Manish Tripathi <mtripathi@pivotal.io>

Thursday Thunder: Week 3: Learning 1 & Question 3

Manish Tripathi <mtripathi@pivotal.io> Fri, Oct 30, 2015 at 10:47 AM
To: Pivotal Data Science - All <PDS-ALL@pivotal.io>

Disclaimer: This post is ultra long. So read at your own leisure

Hey

What good day to send an email than Halloween. I thought Thunder on Halloween might be more scary!!
Yeah another of my Poor Jokes.

Anyway so coming to the point. No one responded to my last week's question. So no one gets the coffee. And as promised, here is the answer to that.

"Q: For n << d case (think text data, genomics data etc.), what are the maximum no. of dimensions one can reduce the data to?."

>> When we have a normal data where **n >> d**, we know the max no. of principal components is equal to the dimensions of the dataset which is d. So isn't the same should be applicable for n<< d case?. Naah!!. If it was that easy I won't be asking the question. :D.

Right answer is that for centered data , one would not have more than **n-1** principal components or dimensions to reduce the data to. So that means if you have 1000 observations and 5000 dimensions, the maximum dimensions you can reduce your data to is going to be 1000-1=999. Why is that?.

Well it has to do with the no. of eigenvectors we have in the dataset.

Aaaaahhhhh! Again Eigen vectors. What the hell is this eigenvector?

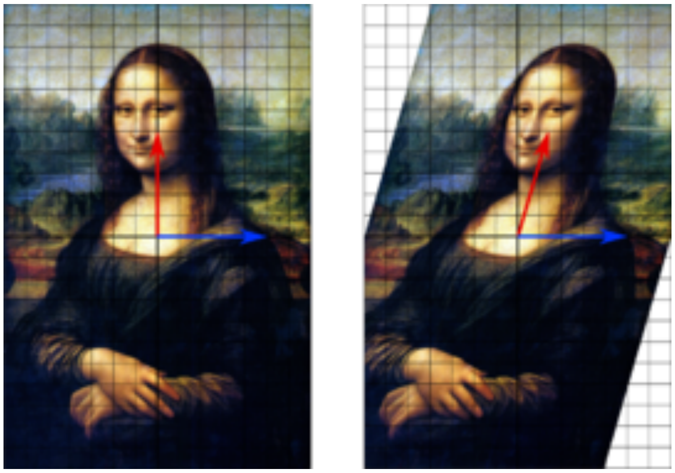
So lets talk what eigenvectors are:

Eigenvectors are the soul of your matrix. What does that mean.

Eigen in German means "self" or "own". Putting that it translate to Self-vector. Hmm..that's weird. What would that mean?.

Well so here is the thing. In Linear algebra 101 class (I never got to take one in my undergrad or even grad school :-( ) we know that a Matrix -Vector multiplication is a linear transformation. What it means is that if you multiply a vector with a matrix , it's direction and scale will change. But Eigen-vectors are those stubborn vectors (just like me) which don't change their direction even after a linear transformation is applied to them. They only scale. They retain their "self"- identity. Here is a diagram to explain that:

To compare both normal and eigen vectors together see below:



The red vector is your normal vector. When applied a linear transformation of a matrix it changes it's direction. But the blue is that stubborn vector which remains in the same direction. This is the eigenvector.

So obvious question now arises what's the use of such vectors which doesn't listen like everyone else and not willing to change?. The reason is Eigenvectors define a basis for that vector space.

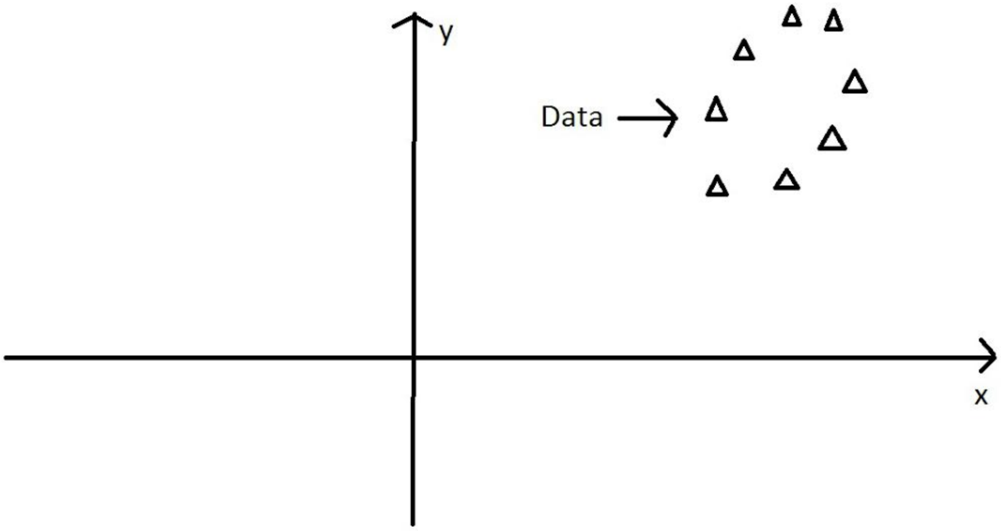
Aaaahhh.... What crap is this Basis now?. I will surely die now :(

I would not delve deep into what basis means, but to give a gist, basis are the set of vectors which are linearly independent of each other and using those set of vectors we can create any other vector in that vector space(span). The normal x, y, z co-ordinate system we have been using is one such basis system. It is called Canonical Basis. Similarly we have infinite such co-ordinate systems. Eigenvectors represent one such co-ordinate system of the underlying data.

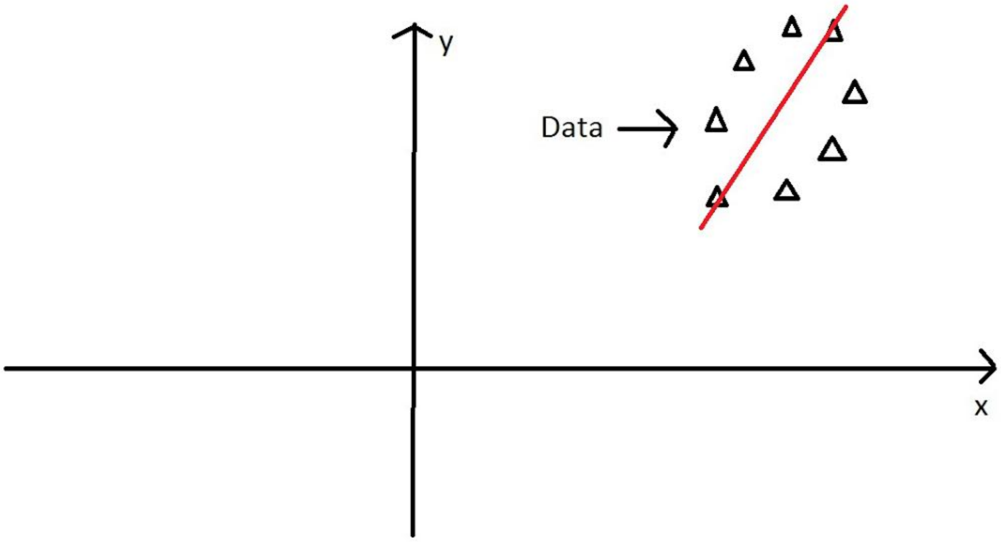
Ok. So looks like I am getting it a bit now. So where does this all lead to in PCA?.

Well if you see PCA, it is nothing but a change of basis of your data. Basically what PCA is doing is it is finding that camera angle by standing in the midst of your data points such data it is able to capture the right angle where the data spread is large. Here is an understanding

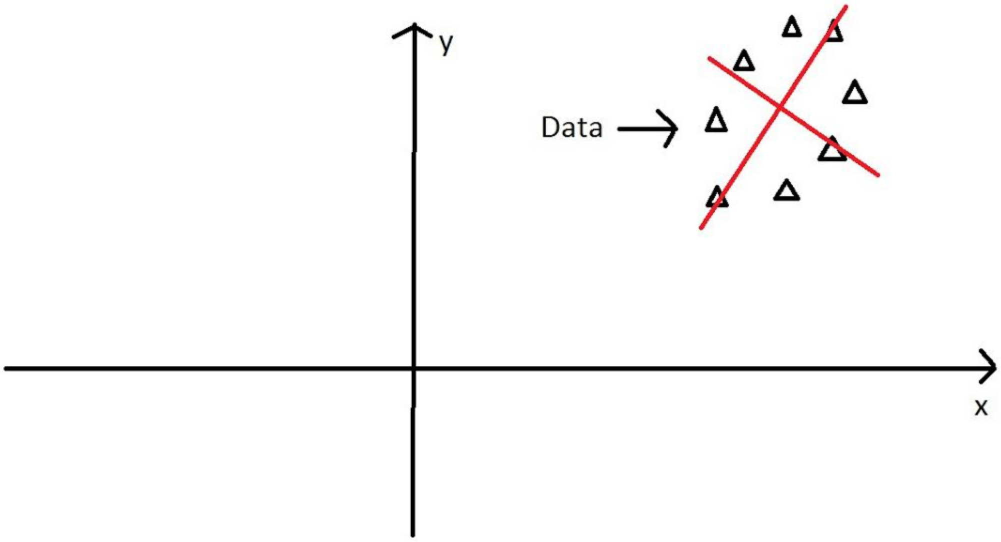
Lets say below is your datapoints in the original X, Y co-ordinate system (Canonical Basis).



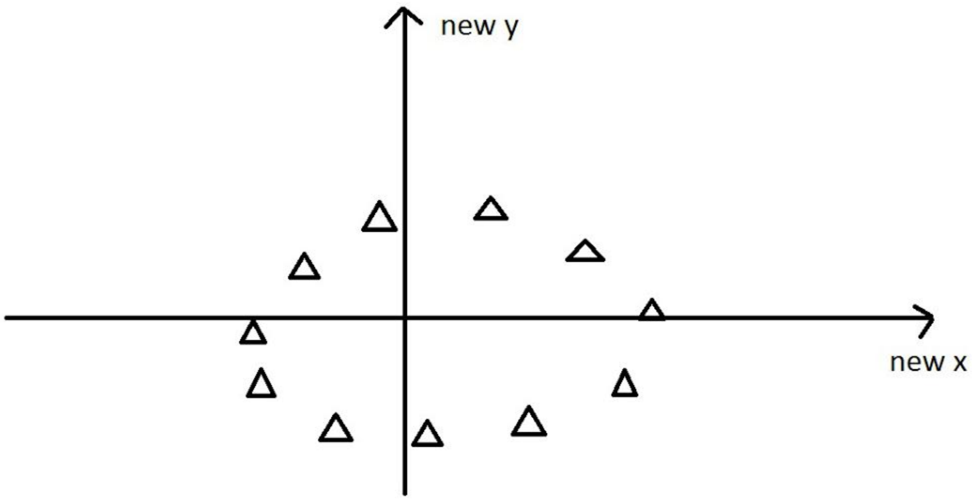
Now imagine you are holding a camera in sitting right in the middle of these points. You want to search for an angle to capture the image of these points such that there is maximum spread. You find the first angle as below.



Now you want to capture another angle so that you are able to capture maximum spread of the whole data. The best direction would be the direction orthogonal to this first red line. Why?. Because orthogonality would make it for a new co-ordinate system where both the new vectors are linearly independent (as they are orthogonal) of each other and hence form a Basis(remember definition of basis I shared above).



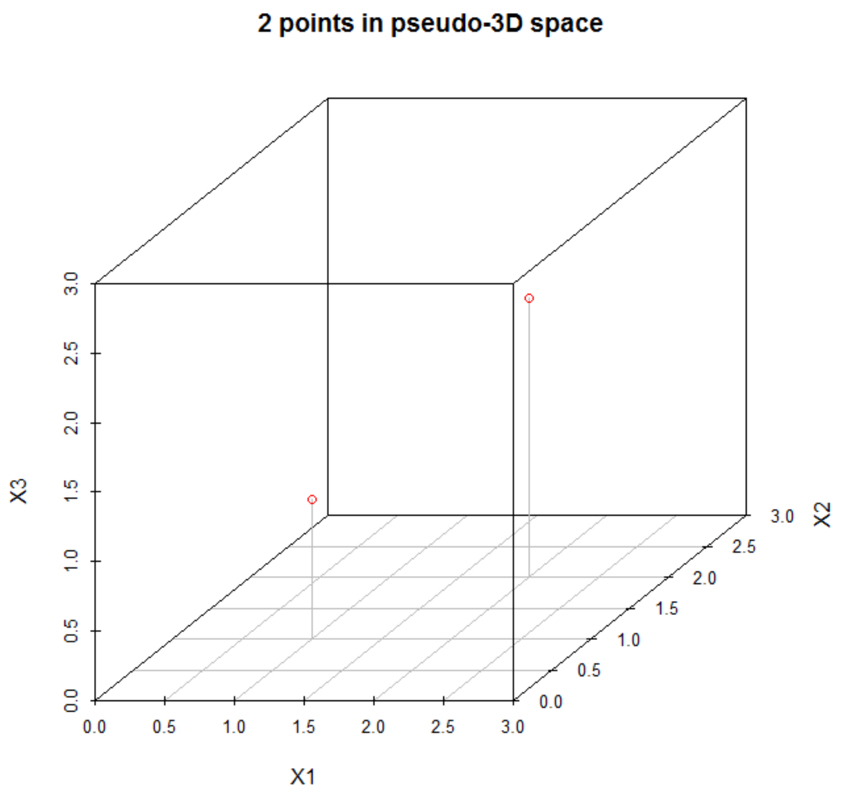
Now what you would just do is, simply rotate your original X, Y axis to the above red lines (new axis). Your transformed data now looks like this.



If you see the new transformed data (which is simply rotated with respect to new co-ordinate system) has zero correlation. This new X and new Y are the eigenvectors which form a new Basis for this new co-ordinate system.

Ok, Mr. Tripathi. I think I get this a bit now. But how does that relate to the question which was asked.

Well if you understand the eigenvectors and basis, then it's intuitive to understand why the no. of dimensions one can reduce the data when **d>>n** is **n-1** and not more. Here is an example:



If you see the above diagram, what is the best direction where the data is spread maximum?. Its diagonally left bottom to upper right crossing both the points. What is the other best direction?. There is none. As there is no other direction where you have the spread of the data. So if you see, 2 points with 3 dimensions just have 1 direction where the data is spread. You can put any of the point anywhere within this cube, there will still be only one straight line connecting them and that is the only direction where the data is spread.

To think other way, eigenvector of a matrix are those linearly independent set of columns which form a basis for a new co-ordinate system. What are the linearly independent set of columns called in a matrix jargon?. Its the **rank** of the matrix. And we know row rank of a matrix= column rank of the matrix. So even if your data had 5000 dimensions, and 1000 rows, the rank of the matrix would not exceed 1000 as the row rank would be not more than that.

Phew!.. Finally you are done with your rant. Good Riddance!!

[Quoted text hidden]