2.1. PCA

2.1.1. PCA overview

The principal component analysis (PCA) is a technique to reduce the dimensionality of data set (sample) composed of many interrelated variables, while retaining the sample’s information as much as possible. The PCA analysis is performed on one of a square symmetric matrix which can be SSCP matrix (pure sums of squares and cross products), Covariance matrix (scaled sums of squares and cross products), or correlation matrix (sums of squares and cross products from standardized data). The results after analyzing using SSCP and Covariance will not be difference because those matrix types are only different in a global scaling factor as well as the results will be different in correlation matrix. This is because the correlation matrix is used if the variances of each variates is different significantly, or the scale of measurement of each variate differs. In this project, the covariance matrix is used to get reduced dimensionality of data set.

A close up of a map

Description automatically generated

The figure above shows bi-variate scatter plotting with first and second principal component. The point (x1, x2) is projected onto the first principal component which coincide with directions of maximum variation of the original observations to make projected point y1. All the points of X (X is set of all points in the figure) are projected onto the principal component to reduce the dimensionality of 2D original data set. The first principal component has the property that the variance of the projected points Y (Y is set of all projected points onto the axis) is greater than the variance of the points projected any other axis that passes through the centre of the elliptical of points. Likewise, the second principal component indicates the second biggest variance of any other possible axis or line.

2.1.2 Steps to get PCA

Step 1: Zero-mean data

This step is to transform all variables in same scale not to lead biased results. This step helps to reduce the difference of variance between initial variables and the variables of points projected onto principal component. If PCA is done without this step, points with a variance, ranges between 0 to 100, will dominate over a variance, ranges between 0 and 1. Normalization to corresponding scales will help to avoid this problem. This can be done by subtracting mean of variables from each variable.

z = value(x) – mean(μ)

Step 2: Covariance matrix

The covariance matrix is a symmetric matrix which is used to understand the difference between two variables. It means if there are variables highly correlated, the redundant information is contained. In this project, 3D coordinates will be projected onto first principal component. So, one vector will be composed of 3 variables which are x, y and z axis. 3×3 covariance matrix for a vector will be:

The covariance matrix for N number of vectors will be sum of squared divided by the number of the vectors.

Step 3: Feature vector

At this stage, eigenvalue and eigenvector are computed using covariance matrix calculated in the previous step by simple equation. The x and λ indicate eigenvector and eigenvalue of covariance matrix respectively.

(C – λI) = x

Each column of eigenvector is principal component and the first principal component is computed using the biggest eigenvalue as well as the first principal component maximizes the possible variance. The second principal component is also decided by the second biggest eigenvalue and the rest principal components are taken in the same way (i.e. each column of eigenvector has eigenvalue as well as the larger the value eigenvalue is, the bigger the variance eigenvector has).

Step 4: Projection

To project vectors onto principle component, zero-mean vectors are used. Desired principle component is chosen, and projected vector will be simply computed by multiplication between zero-mean data and principle component. Needless to say, first principle component has to be used.

*Projected data set* =

2.2. Correlation

2.2.1. Pearson’s correlation

The Pearson’s coefficient r, a.k.a. the Pearson Product-Moment Correlation is a measure of association between two variables. The Pearson’s r coefficient is defined as the percentage of the covariance of two vectors composed of numerical data after normalized using the square root of those variances. The coefficient is determined in the range between – 1 and 1. The coefficient value 1 indicates two variables are perfectly linear relationship as well as the value of -1 shows two variables are negatively related. The computation of the coefficient can be represented mathematically.

2.2.2. Cross correlation

The cross correlation is a method to measure interrelation of series as a function of the displacement of one signal in relation to the other, and generally used to search similar pattern inside a long signal for a shorter signal. In the simplest example, there are signals labeled ,, n = frame number between identical time gap (i.e. number of samples). Mathematically, the cross correlation can be computed using the equation.

2.2.3. Autocorrelation function

The correlation analysis is usually aimed to identify the relationship between variables, which can be between two variables (Pearson’s correlation), one to multiple variables (Multicollinearity) and identical two variables (Autocorrelation function). The basic idea of Autocorrelation function is cross correlation function. Cross correlation is a measure of similarity of two signals, or it can be numerical variables, as a function of a time-lag applied to one of them. In autocorrelation function, identical two variables are used (i.e. the variables used in previous section , become , on the same equation) based on cross correlation method. Autocorrelation function generates correlation coefficient at every different time lagging. So, this can be utilized to assume the period (time-series) of signal which can be critical property of the signal.

2.3. Mutual Information

2.3.1. Mutual information definition

Mutual information (MI) is used to measure dependency between two random variables. Likewise, this method can be utilized as a measure of the similarity between two variables. For the discrete distributions, the following equation is used to compute mutual information score. Let X and Y is discrete random variables.

* If X and Y are independent, mutual information will be 0 since p (x, y) = p(x)×p(y)

Things we can learn after using mutual information for signals of two sensors.

* The stored information in one variable to another variable
* The degree of the predictability of the second variable by knowing the first.

2.3.2. Difference of Mutual information from correlation

Mutual information is used to measure the general dependence. On the other hand, the correlation function is to measure the linear dependence. For the measure of the dependence, the mutual information is a finer method. Additionally, the correlation function is used only for numerical sequences. However, the mutual information can be implemented to symbolic sequences along with numerical sequences, which means mutual information can be applied to much more various types sequences of events from science and engineering to natural and social developments such as biological sequences, shopping list, program execution sequences and so on. Thus, the mutual information is good alternative to the correlation function for symbolic sequences.

2.4. Properties of signal

* Mean

Mean value of discrete signal can be computed in simple equation. Where f(t) is input signal, and T is a reciprocal number of fundamental frequencies

* Variance

Variance is a distance between mean value and variable, which is like amplitude of signal with offset 0.

* Max value
* Min value
* Offset
* Estimated time series (periodic) of random discrete distribution by autocorrelation

A screen shot of a social media post

Description automatically generated

The autocorrelation is to find out the correlation of between a variable and itself at different sequent phases. The autocorrelation can be defined using the values or . This can be computed using formula.

Where n is phase difference between a signal and itself. At each phase difference correlation is calculated to find the highest correlation point where becomes time series of random discrete signal.

* Skewness and Kurtosis

The measurements mentioned above that are mean, variance, max and min value are numerical measures of signal or distributions. However, the skewness and kurtosis can be used to estimate the shape of a distribution and give us more precise shape information. Skewness shows how much and where to the distribution skews as well as the kurtosis indicates how tall and acute the centre of peak is. The estimation of shape can be done based on normal distribution. The skewness and excess kurtosis of a normal distribution are 0. Based on a normal distribution, the values of skewness and excess kurtosis tells how different from normal distribution in terms of shape (i.e. if values are close to 0, the shape of measured distribution is close to a normal distribution). Both of those can be computed with simple equation.

* Dominant frequency of discrete Fourier transforms and its energy