PROBLEM 2

1) Centering

What it means

Centering is when you subtract the mean of each feature of the data from each individual data point.

Why it's important

Centering is important because it allows the matrix decomposition in PCA to find the actua covariance matrix of the data and therefore find the eigenvectors with the highest variance.

How to apply it

To apply centering:

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i}$$
$$\boldsymbol{x_i} = \boldsymbol{x_i} - \boldsymbol{\mu}$$

2) Scaling

What it means

Scaling is the transformation of data to a desired interval. It typically refers to normalization or standardization (discussed below).

Why it's important

Scaling is important for PCA because simply having a feature vector with a larger scale than the other features will cause the PCA algorithm to identify that eigenvector to have the highest variance, although that may not be true.

How to apply it

Scaling is typically done by applying a scalar multiplication to each feature vector, which can be shown as a Hadamard product:

$$\boldsymbol{v}' = \boldsymbol{v}^{\mathbb{R}^d} \odot [s_1, s_2, s_3, ..., s_d]$$

After these features are scaled they can then be shifted by adding a unique scalar value to each feature dimension. Any other linear operation or function can also be applied in a similar fashion.

3) Normalization

What it means

Normalization is when each feature dimension is scaled to be within the interval [0, 1]

Why it's important

Normalization is important for a similar reason as discussed in the scaling section. If all of the features are normalized, then the PCA algorithm will treat them all as 'equal' when attempting to find the direction of highest variance. Using normalization avoids the algorithm assuming one feature is better than another simply because it's values are on a large scale.

How to apply it

Normalization can be applied by shifting all of the data for each feature such that the minimum value is 0 and then dividing by the maximum value of that feature.

4) Standardization

What it means

Standardization is when each feature dimension is shifted to have a mean of zero and scaled to have a maximum range of one standard deviation in each direction.

Why it's important

Standardization attempts to convert each feature dimension into a normal distribution. This is useful because the normal distribution is well understood and documented, as well as provides us a way to calculate things like the variance more easily.

How to apply it

To apply standardization, the mean (μ_d) and standard deviation (σ_d) of each feature dimension is calculated, and then the following equation is applied to all data points:

$$x'_{n,d} = \frac{x_{n,d} - \mu_d}{\sigma_d}$$