We will derive a notion of loss of information in the projected space.

Let us say that … be the 1st data point (in D- dimensional space) and be the direction of projection, then the linear combination i.e. gives us a scaler value (i.e. projected value). The scaler value is the projection of first data point … along the direction .

So, we can say that the projection of a d-dimensional data point . (a scaler value)

Similarly, \*\*mean of the data in the projected space\*\* would be:

Each data point, which is a d-dimensional vector in the original space, gets projected into the projected space.

The idea of information is equal to [variance](https://en.wikipedia.org/wiki/Variance). If, in a dataset, all the columns have constant values, then there is no information gain because every data point would be the same.

We have got original data point in the projected space, and the original mean in the projected space. Now, we would want to maximise the variance in the projected space in order to capture the maximum information.

\*\*Variance in the Projected Space is -\*\*

= Mean in the projected space

= nth data point in the original space

= nth data point in the projected space

= Mean in the original space

= Variance in the projected space

=

=

=

=

=

= Eigen Vector ( )

Find the that maximizes . is the direction where we project the data and obtain the maximum variance in projected space.

The angle is the best angle which minimizes the loss of information in the covariance matrix ( C ) of the data.

The can be computed as:

= argmax = Eigen Vector (C )

We take the covariance matrix of the data and take the first eigen vector of that matrix.

We have written our objective function, both in terms of data and parameter –

i.e. Find the that maximizes

Where

is a parameter

, and can be estabilished from the data. Put simply, these are not parameters.

We solved the equation to get the best value for .

I have D-dimensional data and we do first K projections/(eigen vectors), then how many parameters am I learning?

Parameters Learned = D$\*$K which, in the above case accounts to 784$\*$2 = 1568 parameters. Through PCA, we have converted the dimensions of the original N$\*$D data into D$\*$K projections or representation, which express the variance in data.

When to use PCA:

* When the data is multivariate and Numeric
* When the number of features are large
* When the data is Unimodal
* When the Class labels are not present
* To reduce noise and outliers
* To visualize the data

Build a map from the Pair wise distance. We get the rotated version of the map. We fix the first point and compute the distance of this point from all the other points.

Loss of Information in the Pair wise Distance-