To scale or not, before PCA

PCA represents the data in a new coordinate system so the first direction/dimension has the maximum variance, with maximum information.

If one variable, v_1 , is in a scale that varies from -10M to 10M, like distance between cities in meters, and another, v_2, from -1 to 1, then just due to scale difference, variance of v_1 would be larger, and picked by PCA process.

Hence, one should do the standardization before applying PCA.

(standardization keeps correlations fixed, mix-max scaling changes it.)

Let's take a look:

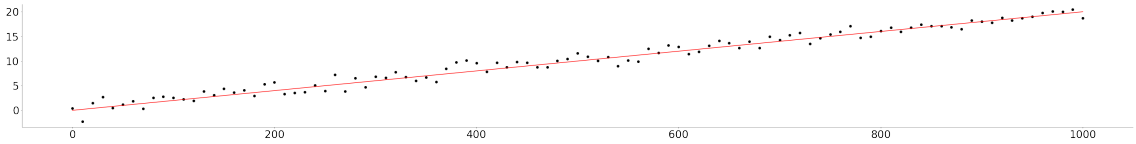

Figure 1

In Fig. 1, the range in the x-direction is 1000 and in the y-direction is 20, so, variance in x-direction will be larger as well. Hence we normalize. (Fig. 2)

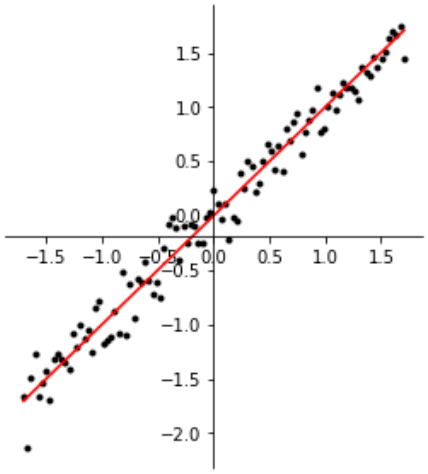

Figure 2

Now the data are in almost the same range, we rotate the coordinate system so first axis is in the direction of the derived variable with largest variance and second will be in the direction of second largest orthogonal to previous one.

In the above plot the red line would be the first PCA, let me call it PC_1 direction and a line orthogonal to it will be the second direction, PC_2.

However, even after mapping the data to PCA space (Fig. 3), in the new coordinate system you can see, the new variables will be in different scales. The data in direction of the PC_1 has larger variance as opposed to the PC_2.

Hence, if we want to compute distances in the PCA space, the distances will be affected by such a difference in scale. So, we have to do another scaling, after the PCA step.

In conclusion

- The scaling done in the first step lets PCA does its job correctly.
- The scaling in the last step makes the computed distances to not be affected by difference of scales

Representing the data in PCA space kills the correlation among variables, hence the covariance in PCA space is diagonal:

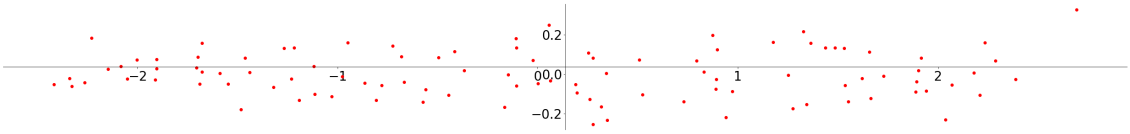

Figure 3

And the second scaling kills the bad effect of different scales in computing distances.

The three sptes described above, scaling, PCA step, scaling, is equivalent to *Mahalanobis* distance. It kills the effect of scales and correlations among data.

To see the proof of the equivalency please take a look here.