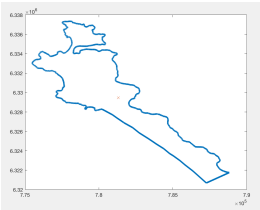


Principle Component Analysis

By: Miriam Galecki

Principle Component analysis is a linear projection method used to reduce the number of parameters. It works by transferring a set of correlated variables onto a new set of uncorrelated variables. This is also to say that the direction of most variance, for the new data after PCA, is along the x-axis. Principal component analysis can be viewed as a way to rotate the existing data to a new position whose axes are orthogonal and represent the direction with the maximum variability. PCA is also a very useful tool in reducing dimensionality.

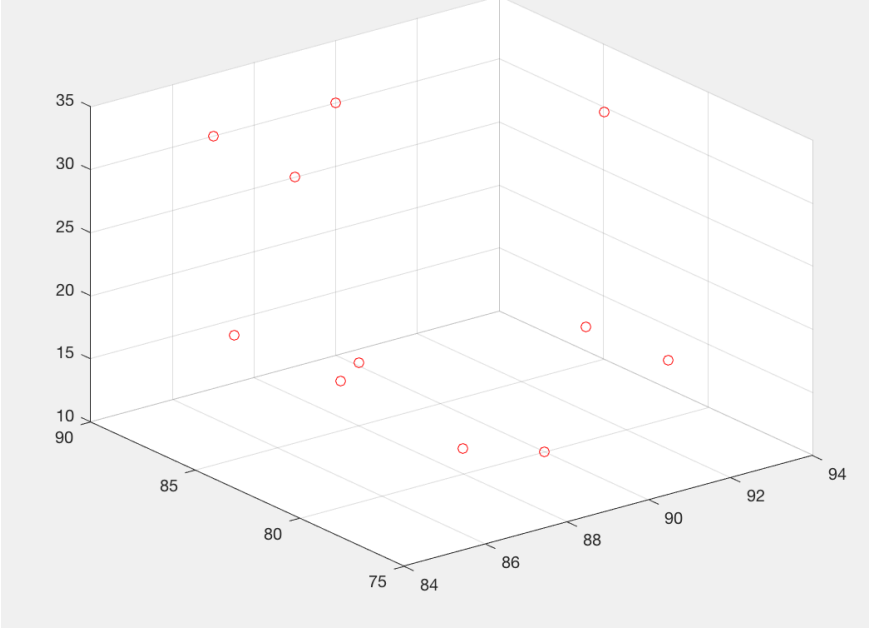
I will start out with a quick explanation on how to do PCA. For this example I have created the random data set below of 11 students who each took 3 tests. I will call this data set the Raw Data Matrix.

Raw Data Matrix

Student	Test		
	90	90	30
	85	80	20
	87	90	30
	90	80	10
	88	85	15
	93	83	32
	89	90	25
	88	80	12
	92	78	17
	87	89	15
	91	80	19

Below is a plot of the original data

Plot of Mean Original Data

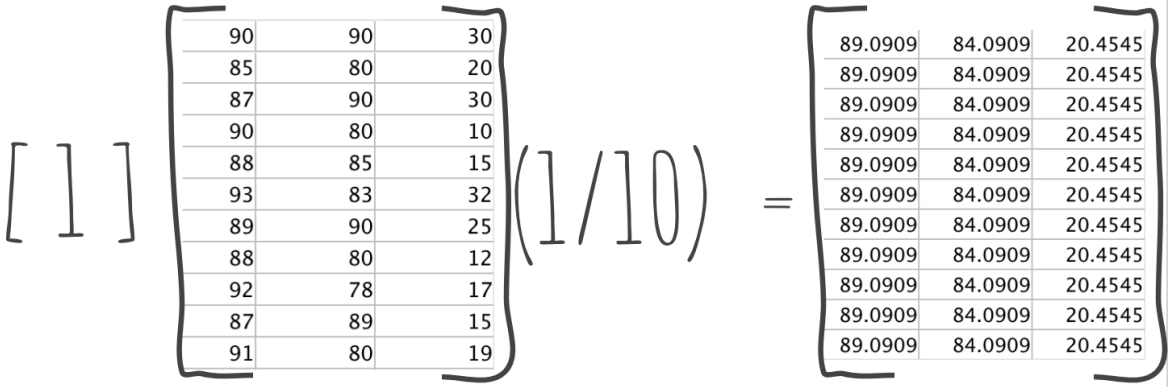


The first step in the PCA is to calculate the Mean Adjusted data. Since PCA transforms the data so that it is centered at the origin we are going to find a matrix that is the mean of each each test filled in the columns of a matrix. This matrix will be called the Mean Adjuster Matrix. A is the raw data matrix. A is an n x m matrix. [1] is an n x n matrix filled with ones.To find the Mean Adjuster matrix we will perform the following calculation. Where MC1 is the mean of column 1.

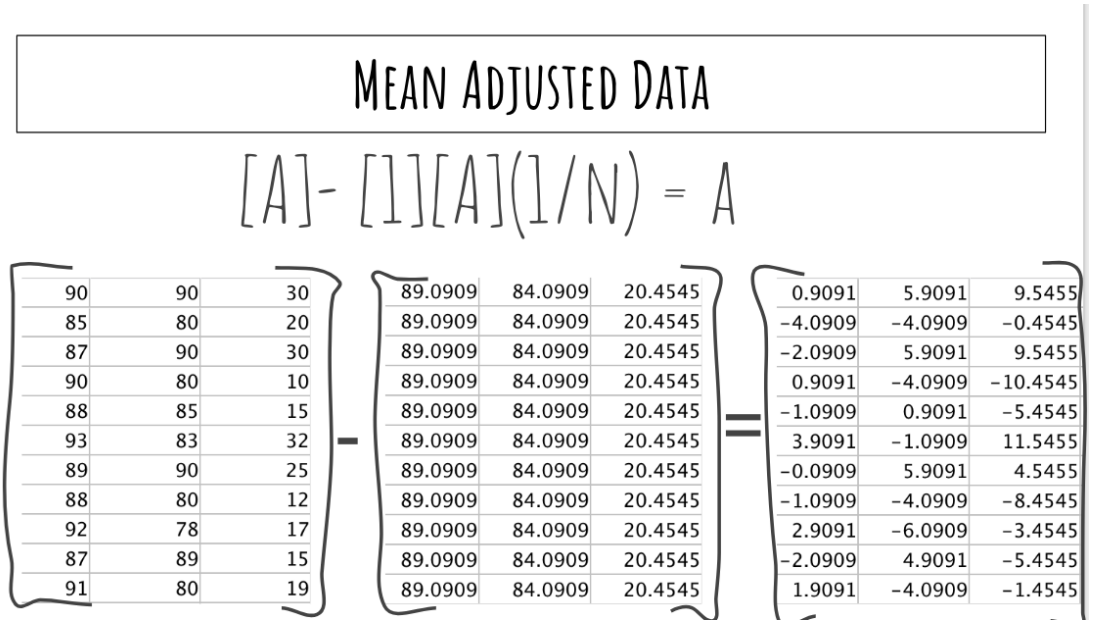
Find mean adjuster Matrix

$$[1][A](1/N) = \begin{bmatrix} MC1 & MC2 & MC3 \\ \downarrow & \downarrow & \downarrow \\ & (nX3) & \end{bmatrix}$$

Above is the general equation. Below is a picture of the calculation for our example.

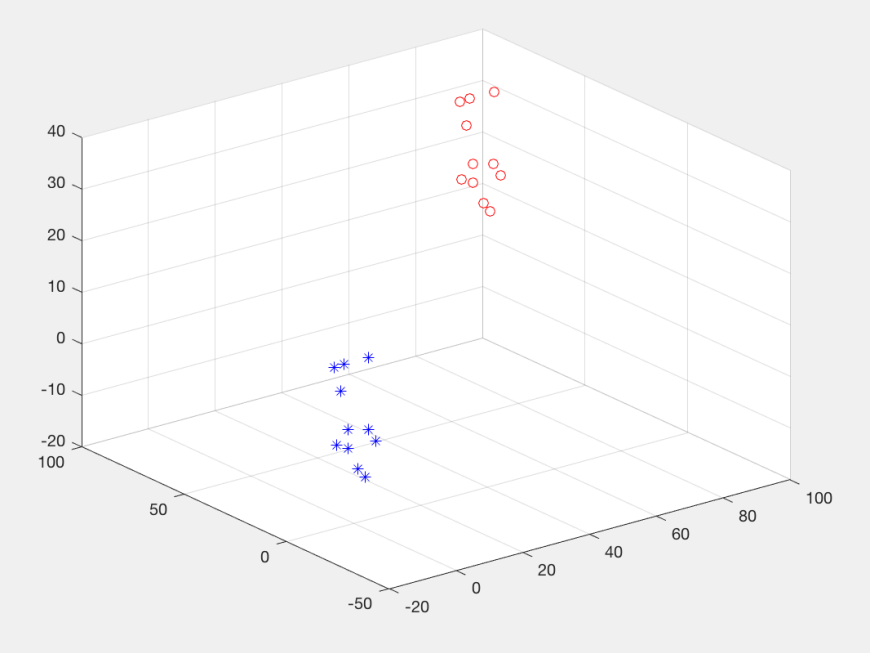


Now to actually find our mean adjusted data we will subtract our mean adjuster matrix from our original data matrix. This calculations makes intuitive sense. It makes sense to subtract the mean from each point if we want the data to be centered at the origin/the mean of our new data to be zero. Below is the calculations to find our mean adjusted data.



Below is a plot of our mean adjusted data. The original data is in red and the adjusted data is in blue. As you can see this data maintains the same shape and general orientation. It is just shifted so that the mean of this data is (0,0,0)

Plot of Mean Adjusted Data



The next step is to compute the covariance/variance matrix. Note that the covariance between n and n is the same thing as the variance of n. The covariance matrix will have the variance of 1 in the top left corner denoted as sigma 11 and the covariance of 1 and 2 denoted as sigma 12 in the first row second column and so on and so forth. Below are the covariance and variance equations along with the layout of the covariance matrix for a general problem.

Covariance/ Variance Matrix

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{(n-1)}$$

$$\text{Cov(X)} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

$$variance = \frac{\sum(X-\mu)^2}{N}$$

Below is the filled in covariance matrix for our example.

$$\begin{bmatrix} \text{Var}(1) & \text{Cov}(1,2) & \text{Cov}(1,3) \\ \text{Cov}(2,1) & \text{Var}(2) & \text{Cov}(2,3) \\ \text{Cov}(3,1) & \text{Cov}(3,2) & \text{Var}(3) \end{bmatrix}$$

$$\begin{bmatrix} 5.6909 & -3.1091 & 3.9545 \\ -3.1091 & 23.4909 & 20.1545 \\ 3.9545 & 20.1545 & 59.0727 \end{bmatrix}$$

We computed the covariance matrix because we are interested in which direction the data varies the most. To do this we will compute the eigenvalues and the eigenvectors of the covariance matrix. We do this by first computing the eigenvalues. As shown below

Find Eigenvalues

$$|\text{Covariance} - \lambda I| = 0$$

$$\begin{vmatrix} 5.6909 - \lambda & -3.1091 & 3.9545 \\ -3.1091 & 23.4909 - \lambda & 20.1545 \\ 3.9545 & 20.1545 & 59.0727 - \lambda \end{vmatrix} = 0$$

Characteristic Equation :

$$88.2545\lambda^2 - \lambda^3 - 1426.02\lambda + 4151.46$$

$\lambda 1 = .37403$
$\lambda 2 = 16.2619$
$\lambda 3 = 68.2523$

Using the eigenvalues we now calculate the eigenvectors. Below are the eigenvalues and their associated eigenvectors.

Find Eigenvectors

$\lambda 1 = .37403$	V1 =	0.9220 0.3380 -0.1890
$\lambda 2 = 16.2619$	V2 =	-0.3855 0.8480 -0.3636
$\lambda 3 = 68.2523$	V3 =	0.0374 0.4081 0.9122

Because we computed these eigenvectors from the covariance matrix they are useful in determining in which direction our data varies the most/ least. The eigenvectors with the largest eigenvalue corresponds to the dimension that has the strongest correlation in the data set. In other words the direction in which the data varies the most. In this case the data varies the most in the direction of V3.

To find the final data we will create two new matrices. One matrix is the RowFeatureVector which is a matrix with the eigenvectors in order from highest variance first to lowest variance. The second Matrix we need to calculate is the RowZeroMeanData which is the mean adjusted data transposed. Below are these two matrices in the context of our problem.

Find The Final Data

RowFeatureVector =

0.0374	-0.3855	0.9220
0.4081	0.8480	0.3380
0.9122	-0.3636	-0.1890

RowZeroMeanData =

0.9091	-4.0909	-2.0909	0.9091	-1.0909	3.9091	-0.0909	-1.0909	2.9091	-2.0909	1.9091
5.9091	-4.0909	5.9091	-4.0909	0.9091	-1.0909	5.9091	-4.0909	-6.0909	4.9091	-4.0909
9.5455	-0.4545	9.5455	-10.4545	-5.4545	11.5455	4.5455	-8.4545	-3.4545	-5.4545	-1.4545

To find the final data we simply take the transpose of the multiplication of RowFeatureVector X RowZeroMeanData.

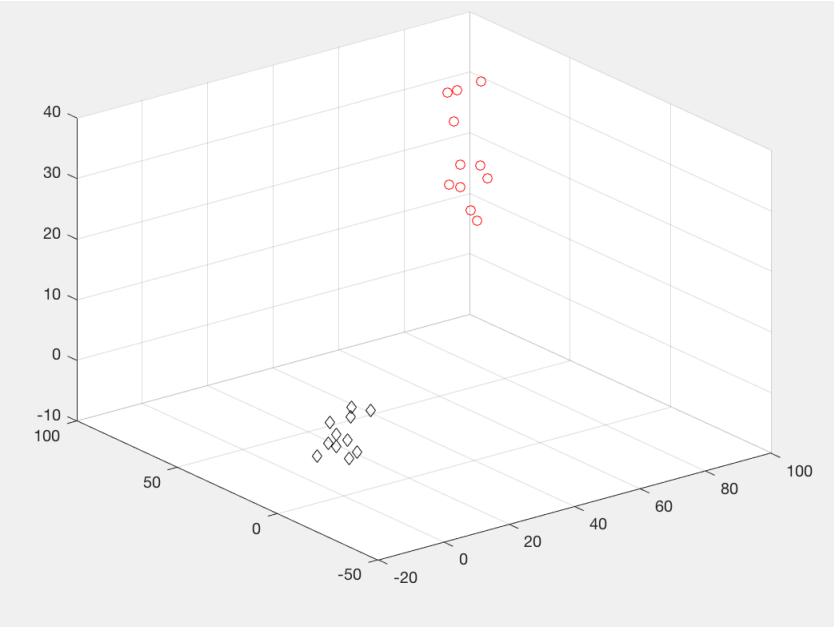
Data = RowFeatureVector X RowZeroMeanData

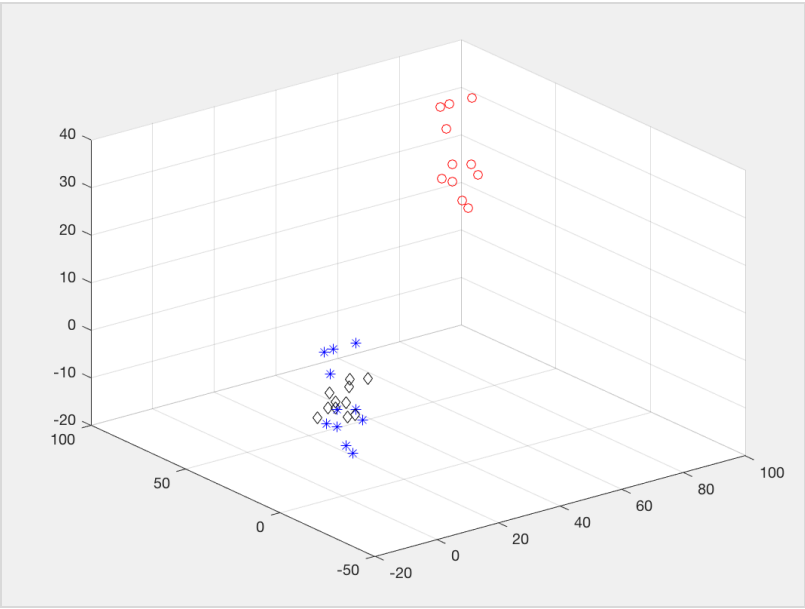
Transpose(Data) = FinalData =

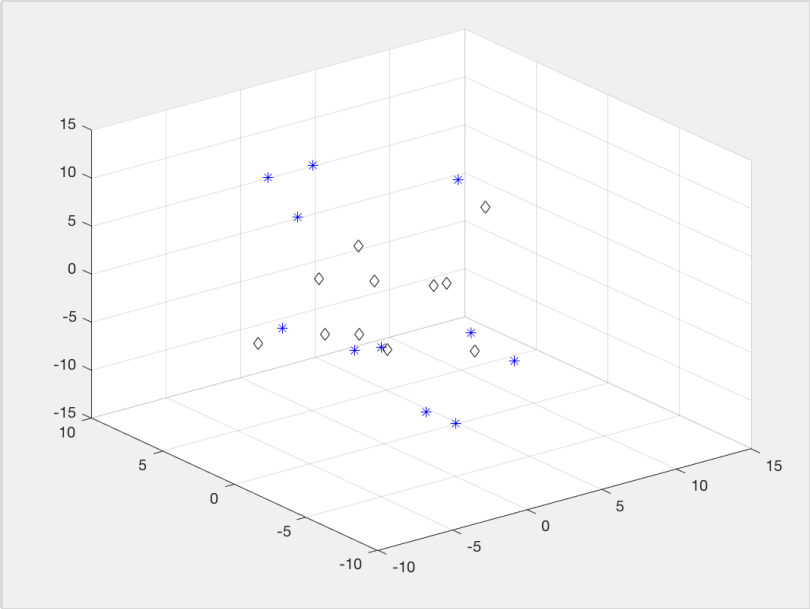
6.5569	8.6087	-3.1237
1.0049	-5.2925	-2.1580
6.4447	7.3843	-5.8602
-8.0279	-6.6320	4.2929
-5.4201	-1.5180	-0.2947
11.2111	4.5727	1.7802
1.9096	6.5105	-3.0908
-6.2587	-6.7722	2.0905
-0.7284	-5.1458	5.5214
-6.9993	1.4661	-2.6614
0.3072	-3.1818	3.5039

Below are graphs representing this transformation. The final plot is in black. The Original Data is in red. The mean adjusted data is blue.

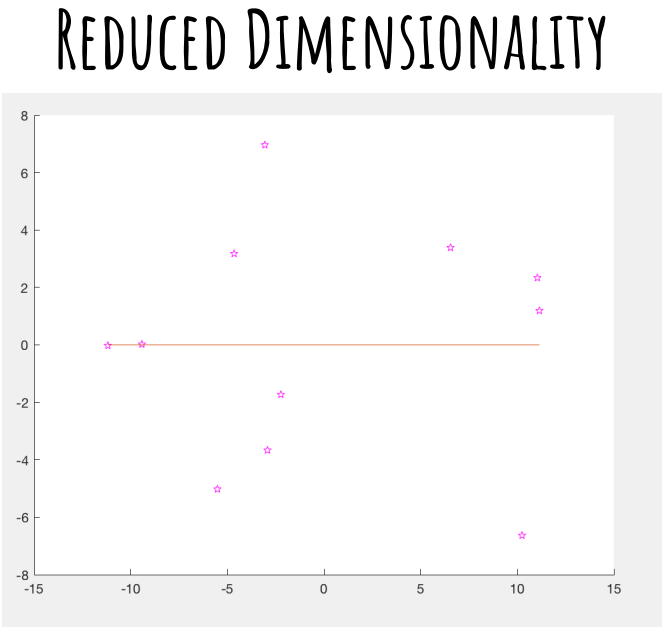
PCA Final Results - Plot



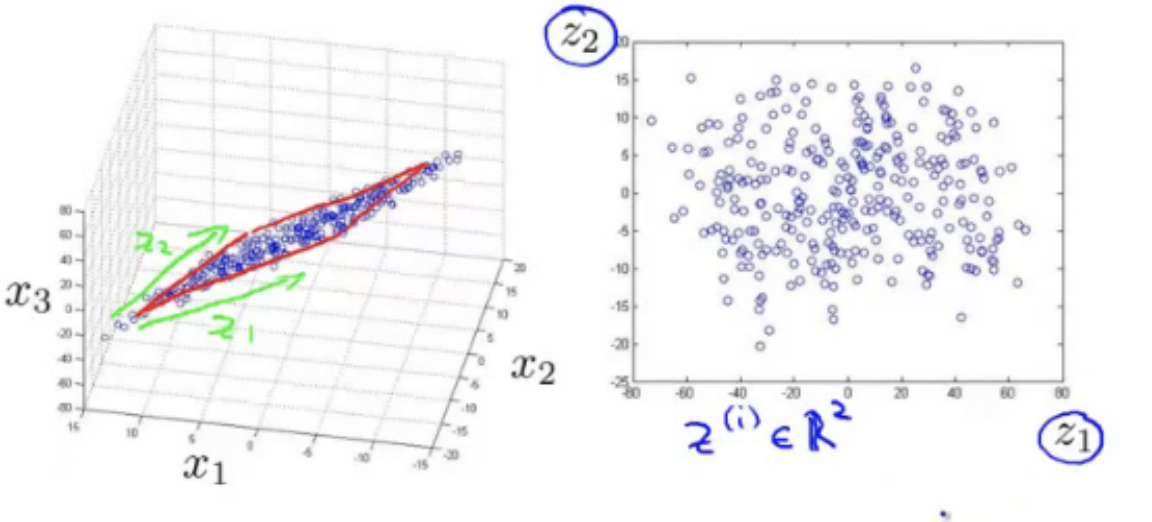




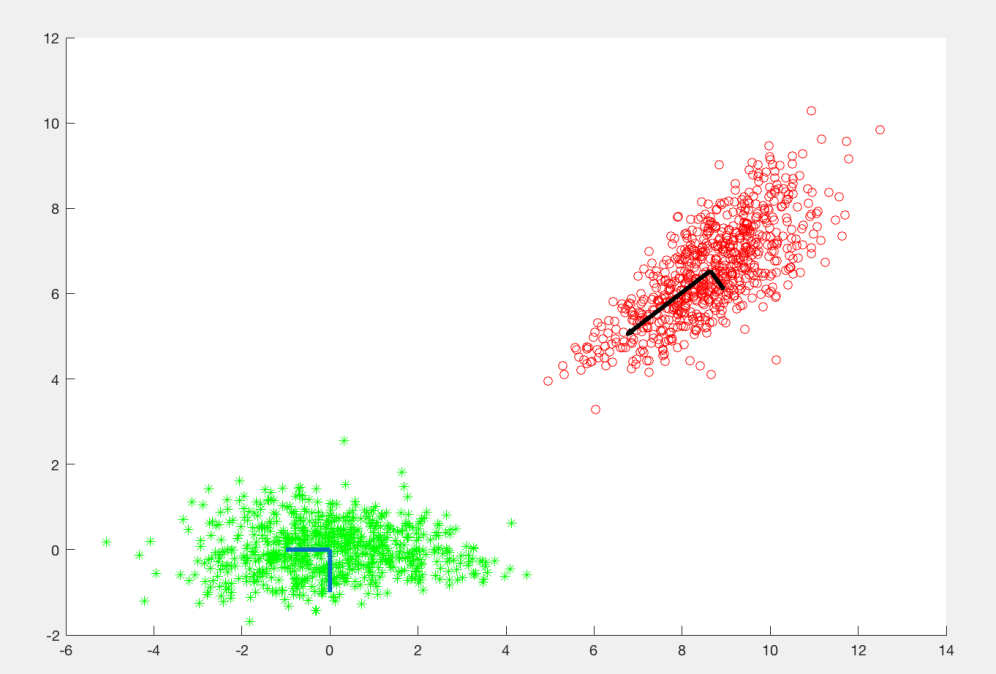
As I mentioned earlier a useful thing about PCA is that you can use it to reduce dimensionality. You can reduce the dimensionality of the graph/data by removing the direction of least variance, this is to ignore the eigenvector with the smallest eigenvalue. {reduced_data= meanadjusted*[V(:,3,V(:,2)]} Below is a graph from our example when we reduce the dimensionality from 3 dimensional to 2 dimensional. On this graph I included a linear regression line. This linear regression line has slope zero and lies on the x-axis. This shows that the data is now uncorrelated/the direction of most variance is along the x-axis.



Below is another picture that demonstrates reduced dimensionality. This can help in visualization as to why reducing the dimensionality can be helpful.



I was interested in principal component analysis so that I could use it to analyze information of the paleomorphogeological relationships regarding the shape of barrier islands. For this research I will use PCA to compare data of ancient and modern barrier islands. The paleomorphogeologists are interested in the shape of these island and what aspects affect the shape. By using this transformation they will be able to guess where petroleum deposits are located in ancient barrier islands. We do this by looking at the conditions under which the modern barrier islands are being formed and by comparing the shapes of the two data sets we can assume under which condition the ancient islands were formed.



Below is my code that I used for my barrier island PCA transformation. For the plot above 1,031 different barrier islands are analyzed. The original data is in red and the transformed data is in green. The lines drawn are the eigenvector with the greatest variance and the corresponding orthogonal vector.

```
test=[l,w];

test = test(test(:,1)~=0,:);
test = log(test);
meanTest = mean(test);

scatter(test(:,1),test(:,2),'ro')
hold on

[coeff,score,latent] = pca(test);
testcentered = score*coeff;
[U,S,V] = svd( test - repmat(meanTest,size(test,1),1) ,0 );

quiver(meanTest(1),meanTest(2),score(1,1),score(2,1),'k','Linewidth',3)
quiver(meanTest(1),meanTest(2),score(1,2),score(2,2),'k','Linewidth',3)


covariance = cov(testcentered);

[V,D] = eig(covariance)
EigenVectors = [V(:,2),V(:,1)]
RowFeatureVector = transpose(EigenVectors)
RowDataAdjusted = transpose(testcentered)
newdata = RowFeatureVector* RowDataAdjusted
finaldata = transpose(newdata)
scatter(finaldata(:,1),finaldata(:,2),'g*')
line([0, 0], [-1,0],'Linewidth',3);
line([-1,0], [0 0],'Linewidth',3);
```

Principal Component analysis is very useful for all types of research, as demonstrated by using it for the statistical analysis of barrier islands. However, the component of reducing dimensionality has many uses too including, but not limited to face recognition, handwritten digit recognition, text mining, image retrieval, image compression, and protein classification.