# Homework 1

Instructions: This homework is made of two parts: programming and written problems.

A separate Brightspace assignment is created for each part.

**For the programming part:** You should submit your solutions in one single zip file, named "[yourFirstName-yourLastName].zip", to the homework1-code assignment through Brightspace. The zip file should include one folder, named "code" and one jing.txt file that includes the link to your JING screencast video (look at JING instructions on Brightspace). All the code for programming problems should be in the "code" folder. Make sure to mention the Python version you are using. Make sure you can import your .py files into Python without error, and test your solutions before submitting.

**For the written part:** Submit a PDF file named "[yourFirstName-yourLastName].pdf" that includes your solution, typed in an editor like MS Word or using LaTeX. Do **not** submit images of handwritten solutions.

## 1 [40 pts.] Data analysis

The number of students admitted to college A as freshmen is 1000 and on average 15% of the students drop out every year of college (assume all degrees take 4 years and that 15% of students in Y1 do not continue to Y2, and then 15% of students in Y2 similarly do not continue to Y3 and so on).

**(a)[5 pts]** Plot the Probability Mass Function (PMF) and Cumulative Distribution Function (CDF) of the random variable corresponding to **year in college** of a student (Hint: $X \in \{1, 2, 3, 4\}$).

**(b)[5 pts]** Calculate the expected **year in college** of a student (mean) and its variance. Show the steps of your calculation.

**(c)[10 pts]** Suppose that instead of 15%, $\alpha$% of students drop out per year ($0 \leq \alpha \leq 100$). What is the mean as a function of $\alpha$.

**(d)[10 pts]** Suppose college A merges with nearby medschool B to create university C. In the year of the merge, medschool B has 40 students in year 6, 50 students in year 7, and 100 students in year 8.

Calculate the new expected value of **year in college** at the combined university, C. Is the mean (as compared to part **b**) stable or sensitive to these new data points? What other statistical measures are there to estimate the average behavior? Are they less or more stable in regard to the outliers introduced in the merger? Justify your answer by computing those alternative measures for the original college A and the merged university C.

**(e)**[**10 pts**] Create a box plots for the **years in college** variables $X_A$ for college A and $X_C$ university C. (don't use python libraries for this - just draw it by hand or in your submission). Look here: `http://www.physics.csbsju.edu/stats/box2.html` for examples of box plots. You should have two boxes, one for $X_A$ and one for $X_C$ with their specific statistics.

## 2 [10 pts.] Irreducible data example

In class we discussed that not all datasets' dimensionality can be successfully reduced using PCA.

**(a)**[**2 pts**] Discuss the cases when PCA will fail.

**(b)**[**2 pts**] How do we quantify that it fails?

**(c)**[**6 pts**] Provide a minimal example of a dataset (specify the points as vectors of numbers) in which PCA will not work well for dimensionality reduction. Explain why. *Hint: Think of 2D points and reduction to 1D.*

## 3 [50 pts] Dimensionality reduction

For this question, you will use the hw1_data.txt dataset. Use Python for all your programming. You will have to submit your code in the zip file and your writing with the rest of your write-up.

**(a)**[**5 pts**] Load the data into a Python program and center it. Note: There should be a function called *center()* in your code that achieves this.

**(b)**[**10 pts**] Compute the covariance matrix of the data $\Sigma$ in three different ways.

*Hint: by using the definition of sample covariance, as a matrix product or as a sum of outer products. See book for details.* Use NumPy for linear algebra computations (https://docs.scipy.org/doc/numpy-1.13.0/reference/routines.linalg.html).

As a result, you should have three functions *cov1()*, *cov2()* and *cov3()* in your code. Measure the time that each function takes to compute $\Sigma$ for the dataset and report it in your Solution.pdf document. Discuss the differences in terms of algorithm complexity and explain the difference in measured times.

**(c)[5 pts]** Compute the eigenvectors and eigenvalues of $\Sigma$. The NumPy linear algebra module referenced above has a function that can help.

**(d)[10 pts]** Determine the number of principal components (PCs) r that will ensure 90% retained variance. How did you compute this? Provide a function in your code that determines $r$ based on an arbitrary percentage $\alpha$ of retained variance.

**(e)[10 pts]** Plot the first two components in a figure, with horizontal axis (x) corresponding to the dimensions and vertical axis (y) corresponding to the magnitude of the component in this dimension. There will be 2 traces with d points in this figure. Include the figure in your PDF solution. Also save the top two components in a text file "Components.txt" in the code folder, with each component on a separate line and represented as dd comma-separated numbers (i.e. the file should have two lines with dd numbers separated by commas).

**(f)[10 pts]** Compute the reduced dimension data matrix $A$ with two dimensions by projection on the first two PCs. Plot the points using a scatter plot (a two-dimensional diagram that places each sample $i$ according to its new dimensions $a_{i1}, a_{i2}$). Discuss the observations. Are there clusters of nearby points? What is the retained variance for $r = 2$? Argue for or against whether these are sufficient dimensions.