Problem 1

A.

We can find the covariance matrix by $c_{jk} = \frac{1}{m-1}\sum_{i=1}^{m}(m_{ji} - \bar{m}_j)(m_{ki} - \bar{m}_k)$

B.

To find the principle components, we calculate the covariance matrix as above, then we find the eigenvectors/values for the covariance matrix. Then, we take the eigenvectors that correspond to the eigenvalues, which are the principle components. We can then take the eigenvectors that correspond to the top k eigenvalues as needed, and transform our data accordingly.

C.

The principle components represent the vectors that capture the highest variance between cells. It maximizes the variances between the gene expression profiles.

D.

The maximum number of principle components with non-zero variance is the minimum of n and m-1. We mean centered the matrix to find the PCA, which means one column is a linear combination of the others then, so we have m-1 principle components from the features. in the event that n is smaller than m-1, then the rank is lower and the maximum number of eeigenvalues would depend on n. Thus, we have min(m-1, n)

Problem 2

Data points: [[1, 1], [-1, -1], [1, -1], [-1, 1], [2, 2], [-2, -2], [-2, 2], [2, -2]]

This gives us the following covariance matrix: array([[2.85714286, 0.], [0. , 2.85714286]])

which gives us the following eigenvalues: [2.85714286 2.85714286] as desired, they are the same.

Problem 3

A.

PCA is a decorrelation operation, which removes correlation between variables. That does not necessarily mean that the variables are then independent.

B.

if X is a multivariate gaussian random variable, then by doing PCA transformation, we are finding y=px, which is a linear operation. Since X is gaussian, and we're applying a linear operation on a gaussian, the result of that is still gaussian. Thus, if X is a gaussian that is uncorrelated, that means the values are also independent, and the operation on them is also independent. Thus, equation 1 holds.

Problem4 Link