ICSI431HW1

Seoyeon Choi

February 2024

1 Part1-a

Part 1-a: Calculate PMF and CDF

Let X be the year in college for a student at college A, where $X \in \{1, 2, 3, 4\}$.

Since every year 15% of students will drop out, we need to calculate the Probability Mass Function (PMF) and Cumulative Distribution Function (CDF) for X.

PMF:

$$f(x) = P(X = x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = x)$$

CDF:

$$F(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$$

PMF: P(X=1) = 1.0,
P(X=2) = 0.85,
P(X=3) = 0.72249,
P(X=4) = 0.61412.

CDF: $F^{(1)} = P(X \leq 1) = 1$,
$F^{(2)} = P(X \leq 2) = 1 + 0.85 = 1.85$,
$F^{(3)} = P(X \leq 3) = 1 + 0.85 + 0.72249 = 2.5725$,
$F^{(4)} = P(X \leq 4) = 1 + 0.85 + 0.72249 + 0.61412 = 3.186625$.

2 Part1-b

Calculate Mean and Variance

Mean (μ):

$$\mu = E[X] = \frac{1}{n} \sum_{i=1}^{n} x_i f(x)$$

$$\mu = (1 \times P(1)) + (2 \times P(2)) + (3 \times P(3)) + (4 \times P(4))$$

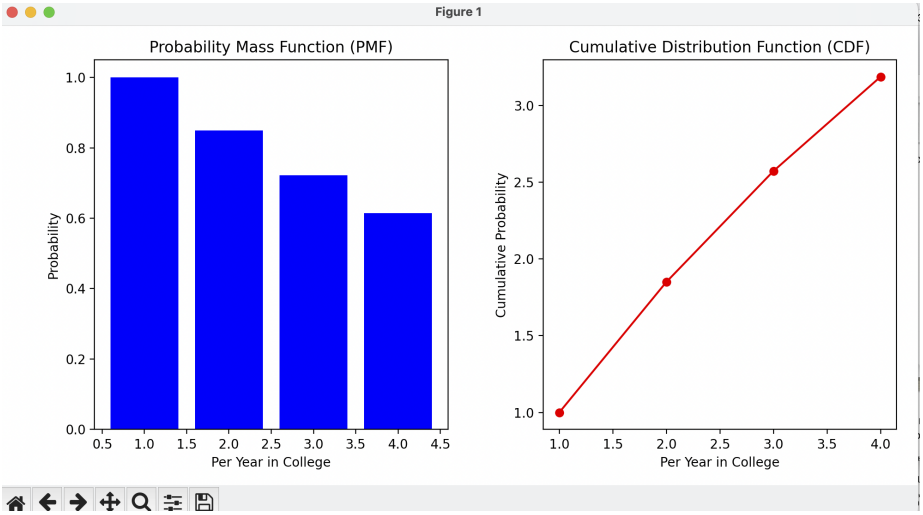

Figure 1: PMF vs CDF plot

$$= (1 \times 1.0) + (2 \times 0.85) + (3 \times 0.72249) + (4 \times 0.61412) = 7.324.$$

Variance (σ^2):

$$\sigma^2 = var(X) = E[(x-\mu)^2] = \sum_{i=1}^{n}(x_i - \mu)^2 f(x)$$

$\sigma^2 = (1-\mu)^2 P(X=1) + (2-\mu)^2 P(X=2) + (3-\mu)^2 P(X=3) + (4-\mu)^2 P(X=4)$.

$= (1-7.324)^2(1.0) + (2-7.324)^2(0.85) + (3-7.324)^2(0.72249) + (4-7.324)^2(0.61412) = 84.3802$.

Standard deviation (σ): $\sigma = 9.1858$.

3 Part1-c

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

When the dropping rate is α, let $p = \frac{\alpha}{100}$. For year k, PMF will be $(1-p)^{(k-1)}$. In order to calculate the mean, we will calculate each year's PMF value and use the formula to calculate the mean expected value.

Mean Expected Value:

$$\mu = E[X] = \frac{1}{n}\sum_{i=1}^{n} x_i f(x)$$

When dropping rate = α%

$$pmf(1) = \left(1 - \frac{\alpha}{100}\right)^0 = 1$$

$$pmf(2) = (1 - \frac{\alpha}{100})^{2-1} = (1 - \frac{\alpha}{100})^1$$

$$pmf(3) = (1 - \frac{\alpha}{100})^{3-1} = (1 - \frac{\alpha}{100})^2$$

$$pmf(4) = (1 - \frac{\alpha}{100})^{4-1} = (1 - \frac{\alpha}{100})^3$$

When the year is up to 4 years, the Expected Value (μ) is calculated as:

$$\mu = (1 \times P(x = 1)) + (2 \times P(x = 2)) + (3 \times P(x = 3)) + (4 \times P(x = 4))$$

$$= (1 \times 1) + (2 \times (1 - \frac{\alpha}{100})^1) + (3 \times (1 - \frac{\alpha}{100})^2) + (4 \times (1 - \frac{\alpha}{100})^3)$$

4 Part1-d

Calculate the new expected value of the year in college at the combined university, C.

College C Expected value (year = 8): $E[x] = \sigma x \cdot f(x) = (1 \cdot P(1)) + (2 \cdot P(2)) + (3 \cdot P(3)) + (4 \cdot P(4)) + (5 \cdot P(5$

I recalculated PMF values since the total number of students for college C is the sum of the total number of students in college A and college B.

$$= (1 \cdot \frac{1000}{1100}) + (2 \cdot \frac{850}{1100}) + (3 \cdot \frac{722.5}{1100}) + (4 \cdot \frac{614.25}{1100}) + (5 \cdot \frac{614.125}{1100}) + (6 \cdot \frac{654.123}{1100}) + (7 \cdot \frac{664.125}{1100}) + (8 \cdot \frac{714.125}{1100})$$

$$= 22.43749$$

Is the mean (as compared to Part b) stable or sensitive to these new data points?
Therefore, College A has a mean value of 7.324 , while College C has a mean value (expected value) of 22.43749. Since College C's mean value is significantly larger than College A's mean value, this indicates that the mean value is sensitive to outliers and new data points. Therefore, the mean is not stable.

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? Other Statistical Measures

- **Median**: The median is the middle point value in a dataset when it is sorted in ascending order. For example, if the dataset is 1, 2, 3, 4, 5, then the median is 3. The median is less sensitive to outliers compared to the mean.

- **Mode**: The mode is the value at which the probability mass function or probability density function attains its maximum value, depending on whether the variable is discrete or continuous.
- **Variance**: Variance ($Var(x)$) provides a measure of how much the values of X deviate from the mean or expected value of x. It helps observe how data points are spread around the mean value.
- **Standard Deviation**: The standard deviation provides information about the spread or deviation of data around the mean. It is useful for observing how much variation exists from the mean.

Justify your answer by computing those alternative measures for the original college A and the merged university C. College A

Mean of College A:

$$E[X_A] = \sum x \cdot f(x) = 7.324$$

Variance of College A:

$$\sigma^2_{XA} = var(X_A) = E[(X_A - \mu)^2] = \sum (X_A - \mu)^2 \cdot f(x)$$

$= (1-7.324)^2 \cdot 1.0 + (2-7.324)^2 \cdot 0.85 + (3-7.324)^2 \cdot 0.7224 + (4-7.324)^2 \cdot 0.6141 = 84.38$

Standard Deviation of College A:

$$StandardDeviation_{XA} = \sqrt{\sigma^2_{XA}} = \sqrt{84.38} = 9.185$$

College C

Mean of College C:

$$E[X_C] = \sum x \cdot f(x) = 22.43749$$

Variance of College C:

$$\sigma^2_{XC} = var(X_C) = E[(X_C - \mu)^2] = \sum (X_C - \mu)^2 \cdot f(x)$$

$= (1-22.43749)^2 \cdot 0.909 + (2-22.43749)^2 \cdot 0.7727 + (3-22.43749)^2 \cdot 0.6568 + (4-22.43749)^2 \cdot 0.5583$

$+ (5-22.43749)^2 \cdot 0.5583 + (6-22.43749)^2 \cdot 0.5947 + (7-22.43749)^2 \cdot 0.6037 + (8-22.43749)^2 \cdot 0.6492 = 1788.1289$

Standard Deviation of College C:

$$StandardDeviation_{XC} = \sqrt{\sigma^2_{XC}} = \sqrt{1788.1289} = 42.286$$

5 Part1-e

Create box plots for the years in college variables XA for college A and XC university C.

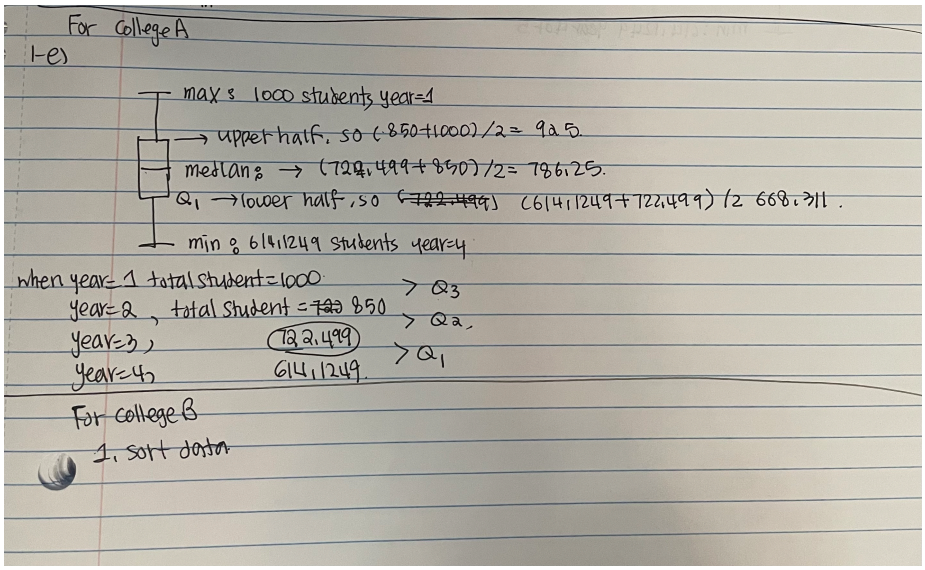

Figure 2: College A box plot

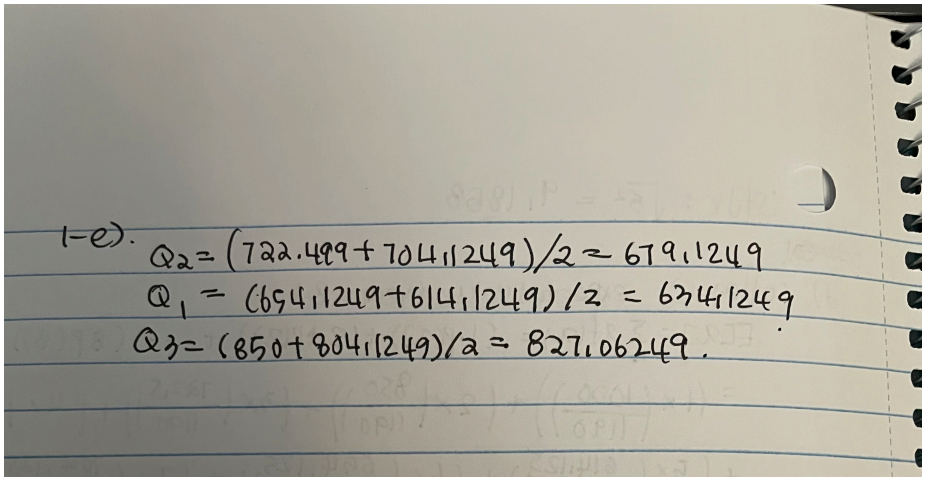

Figure 3: Calculation to get Q1, Q2, Q3 for college C.

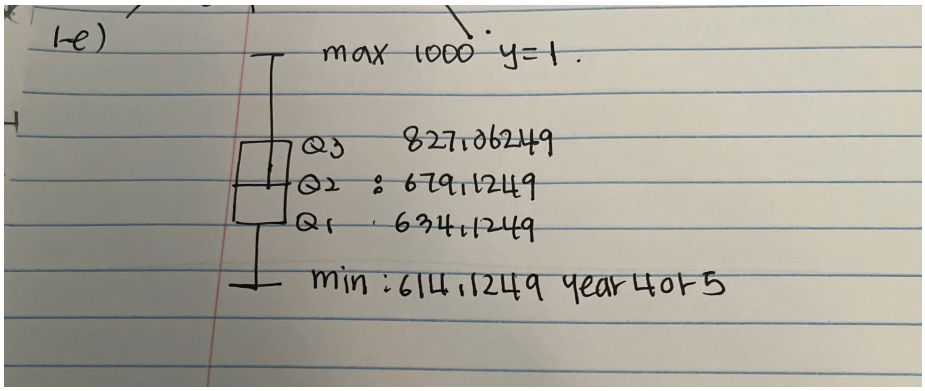

Figure 4: College C box plot

6 2. Irreducible data example

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

PCA is a method to reduce dimensionality and find an optimal basis that represents most data points. PCA will fail if the data structure is not linear or the variance of the data is not aligned with the principal components. To perform PCA, we can find the optimal basis with the highest variance or lowest mean squared error. Therefore, if the data points do not have a linear relationship, PCA will fail. Additionally, if all data points are the same across all variables and do not have the highest variance, PCA may have difficulty capturing dimensionality and the optimality criterion.

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

The variance is high, we can say that we have successfully performed PCA to reduce the dimensionality and obtain an optimal basis that captures most of the data points. However, if the variance ratio of the principal components is low, it means that PCA did not capture most of the variance. I can also view a scatterplot of the data projected onto the optimal basis (principal components) to determine whether the data exhibits a pattern along the optimal basis found

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. When we reduce dimensionality, most data points should

be within subspace with a new basis (u1, u2, etc). So optimal subspace maximizes the variance and minimizes the squared mean error. Think of this two

datasets with 2 dimensional points : (1,1,) (2,2) and (1,1), (2,2) These points are in 2D space. If I reduce 2D into ID using PCA to find maximum variance or minimum squared mean error, PCA dimensionality reduction will not work well. Because all points in those two datasets lie on the exactly same line and there will be no variance that is perpendicular to this line. (We can not find orthonormal vectors u1, u2 ,,, ud where x =a1u1 + ... + AuUd). So reducing the dimensionality of this dataset is ineffective if we cannot capture the variance

7 Part3-b

Measure the time that each function takes to compute Σ for the dataset and report it.

Covariance Method 1 uses $E[(x-\mu)^T(x-\mu)]$, the definition of sample covariance to calculate the covariance matrix by using centered data points. It took 0.00013154902262613177 seconds.

Covariance Method 2 uses $\Sigma = \frac{1}{n}Z^TZ$, a matrix product method. It took 0.00005728198448196053 5 seconds as elapsed time.

Covariance Method 3 uses $\frac{1}{n}\sum_{i=1}^{n} Z_i \cdot Z_i^T$ as a sum of outer products. It took 0.3899385699769482 seconds as elapsed time.

Discuss the differences in terms of algorithm complexity and explain the difference in measured times.

Covariance Method 1 uses $\Sigma = E[(x-\mu)^T(x-\mu)]$, the definition of sample covariance, to calculate the covariance matrix. This method has an algorithm complexity of $O(d^2n)$ as algorithm complexity where d is dimension of data and n is representing number of data points. This method need to calculate mean of each column vector and calculating product of each centered data points . so this makes $O(d^2n)$ to compute. Also elapsed time will be depends on the dimension of data.

Covariance Method 2 calculates the covariance matrix using a matrix product method, computing the matrix product Z^TZ, where Z represents the centered data points. This method also has an algorithm complexity of $O(d^2n)$ operations. Based on the elapsed time output, this method was slightly faster than the first method.

Covariance Method 3 calculates the covariance matrix using $\frac{1}{n}\sum_{i=1}^{n} Z_i \cdot Z_i^T$ as a sum of outer products. For each centered data point we do outer product and summing them up up to n and this requires $O(d^2n)$ operations. But it took longest elapsed time output compare to method1 and method2. Even if algorithm complexity is same as method2, outer product may requires more product computations and summing all of them so it may takes more time.

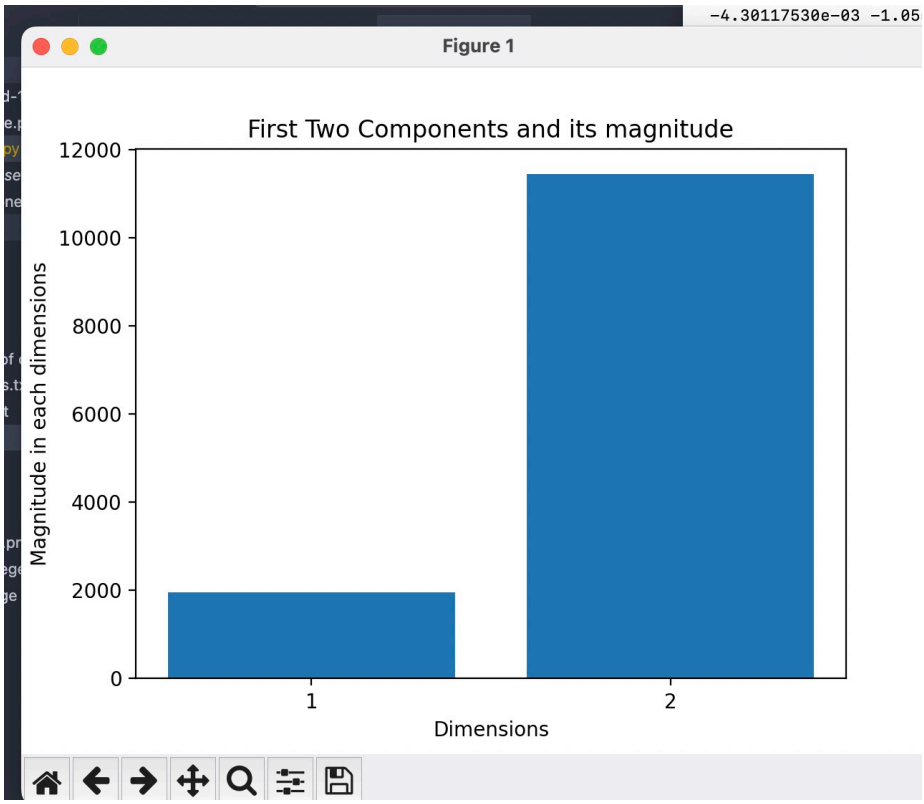

Figure 5: Task e : first two components and its magnitude

8 Part3-e

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 com-ponent 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). Componenets.txt file is created in code folder. Plots are included at the end

(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 ai1, ai2). 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. Clusters of nearby points show patterns in data points. Clusters of nearby points

can be used to reduce dimensional by capturing high projection variance over the original data points (capturing principal components). In this case, it captures the first two principal components r=2 (dimension is 2). We captured the first two principal components using the eigenvalues of the covariance matrix (1 and 2 with all eigenvalues non-negative and sorted). Therefore, if the expected variance for r =2 is high enough, it means that the first two principal components represent most of the data points well in the low-dimensional space. For scatter plots, r=2 appears to capture high prediction variance. Therefore, an r =2 dimension appears to be sufficient to capture patterns in the data. You can also find the minimum mean squared error value to obtain an optimal reference for representing your data points.

In order to calculate the retained variance for $r = 2$, we can use the formula:

$$variance = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$$

where r represents the number of principal components (in this case, $r = 2$), and λ_i are the eigenvalues obtained from PCA.

When the retained variance for $r = 2$ is 0.9905664757824717, it indicates that these two principal components capture approximately 99.05% of the total variance in the data.

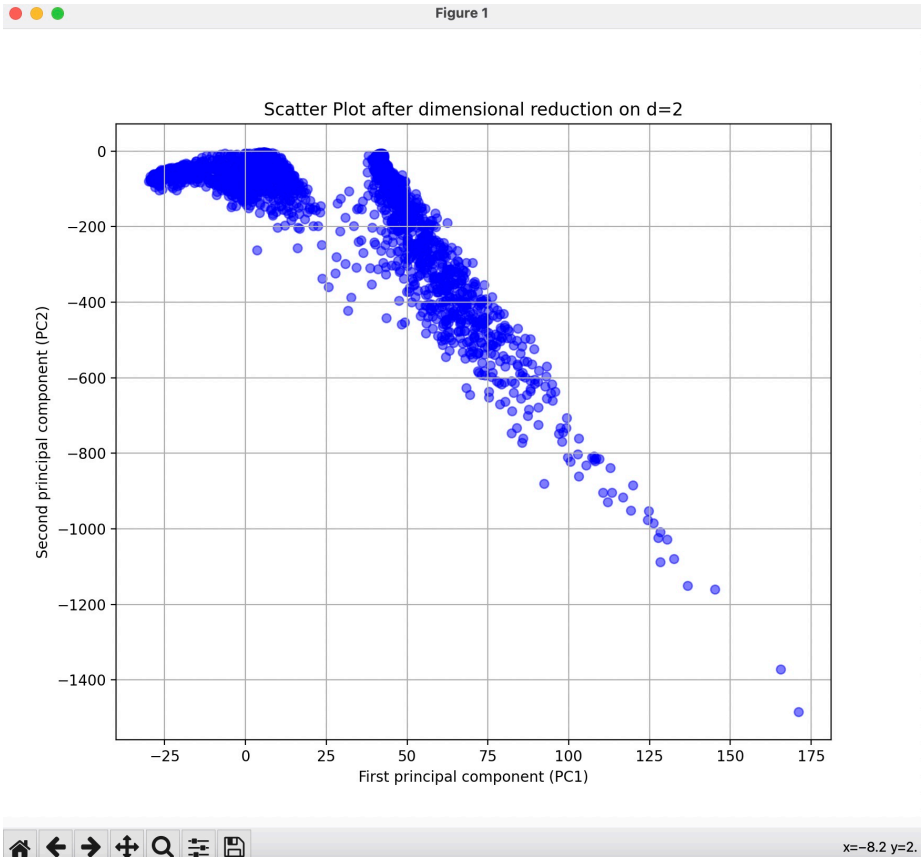

Figure 6: 2 Dimention Scatter Plot : task f