# The effect of the centering for PCA and whitening

Aodong Li

December 2, 2017

**This document only serves as informal personal thoughts.**

The range of this post is within descriptive statistics, where we do not assume any data generation process probabilistically.

## Relation between the rank and eigenvalues of a matrix

Suppose we have a $n \times n$ square matrix $A$. If $A$ has eigenvalue of zero, then, by definition,

$$A\mathbf{x} = \mathbf{0}$$

It means the Null space of $A$ is the eigenspace $W$ with 0 as one of the eigenvalues, so we have

$$dim\ NulA = dim\ W$$

$$rank\ A + dim\ W = n$$

Also, if $A$ has eigenvalue 0, then the columns of $A$ are linear dependent.

## What does it mean when we talk about centering in PCA?

PCA is the technique built on the eigendecomposition of *empirical* covariance matrix (*or not*). The empirical covariance matrix is always the same whether we do centering or not, because the computation of covariance matrix implicitly employs centering. See the empirical covariance matrix definition below.

$$\bar{\boldsymbol{x}} = av(\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}) \tag{1}$$

$$\Sigma = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^\top \tag{2}$$

Now, because the average of all features are zero and we can compute the $n^{th}$ data by the first $n-1$ data, the covariance matrix $\Sigma$ has the rank at most $n-1$ (linearly dependent).

*Thus, when we use empirical covariance matrix to compute the direction of the maximum empirical variance in the data, we always get 0 as one of the eigenvalues of* $\Sigma$.

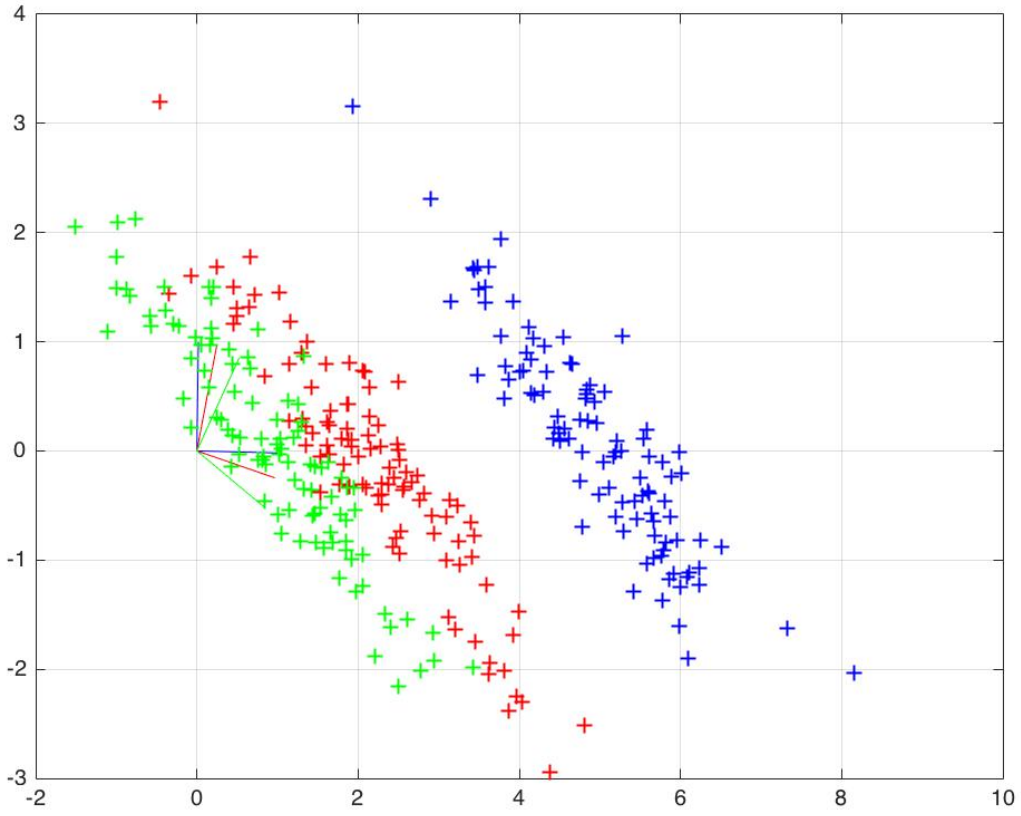
All above statements or phenomena are related to centering.

The uncentering actually means direct eigendecomposition of the matrix $XX^\top$ without subtracting the empirical mean, where $X$ is a $p \times n$ matrix with each data located on different columns. As for the PCA optimization problem, we are doing

$$\max \ \mathbf{u}^\top XX^\top \mathbf{u} \tag{3}$$

$$s.t. \ ||\mathbf{u}|| = 1 \tag{4}$$

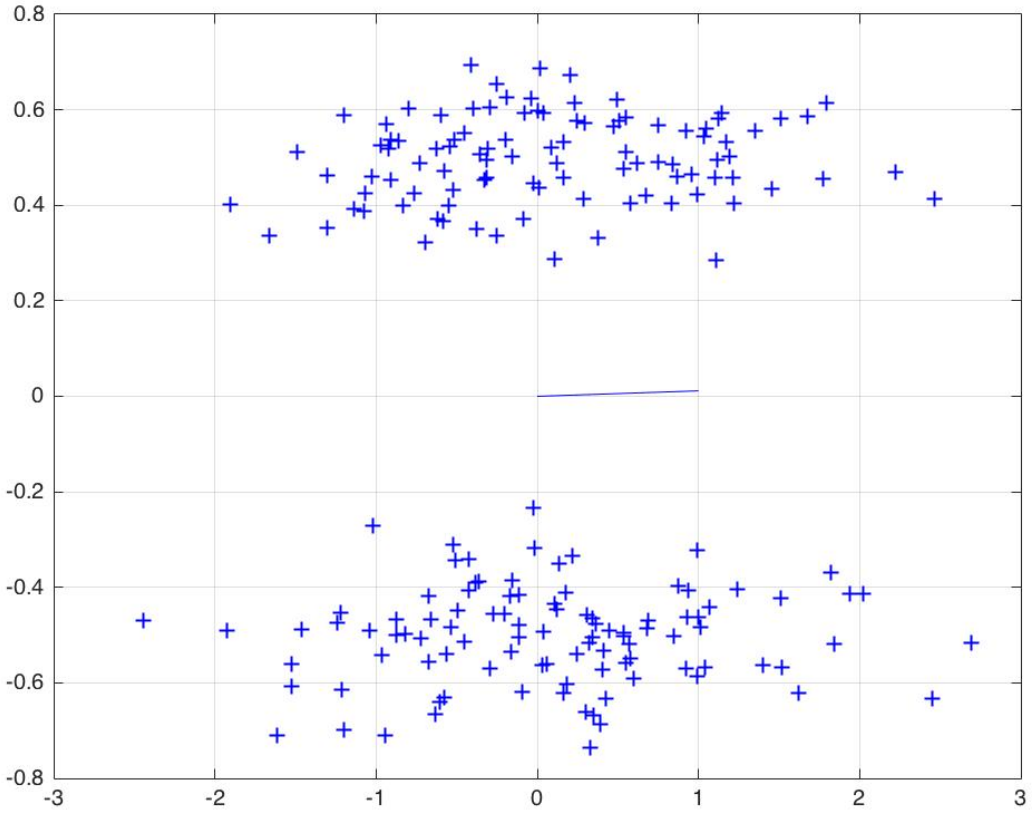
If we solve this problem (through quadratic form or power method), the first eigenvector roughly points to the center of the data, which makes senses since more energy is captured along that direction. See the figure below. As the cloud of the data reaches the origin, the eigenvector captures the variance better.




## PCA *doesn't* help on classification task

When we utilize PCA to reduce the dimension of the dataset, we rely on the motivation that the direction with higher variance is more informative and could capture the rough structure of the data *in most cases*. But PCA doesn't guarantee that the dimension-reduced dataset can be classified better. The main reason is that PCA is agnostic to classes, i.e., it treats data of different classes blind to the underlying distribution (PCA is one method of descriptive statistics!).

The following figure shows that if we project the dataset onto the first principle component, then the two clouds of data are messed up and no longer separable.




## The centering on high-dimentional data (NEED FURTHER TEST!)

The centering renders the same effect on high-dimensional data – at most $n-1$-rank empirical covariance matrix. But the uncentering has less influence on high-dimensional data.

It is true that the first direction roughly points to the center of the data, but the bias will only affects the first two to three directions or so (remember the directions are orthogonal) and the other directions work as normal.

The take-away is that for high-dimensional data, we can simply use the eigendecomposition of $XX^\top$ to get rid of the singular empirical covariance matrix. The premise of all these statements is that we have enough data to fill up the vector space.

## For whitening (NEED FURTHER TEST!)

If we use the empirical covariance matrix for whitening, then we need to leave out of the final 0 eigenvalue and its corresponding eigenvectors, because this may simply explode resulting

data in the orthogonal direction of the space where the data scatters. On the other hand, if we use the eigendecomposition of $XX^\top$, we can leave out of these worries.