $\tilde{x}$ and u (i.e. the amplitude of the projection of $\tilde{x}$ onto that direction) has variance greater than 1?

(c) Find three constants a_1, a_2 and a_3, such that at least one of them is nonzero and $\mathrm{P}(a_1\tilde{x}[1]+a_2\tilde{x}[2]+a_3\tilde{x}[3] = 0) = 1$. Justify your answer mathematically, and interpret it geometrically.

4. (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.