Principal Component Analysis for dimensionality reduction

PROJECT TITLE: 3D Geometry HULL Calculations

SEP Group: HULL2

BY: Gordon Tang (a1743596)

Research scope:

Given a complex 3D geometry, the program will need to calculate the hull of a 3D shape. To perform this calculation, implementing a Principal Component Analysis (PCA) algorithm will reduce the dimension of the 3D geometry.

Basic concept:

PCA is a dimensionality–reduction method that reduces the dimensionality of large data sets into smaller sets that still contains most of the information in the larger set. However, reducing the number of variables naturally decreases the accuracy of the data set but is traded for the simplicity of the data set.

The pseudocode of PCA:

1. Standardize the range of continuous initial variables
2. Compute the covariance matrix to identify correlations
3. Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
4. Create a feature vector to decide which principal components to keep
5. Recast the data along the axes of the principal component

Explanation of each step:

1. Standardize the range of continuous initial variables so that each variable will contribute equally to the analysis. Without standardizing the initial variables, there will be large differences between the ranges of initial variables, where these variables will dominate over the variables with smaller ranges. Leading to biased results for the reduction algorithm. Standardization is calculated by
$$z = \frac{value - mean}{standard\ deviation}$$
2. A covariation matrix is computed to identify the variable's correlations with one another. If the variables have a high correlation shared between one another, these variables will contain redundant information. The signs of the covariances distinguish the correlations between variables such that the covariance being positive shows the two variables increase or decrease together, which means they are correlated. However, covariance that are

negative means that one variable increase while the other decreases, which means they are inversely correlated.

3. Eigenvectors and eigenvalues of the covariance matrix are used to compute the principal components of the data. The principal components are constructed of new variables as linear combinations or mixtures of the initial variables. The principal component is computed of variables that are uncorrelated and most of the information within the initial variables is compressed into the component.
4. To calculate the feature vector, the components calculated from the eigenvectors and eigenvalues are either decided to be kept the same or discard the lesser significant components. The remaining vectors from the matrix are formed together to get the feature vector.
5. The reduced data set is formed by using the feature vector to reorient the data from the original axes to the ones represented by the principal component. This can be calculated by multiplying the transpose of the feature vector with the transpose of the original data set.

$$Final\ Dataset = Feature\ Vector^{T} * Standardized\ Original\ Dataset^{T}$$

Source

A Step-by-Step Explanation of Principal Component Analysis (PCA), A Step-by-Step Explanation of Principal Component Analysis (PCA) (2021). Available at: https://builtin.com/data-science/step-step-explanation-principal-component-analysis (Accessed: 22 August 2021).