# How Robust is Principal Component Analysis? Not Robust at All

# Written on May 16, 2019 by Andreas Steiner

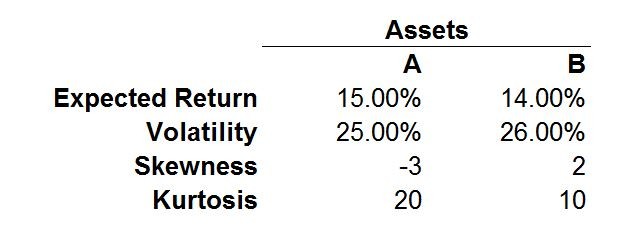
Principal Component Analysis (PCA) is used for all sorts of purposes in quantitative finance. Especially in portfolio construction, it is a very popular statistical technique to manage "Big Data" (i.e. dimensionality issues) and to manage dependency between assets (i.e. one determinate of the resulting portfolio risk).

Most of the time, PCA is used rather naively, mainly by applying established algorithms for example in MATLAB (tm) or other computing environments like for example R. Applying typically means executing a built-in function or particular function module. The PCA results are then further processed, for example, by turning them into statistical factors and feeding them to a mean-variance optimizer. Quality statistics like for example an R-squared in linear regression analysis are typically not used for PCA results. Users also tend to ignore the assumptions underlying PCA. One of these assumptions being that the underlying data follows a multivariate Gaussian distribution.

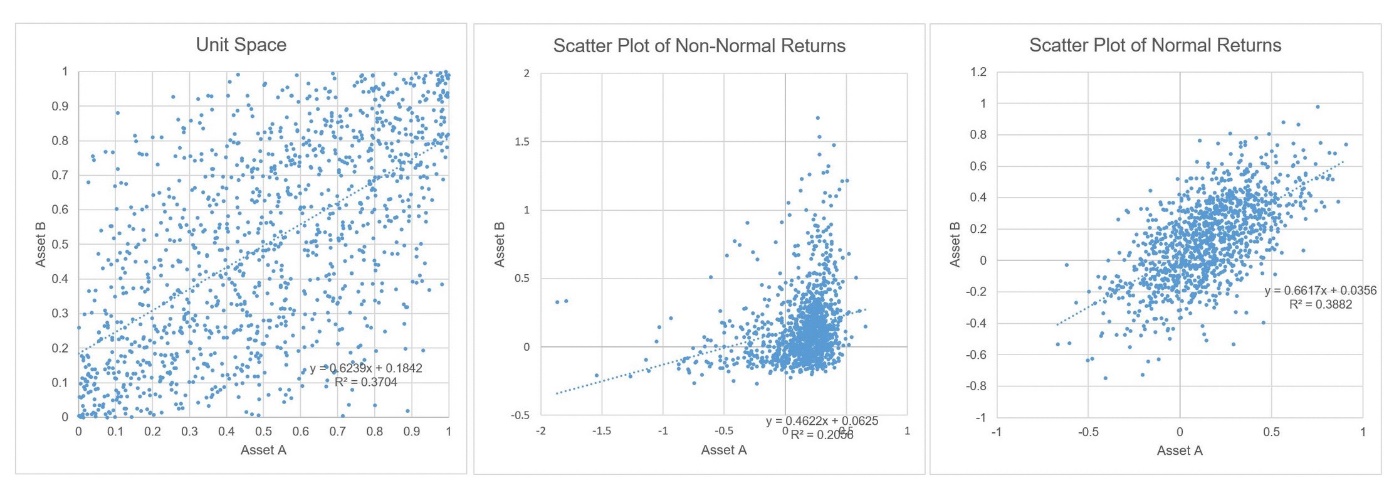
But then, it is rather well known that empirical financial data deviates from the normal distribution in important ways, namely that the left tail which represents "risk" is longer and fatter than the tail of the Gaussian distribution. Additionally, empirical tail dependency is typically higher than what the multivariate normal distribution implies (a so-called Gaussian copula).

In order to illustrate what can happen to PCA results when using real-world data, we have conducted a very simple simulation experiment using two theoretical assets. We assume that the assets are dependent as implied by a bivariate normal distribution in which the assets are correlated 60%.

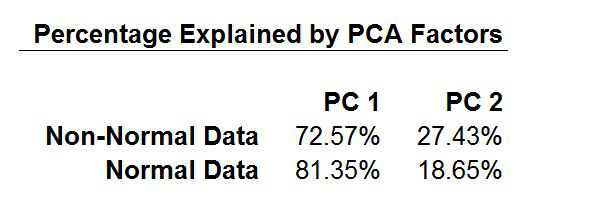
Using standard techniques from copula theory, we simulate 1) asset returns following a normal distribution and 2) assets returns with equal first and second moments. But non-zero skewness and kurtosis (i.e. non-normality)...



Note that the underlying dependency structure is assumed to be Gaussian in both simulations. We summarize the results of the simulations in three scatter plots, one showing the simulated dependency in unit space and the other two the assumptions regarding the marginal distribution parameters shown above...



If we now run a PCA across the normal and non-normal datasets, we can compare the percentage contributions of the first and second principal components (i.e. underlying statistical factors)...



The lower percentage figure of the first principal component in the case of the non-normal data is not a coincidence. It would be interpreted as providing "more diversification potential" as the higher value in the normal dataset. This is clearly a dangerous and misleading conclusion: the only thing that changed is the tail risk characteristics of the assets.

PCA is not a robust statistical technique, its results can be distorted by non-normal asset risk and non-Gaussian dependency characteristics. Maybe the next time somebody presents results based on PCA techniques, you might raise a hand and ask a question. Or two.

PCA requires the underlying data to be (multivariate) normal distributed. ICA; on the other hand, requires pronounced non-normality, otherwise, it will not work. The problem with financial data is that it is non-normal many times, but not too, too far away from normal (I wrote about this recently - see my article on "degree of normality"). If anything, I would replace PCA with something like Nelsin/Siegel/Svensson (pretty much best practice in yield curve modelling nowadays) or then very simple yield curve factors like for example in the very old Lehman model (although simple, they are 99% correlated with PCA factors).

Peter, Alexander is doing risk analysis only. Also, the use of PCA in risk analysis is nothing new (and not really exciting, because the results are hardly ever a new story). The more interesting question that I am trying to touch in this article: linking risk factors to total returns, i.e. treating them as priced risk factors. While risk modelling is always a topic, the artificial gap between models of risk and return is another major issue in the industry. If anything, my inspiration was John H. Cochrane's work related to the yield curve.