Using principal component regression to reconstruct historical ETF's prices

Dr. Stéphane Sallé
Université Paris Dauphine – PSL, MBA Management, Risques & Contrôle

1. Introduction

ETFs (Exchange-Traded Funds) have experienced exponential growth since their inception in the 1990s. By the end of December 2024, the number of ETFs (over 13,000) and assets under management (around 15 TUSD) had exploded, representing a more than 30-fold increase since 2000. These products are financial assets that are increasingly used by both professional and individual investors, with new offers every months.

However, the products on the market are often recent (less than 10 years old), or have undergone mergers that have eliminated historical data. This can be a problem when you want to analyze performance over a long period, or assess the impact of crises that have occurred in the last 28 years. The author has used the principal component regression method, an extension of classical linear regression, to reconstruct financial asset prices over missing historical periods. By comparing the results obtained by the two methods, he highlights the interest of analyzing the eigenvalues of the information matrix in order to truncate its spectrum and thus reduce the overfitting phenomenon well known to data scientists.

The paper will be structured in four parts, focusing more on parameter identification than on financial aspects. First, we'll take a look back at the modeling of stock market prices for financial assets using lognormal distribution, and highlight the very high levels of noise that can be observed over a long period. We will then present the classic method of linear regression with the main world stock market indices, known as the Arbitrage Pricing Theory (APT) method. We continue with a description of the Principal Component Regression (PCR) method and some results obtained on eigenvectors and eigenvalues. The fourth part will show the results obtained on the reconstruction of the history of indexed and non-indexed ETFs and listed stocks. We'll also look at cases where the approach presented works well and where it doesn't. We will conclude this presentation with potential future work and a presentation on a public site of some real implementation cases.

2. Some empirical feedback on modeling the listed share's return

In this section, we present the model based on the lognormal distribution (Sharpe, 1988), which is the most widely used to approximate the returns of a financial asset, and which is also used to value derivatives via the Black and Scholes model. It assumes that the natural logarithm of a listed share's return follows a normal distribution. There are several known flaws in this model, including thick tails (i.e. large losses are more frequent than predicted by the distribution) and high heads (returns around the mean are more frequent than predicted by the model).

We will present this observation on a large number of assets, and invite listeners to do the same on real data presented on the www.adaminform.com website. So we will show that this lognormal distribution hypothesis is relevant for a large number of assets if the analysis is carried out over a sufficiently long history (in this case, 28 years with weekly returns). We will also present the results of the test of independence between the annual returns of these assets. These findings on real data will enable us to assess the level of noise observable in this type of system.

3. Implementation of APT and PCR methods

The Arbitrage Pricing Theory (APT) method (Ross, 1976) aims to predict the market price of a financial asset as a function of macro-economic or stock market indices. It consists of a linear regression over a time horizon between the valuations y(t) of this financial asset and the values of these n indices $\phi_i(t)$. If the parameters of this linear regression are represented by the n-dimensional vector θ and the index values at time t are in the n-dimensional vector $\phi^T(t)$, the least-squares estimate of θ is derived from the inversion of the observation matrix $\Phi^T \Phi : \widehat{\theta} = (\Phi^T \Phi)^{-1} \Phi^T Y$ (Åström & Wittenmark, 1997).

We have applied this method to financial assets, taking the main stock market indices (up to 25) to which the financial asset belongs. In the case of an index ETF, the values of the replicated index give a good idea of the values and relative variations to be found in the reconstructed past. In the case of some ETFs, we find that the values obtained may be far removed from the values of the replicated index. This phenomenon corresponds to the well-known parametric estimation problem of overfitting.

The Principal Component Regression (PCR) method (Schmidli, 2013) uses the eigenvalues of the observation matrix $\Phi^T \Phi$. Removing the smallest eigenvalues reduces these discrepancies, giving a historical estimate that is more faithful to the asset's economic reality.

4. Achieved performance

We'll present the performance obtained on an extended set of ETF (more than 100). We begin by explaining how we evaluate performance. Then we'll present the overall results obtained with some ETF for which both APT and PCR methods work well, some ETF for which only the PCR works well and some for which none of them give a satisfactory result.
In the second case, we'll expose some rules of thumbs that allowed us to fix the number of eigen values to be used. We'll also expose some surprises in the obtained results that will need further investigations and discussions.

5. Conclusion and potential follow-up

In view of the performance obtained, we believe that this approach is relevant for estimating past values of some financial ETF and other financial assets. These values may be used for stress test purposes of for back-testing of some investing strategies.
The potential applications of this work are many and varied. We have identified three, which we will expose at the end of this presentation.

Bibliography

Åström Karl-Johan & Wittenmark Björn (1997). *Computer-Controlled Systems. Theory and Design, 3rd edition.* Prentice Hall.

Ross Stephen A (1976). *The arbitrage theory of capital asset pricing*. Journal of Economic Theory. 13 (3): 341–360

Schmidli, Heinz (2013). *Reduced Rank Regression: With Applications to Quantitative Structure-Activity Relationships*. Springer.

Sharpe Michael (1988). *General Theory of Markov Processes*, Academic Press.