Optimizing a basket against the efficient market hypothesis

Frédéric Abergel, Mauro Politi

Optimizing a basket against the efficient market hypothesis

Frédéric Abergel*† and Mauro Politi‡†§¶

August 20, 2012

Abstract

The possibility that the collective dynamics of a set of stocks could lead to a specific basket violating the efficient market hypothesis is investigated. Precisely, we show that it is systematically possible to form a basket with a non-trivial autocorrelation structure when the examined time scales are of the order of tens of seconds. Moreover, we show that this situation is persistent enough to allow some kind of forecasting.
Keywords:Market efficiency; Lead-lag correlation; Optimization.

*frederic.abergel@ecp.fr

†Chaire de finance quantitative, Laboratoire MAS, Grande voie des vignes, École Centrale Paris, F-92290 Châtenay-Malabry, France.

‡mauro.politi@google.com

§SSRI & Department of Economics and Business, International Christian University, 3-10-2 Osawa, Mitaka, Tokyo, 181-8585 Japan

¶Basque Center for Applied Mathematics, Bizkaia Technology Park, Building 500, E48160, Derio, Spain.

1 Introduction

The Efficient Market Hypothesis, in short EMH, is technically interpreted to mean that the market is impossible to beat and that there are no autocorrelation (such as systematically repeated price/returns patterns) that can be used for profit [**?**]. The idea is a pillar and a paradigm of classical market finance and it is widely discussed in the influential review of E. Fama in Ref. [**?**], dated 1970. A Markov market, such that future fluctuations depend only on the last observed price, satisfies the condition of a market that is strictly impossible to beat and thus, perfectly efficient. In the past decade, with advancements in trading technologies, it has become clear that the EMH can be violated; a comprehensive survey and discussion can be found in Ref. [**?**]. However, real markets remain very hard to beat and models that generate no autocorrelation in the price increments are very good first step approximations. In such models, in the limit of small time intervals, the correlations of the increments $x(t_1) - x(t_1 - T_1)$ and $x(t_2 + T_2) - x(t_2)$ for a generic price series $x(t)$ vanish

$$\langle (x(t_1) - x(t_1 - T_1))(x(t_2 + T_2) - x(t_2)) \rangle = 0 \tag{1}$$

for each set of finite instants t_1, T_1, t_2, T_1 if there is no overlap in the intervals, i.e. if $(t_1 - T_1, t_1) \cap (t_2, t_2 + T_2) = \emptyset$. This condition is much weaker, and more pregnant, than the assumption of stochastically independent increments. Eq. (1) means that no event that took place during a past time interval can be used to systematically forecast the returns in a future time interval, at least at the level of simple averages and pair correlations. That is, the market is "effectively efficient" in the sense that it is impossible to forecast its direction, paving the way for the application of semi-martingale processes. Contrarily to some other well-known stochastic processes such as the fractional Brownian motion, there is no memory in pair correlations to be exploited. In general, this does not rule out higher-order correlations or cross-correlations between assets that might be used for technical trading. A Markovian market is "efficient" in the strictest sense: it is impossible to beat. Instead, a martingale market leaves the opportunity of exploiting high order dependencies.

It is an acknowledged, tested and recognized fact that individual asset time series do not present features challenging the efficiency - were it not for mean-reverting effects at very short time scales (some seconds), mainly due to the bid-ask bounce. The main question we are addressing in this work is whether interdependences between distinct stock price increments can be employed to produce a single time series presenting a non-vanishing, statistically significant level of autocorrelation. Basically this is not a new idea - in fact, the ideas beyond pair trading, i.e. co-integration [**?**], and most of the statistical arbitrage strategies, are simpler versions of this one. The originality of the work we present here lies, firstly, in the focus on short time scales (of the order of tens of seconds) and, secondly and more important, in the design of an optimized collective dynamics. Precisely, starting with an original set of N stock time series $x_i(t)$ $(i = 1, \ldots, N)$, we are asking whether it is possible to find a suitable set of weights w_i for which the basket

$$B(t) = \sum_{i=1}^{N} w_i x_i(t) \tag{2}$$

presents systematic autocorrelations, or, more generally speaking, a non trivial persistent behaviour.

Unfortunately, assessing the persistent or anti-persistent behaviour of a time series raises various difficulties, and the vast confusion in the literature does not help to overcome them. Often, an

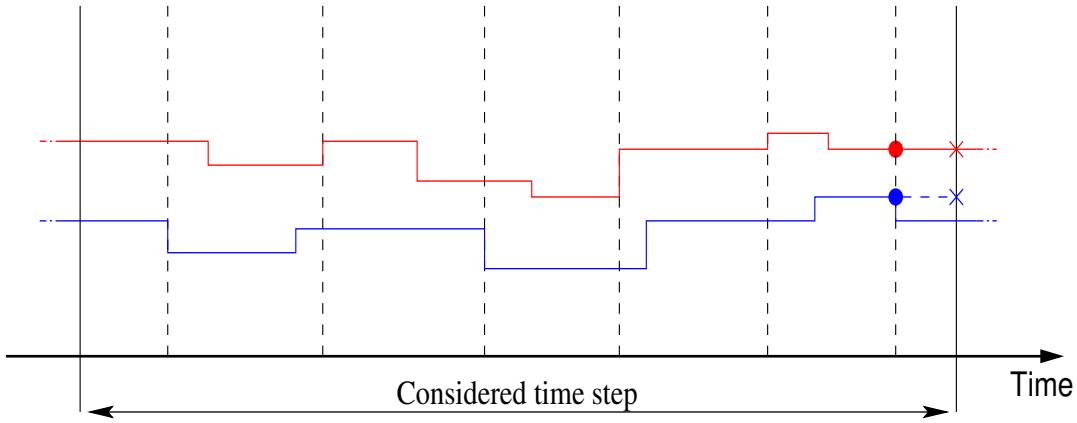

Figure 1: Scheme explaining the nature of the series we study and the procedure used to obtain them. For each time window, we observe the last trade (trades are dashed vertical lines, randomly spaced) and consider the bid and ask prices immediately before the transaction (blue and red large points). The blue line is the bid time series and the red one is the ask time series. The considered midprice is simply the average of the considered bid and ask prices.

estimation of the Hurst exponent is recognized as a good indicator, but this is misleading. Strictly speaking, the Hurst exponent is just a measure of the scaling properties of a time series and, alone, does not give any information on the possible persistent behaviour when the analyzed data are not stationary, as is typically the case in the financial world. Moreover, the estimation of the Hurst exponent is by itself a difficult task, and there is a host of available estimators which seldom allow for cross-validations (see, for example, Ref. [**?**]).

To avoid these problems, or at least to minimize them, we will mainly focus our discussion directly on the autocorrelation, or, more precisely, on the anti-autocorrelation. This choice will help us overcome the tedious and not always solvable issues of considering trends, leaving us only with the problem of non-stationarity in higher order fluctuations. This could still has an effect on the value of the sample autocorrelation, but we will not focus on that value itself. Rather, we simply seek to assess whether it is significant or not, using tests based on the stationarity of the increments. A consistent way to estimate the statistical significance for non-stationary time series is absent, at least to the best of the authors' knowledge.

The paper is organized as follow. In section 2 we present the dataset we use for the study, together with the sampling rules. In section 3 the main statistical objects (autocorrelation and Hurst exponent) are presented and discussed. Section 4 is devoted to the explanation of the tools and procedures we follow looking for the desired non-efficient basket. Finally, Sections 5 and 6 are dedicated respectively to the results analysis and the conclusions.

2 Data set

The original data set consists of all trades registered in the primary markets of the analyzed stocks. The data are stored in the Thomson Reuters TRTH data base made available to the Chair

of Quantitative Finance by Thomson Reuters and BNP Paribas. For the purpose of our study, we extract from the TRTH database records consisting in the time of a transaction, the best bid and best ask prices prior to each transaction, and the traded price. These data, appropriately filtered in order to remove misprints in prices and times of execution, correspond to the trades registered on the NYSE or NASDAQ during 2007, for the 30 shares of the Dow Jones Industrial Average Index, namely, at that time: AA.N AIG.N AXP.N BA.N C.N CAT.N CCE.N DD.N DIS.N GE.N GM.N HD.N HON.N HPQ.N IBM.N INTC.O JNJ.N JPM.N MCD.N MMM.N MO.N MRK.N MSFT.O PFE.N PG.N T.N UTX.N VZ.N WMT.N XOM.N.
28 of those companies were traded primarily on the NYSE. IBM.O and MSFT.O are the only two primarily traded on the NASDAQ. The full meaning of the "RIC" symbols is available from `www.thomsonreuters.com`. The choice of one year of data is a trade-off between the necessity of managing enough data for significant statistical analyses and the goal of minimizing the effect of strong macro-economic fluctuations. However, the consistency of the discussed results during extreme condition periods are beyond the purposes of the present paper, and are left for future studies.

Precisely, for each day the period we consider starts at 10.00 am and ends at 3.45 pm, leading to 4140 increments when considering fixed time steps of $5s$, which will be our basic time scale unit. The choice of restricting the considered periods only to the central part of the trading day, discarding the opening and closing period, is justified by the anomalies data often exhibit during these parts of the trading day: errors tend to occur more often during the first and last part of the continuous trading day, it often happens that some shares are open to trading several minutes after the others, due to potential issues during the opening auction. Moreover, as is well-known, the activity near the opening and closing period is much higher than during the central part of the trading day, adding a non-trivial problem to our study that is strictly based on physical time steps. One could try to work in event time, following a multidimensional approach to trading time that has been recently proposed and successfully used to establish some joint distributional properties of baskets of stocks, see [**?**], but we thought that the added complexity of this change of time would actually conceal rather than emphasize the idea we are putting forward in this work.

Having fixed a time window, we consider the mid-prices at the time of the last registered trades. Figure 1 provides a comprehensive description of the procedure. Finally, we would like to point out to the reader's attention the fact that we are not working with log-returns, but only with increments, as they tend to be a more natural choice in a high-frequency setting where the evolution of the price is essentially discrete due to the "tick" effect. We refer again to [**?**] for another, more statistical justification of this choice.

3 Autocorrelation

Several definitions of the sample autocorrelation coefficients have been proposed in the literature. We consider the most standard one (see Ref. [**?**] for a classical discussion and Ref. [**?**] for a modern one): given n observations of a discrete time series $y_1, \ldots, y_n$ the sample autocorrelation coefficient at lag k is given by

$$\hat{\rho}(k) = \frac{\sum_{j=1}^{n-k}(y_i - \bar{y})(y_{i+k} - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\,, \tag{3}$$

where $\bar{y} = (\sum_{i=1}^{n} y_i)/n$ is the sample mean, and $1 \leq k \leq n-1$. Under the hypothesis of having $y_1, \ldots, y_n$ i.i.d. with mean (resp. standard deviation) equal to 0 (resp. σ), the lag one ($k = 1$) autocorrelation can be rewritten as

$$\hat{\rho}(k) = \frac{\sum_{j=1}^{n-1} y_i y_{i+1}}{\sum_{i=1}^{n} y_i^2} \,. \tag{4}$$

The numerator has zero mean and a variance that can easily be calculated:

$$\begin{aligned} \mathrm{Var}\left[\sum_{j=1}^{n-1} y_i y_{i+1}\right] &= \mathbf{E}\left[\left(\sum_{j=1}^{n-1} y_i y_{i+1}\right)^2\right] \\ &= \mathbf{E}\left[\sum_{h=1}^{n-1}\sum_{j=1}^{n-1} y_i y_{i+1} y_h y_{h+1}\right] \quad = \mathbf{E}\left[\sum_{j=1}^{n-1} y_i^2 y_{i+1}^2\right] = (n-1)\sigma^4 \,. \end{aligned} \tag{5}$$

For large n, the classical Central Limit Theorem shows that $\sum_{j=1}^{n-1} y_i y_{i+1}$ is asymptotically normal, distributed as $N(0, (n-1)\sigma^4)$. The denominator, when divided by $(n-2)$, is an estimator of the variance. Therefore

$$\hat{\rho}(1) = \frac{\sum_{j=1}^{n-1} y_i y_{i+1}}{\sum_{i=1}^{n} y_i^2} \sim N\left(0, \frac{1}{n}\right) \,, \tag{6}$$

for $n >> 1$. For this reason, the 95% confidence interval of the lag one autocorrelation coefficient can be approximated by $\pm 2\sqrt{1/n}$.
Note that we will rely on this confidence level to provide a clearer picture of the obtained results, although, owing to the non-stationarity of our data, the reader may consider such confidence intervals with a somewhat suspicious mind.

3.1 Hurst exponent

The Hurst exponent is defined in the framework of fractional Brownian motion (fBm), a well-known gaussian stochastic process where the second-order moments of the increments scale as

$$\mathbf{E}\left[(x(t_2) - x(t_1))^2\right] \propto |t_2 - t_1|^{2H} \tag{7}$$

with $H \in [0, 1]$. Brownian motion corresponds to the particular case $H = 1/2$. If $H < 1/2$, the behaviour of the process is anti-persistent, that is, deviations of one sign are generally followed by deviations with the opposite sign, an effect that in Finance is usually called mean-reversion. The limiting case $H = 0$ corresponds to white noise, where fluctuations at all frequencies are equally present, or $1/f$-noise (pink noise), where the power spectral density is inversely proportional to the frequency. If $H > 1/2$, the behaviour of the process is persistent, i.e., successive increments tend to have the same sign, a phenomenon termed trend-following in the Finance literature. The limiting case $H = 1$ corresponds to the behaviour of a Lipschitz-continuous function.

As pointed out in the introduction, the correspondence between the values of H and the behaviours described above is true only in the particular framework of the fBm. Otherwise, an estimation of H is only an indicator of the scaling properties of the time series [**?**]. As a limit

example, one can consider the case of the stable Lévy processes. A stability index α leads to a Hurst exponent equal to $\max(1, 1/\alpha)$, but there is no sign of persistent behaviour. Therefore, we will not rely on the mere estimation of the Hurst exponent to assess whether our final time series is efficient or not, but we will rather use it as an indicator of a possibly special behaviour.

4 The procedure

We now describe our method to find the suitable weights w_i leading to the desired "non-efficient" basket. Since we have decided to focus our attention on negative autocorrelation, we will start looking for the basket that minimize this statistic over a fixed period of time. More precisely, we consider a number of consecutive days D and we look for the set of weights producing B_D, the basket with the minimum value of autocorrelation possibly obtainable given the recorded data. Please beware that this is an *a posteriori* evaluation: we first observe the data, and then choose the weights. At this stage, the efficiency of the market is not questioned. The efficiency is threatend only if we are able to provide the weights of a negatively autocorrelated basket without using the information of the data we are testing. In order to do so, immediately after the D consecutive days, that we call *minimization period*, we pick S consecutive days in which we perform an out-of-sample test of B_D. If the value of the anti-autocorrelation remains significantly low during those S days, we now have evidence for non-efficiency.

We can choose among a plethora of methods to optimize the property we are interested in, see [**?**, **?**]. For the sake of simplicity, we will use the autocorrelation matrix. This approach presents some difficulties: for example we are working with scales in which the Epps effect is strong [**?**, **?**], and we will have to estimate $N \times N$ parameters, each one effected by a measurement error. Aside from these purely technical aspects, one could also define other measures of mean-reversion, such as for instance the sign correlation, but the matrix approach would not work in its usual formulation, and we prefer to retain its computational simplicity.

The classical correlation matrix approach to collective behaviour involve equal time correlations. Instead, the construction of a delay correlation matrix (or time-lagged correlation matrix) involves calculating correlations between different assets fixing a time delay [**?**, **?**]. Let us then define a matrix $\mathbf{M}$ of order $N \times T$ where each row is filled by the records of a discretized time series, and N is the number of time series of length T. Suppose that the indices i and j correspond to two distinct assets among the given time series. The delay correlation matrices between asset i and asset j at time lag k is given by

$$\hat{\mathbf{C}}_{i,j} = \frac{1}{T} \sum_{t=1}^{T-k} \mathbf{M}_{i,t} \mathbf{M}_{j,t+k} \tag{8}$$

where $\mathbf{A}_{h,k}$ stands for the element in the h^{th} row and k^{th} column of the $\mathbf{A}$ matrix. The matrix $\hat{\mathbf{C}}$ thus constructed is asymmetric, but the associated quadratic minimization problem $\mathbf{e}^{\mathsf{T}} \hat{\mathbf{C}} \mathbf{e}$ is equivalent to that in the symmetrized case with corresponding matrix

$$\mathbf{C}_{i,j} = \mathbf{C}_{j,i} = \frac{\hat{\mathbf{C}}_{i,j} + \hat{\mathbf{C}}_{j,i}}{2}\,, \tag{9}$$

a classical result trivially justified by observing that $(\mathbf{e}^{\mathsf{T}} \hat{\mathbf{C}} \mathbf{e})^{\mathsf{T}} = \mathbf{e}^{\mathsf{T}} \hat{\mathbf{C}}^{\mathsf{T}} \mathbf{e}$. The minimum eigenvalue of this matrix represents the mode for which the autocorrelation at time lag k is minimized. In the

study we perform, we fix $k = 1$, build the matrix $\mathbf{M}$ with the increments of the analyzed stocks during the D minimization days (increments normalized by the sample standard deviation over the same period) and find the eigenspace corresponding to the minimum eigenvalue. Once this procedure is complete, we have found the basket with the minimal possible lag-1 autocorrelation, among all baskets with unit euclidean norm. w_i is then equated to the i^{th} component of the eigenvector. So, calling these components $e_1, \ldots, e_N$ we obtain

$$B_{Dt} = \sum_{i=1}^{N} e_i x_{it} \tag{10}$$

where x_{it} is the element of the i^{th} discretized time series at discrete time t. Please note that x_{it} are the *raw* increments and they are not normalized by the standard deviation as in the matrix $\mathbf{M}$.

The minimum eigenvalue by itself does not give us the value of the autocorrelation; we must compute it directly from B_D using the estimator in Eq. (3) and we will repeat this operation in both the minimization period and the subsequent test period.

So, schematically, we proceed as follow:

1. We consider D consecutive days and the immediately following S days.
2. The time series are discretized (Sec. 2), normalized by the standard deviation during the D days, and used to build the matrix $\mathbf{C}$ with $k = 1$.
3. From the components of the eigenvector corresponding to the smallest eigenvalue, we build the basket B_D.
4. We evaluate the autocorrelation of B_D during the whole minimization period and, independently, during the test period.
5. Leaving the number of days D fixed, we shift the minimization period ahead by one day, and we go back to point 2.

The considered time scales (the discretization time step) are $5, 10, \ldots, 55, 60$ seconds. The number of test days S is arbitrary. Since we are fixing B_D outside these data we must only take care of having enough statistics, but we cannot increase much the minimization period, for fear of structural changes in the dependencies between the stocks that could cancel the desired effect. Therefore, we will test B_D for S respectively equal to 1 and 5 days.

The choice of D is more complicated. By using a small number of days, we decrease the size of the data set and are more likely to catch some transient dependencies. On the other hand, considering a large number of days D, we take the risk of averaging out the time changes, finding in the end a non satisfactory value of the autocorrelation. As trade off, we fix $D = 10$ days.

Table 1 contains the length of the time series when these numbers are applied.

In addition, we compute the Hurst exponent of B_D using the periodogram method. As stated above, the Hurst exponent alone does not give any information regarding the persistence of the autocorrelation and, moreover, the estimators are well known to be inaccurate (biased) in most of the situations. For those reason we do not use more complex estimators, nor do we want to speculate on the values we find, but rather present them only as indication of "possible" persistence effects. The estimation of the Hurst exponent is carried out each time considering a time scale

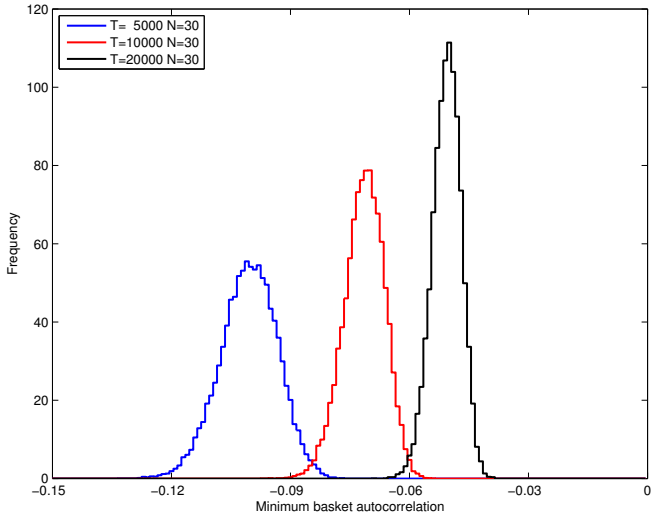

Figure 2: Empirical distribution of the optimal anti-autocorrelation for a basket built from 30 synthetic stock timeseries with i.i.d. normally distributed increments.

of 5s. This is because the Hurst exponent measures a global property of the process. Hence, stretching the time scale would only decrease the statistics and dampen the effects at the short time scales.

The autocorrelations values for the out-of-sample periods can be tested using the confidence interval discussed in Section 3. On the other hand, for the values obtained during the minimization periods, we cannot follow this simple rule. They are the smallest values possibly obtainable from the given data, so the null hypothesis cannot be the absence of autocorrelation. It must rather be the likelihood to obtain such values as minimization of independent time series. A heuristic approach is to run the minimization procedure described in this section using a set of $N = 30$ synthetic stocks with i.i.d., normally distributed increments and zero correlations. Doing so, we are able to plot the distribution of the minimal values obtainable by simulation. In Figure 2 this distribution is reported for the values of $T = 5000, 10000, 20000$ and for a statistic made from a population of 20000 synthetic baskets. Even for the shorter sample (small T), where spurious effects are more likely to occur, the absolute value of the optimal autocorrelation is hardly larger than 0.12, and the typical values drop sensitively for the larger samples.

5 Results

The main results are reported in Figures 3 and 4. Each figure corresponds to a specific time scale. The obtained autocorrelation values are reported as histograms. The red stands for the values during the minimization periods, blue and black for, respectively, the test with $S = 1$ and $S = 5$. The vertical lines indicate the corresponding null hypothesis. Figure 5 plots the mean values of these distributions as functions of the time scale.

The minimal values are far below those obtained from the synthetic independent time series

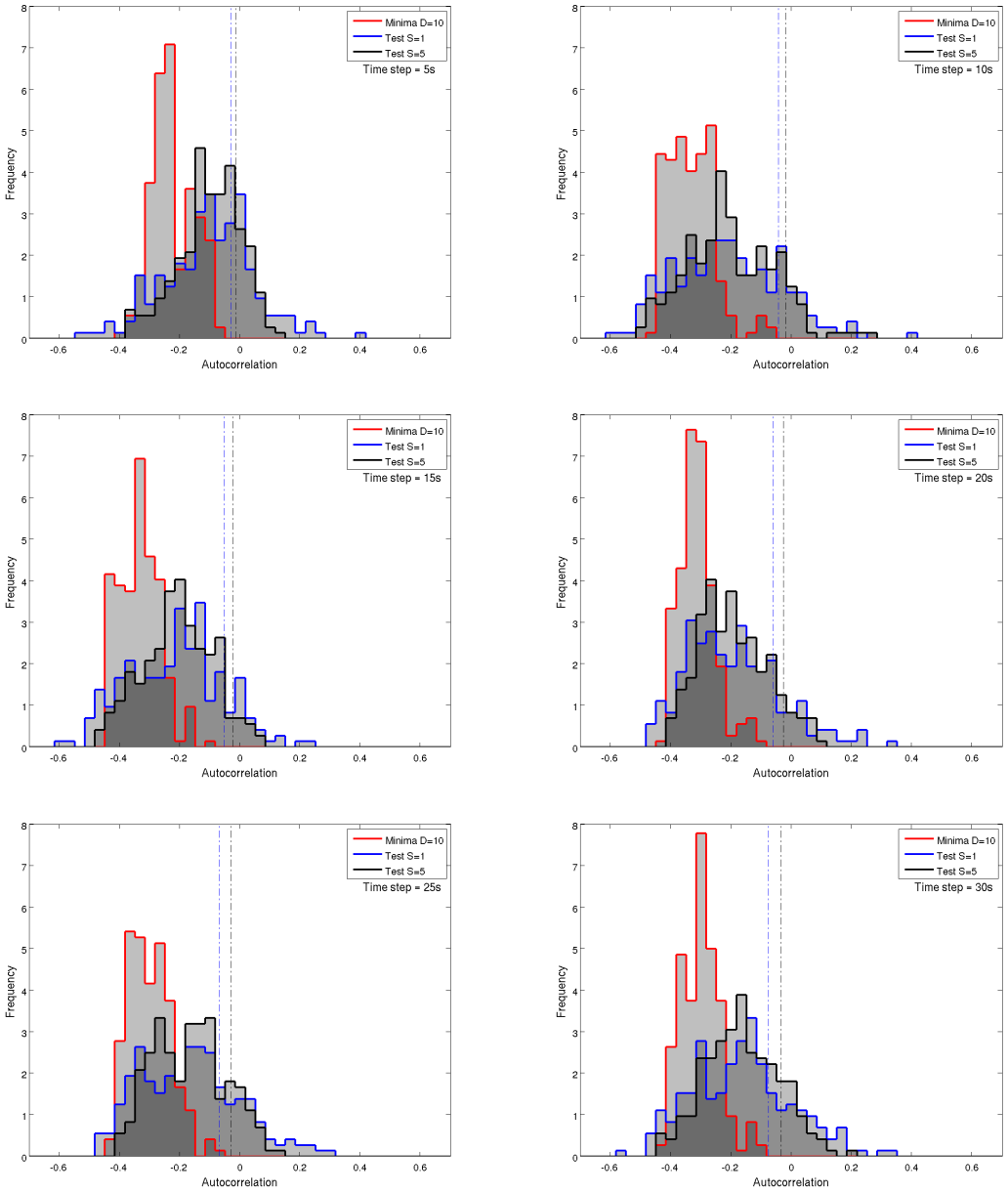

Figure 3: Histograms of the resulting autocorrelation values. In red we have the distributions for the minimal values obtained during $D = 10$ consecutive days, in blue and black we have the distributions for the test performed considering respectively $S = 1$ and $S = 5$ subsequent days. The dashed lines are the critical values for the significance test using the 5% of confidence.

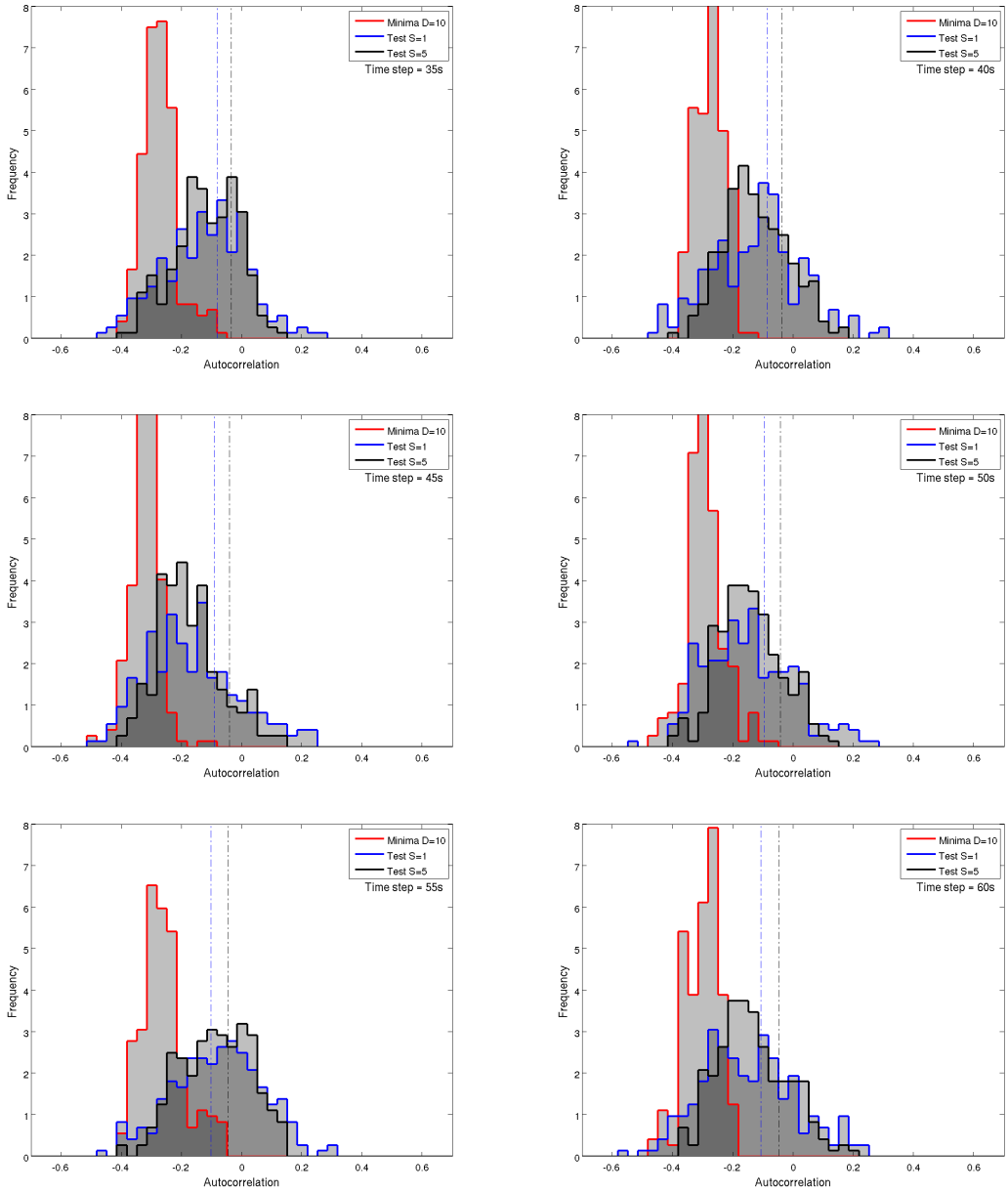

Figure 4: Histograms of the resulting autocorrelation values. In red we have the distributions for the minimal values obtained during $D = 10$ consecutive days, in blue and black we have respectively the distributions for the test performed considering $S = 1$ or $S = 5$ subsequent days. The dashed lines are the critical values for the significance test using the 5% of confidence.

	Minimization $D = 10$	Test $S = 1$	Test $S = 5$
5 s	41400	4140	20700
10 s	20700	2070	10450
15 s	13800	1380	6900
20 s	10350	1035	5175
25 s	8280	828	4140
30 s	6900	690	3450
35 s	5910*	591	2455
40 s	5170*	517	2065
45 s	4600	460	2300
50 s	4140	414	2070
55 s	3760*	376	1880
60 s	3450	345	1525

Table 1: Length of the time series as function of the considered time window size. For the starred values the considered time could have not been extended precisely until 3.45 pm; in these cases the considered period cannot be perfectly divided into the desired time steps, so some seconds preceding 3.45 pm have been discarded.

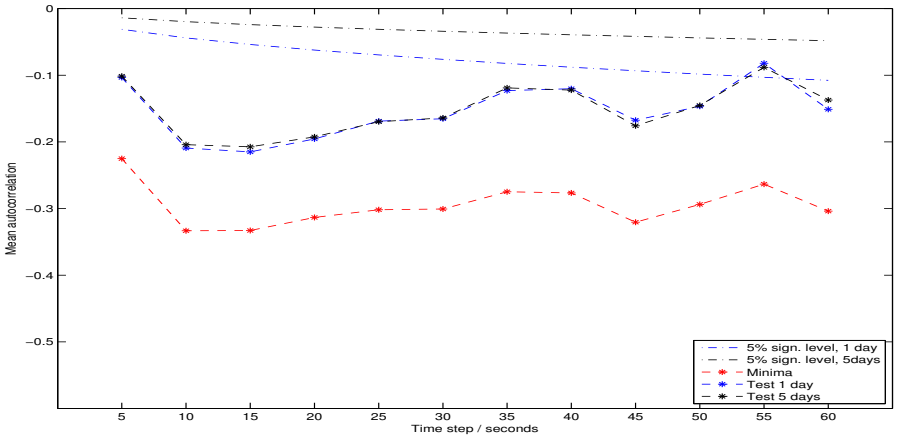

Figure 5: Mean values of the samples as function of the considered time step. The red line stands for the minimized value of the initial sample, the blue line stands for the out-of-sample test using one day of data and, finally, the black line stands for the out-of-sample test using 5 days of data. At the top, the thin dashed lines (blue and black) reports the values of 5% confidence level null hypothesis of non significant anticorrelation.

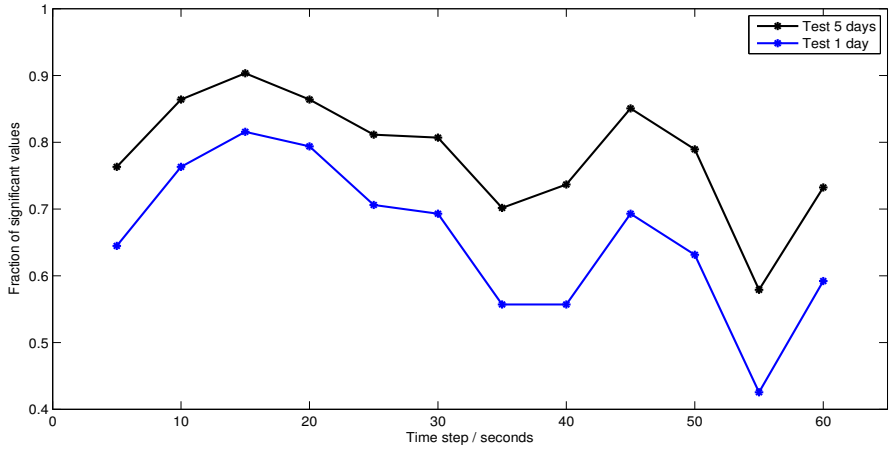

Figure 6: Fraction of the out-of-sample test values that are rejected by the significance test, using the 5% confidence level. Blue stands for the one-day test and black for the five-day test.

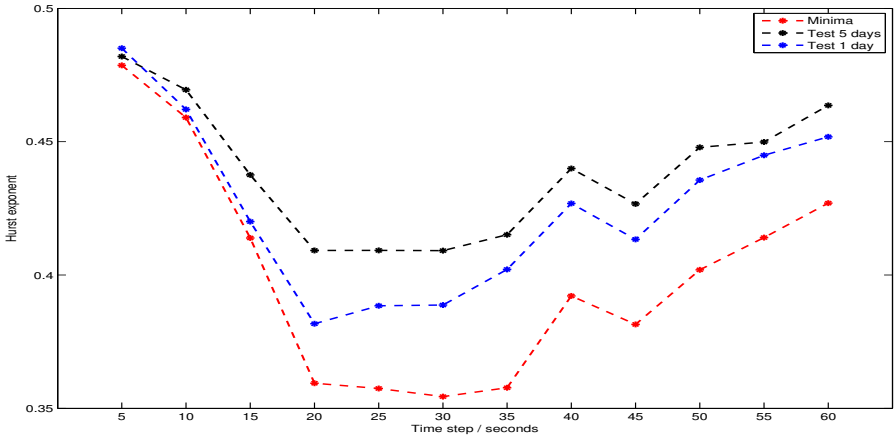

Figure 7: Mean value of the Hurst exponent on the three samples. The minimized portfolio (red) presents a strongest mean Hurst exponent in the range 20 35s. The out-of-sample test using five days (black) still retains a significant value and presents the stronger effect around the scale 20 30s.

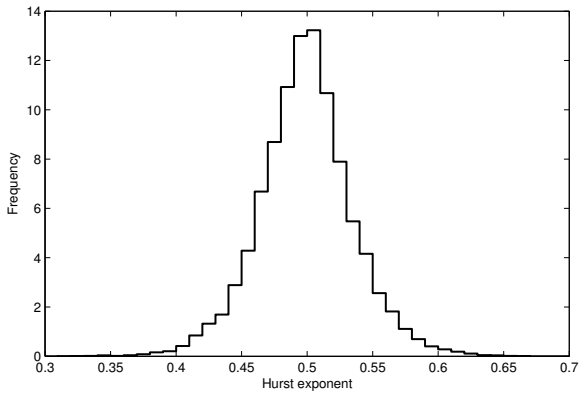

Figure 8: Distribution of Hurst exponent estimated using the periodogram method by randomly extracting ten consecutive day of the DJIA data and building a basket. The population is made by 20000 extractions and the mean value is 0.494.

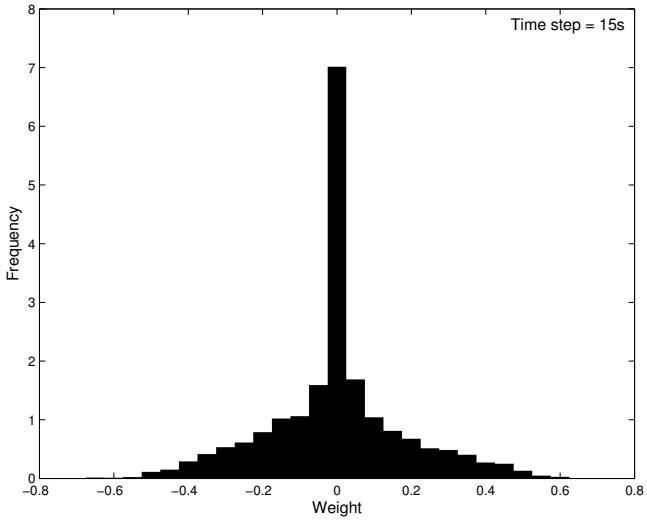

Figure 9: Distribution of weights. The mass around zero is important, but the distribution indeed suggests that a large portion of the considered stocks play an active role in the optimized basket. The depicted case corresponds to a time scale of $15s$; the remaining ones exhibit the same qualitative behaviour.

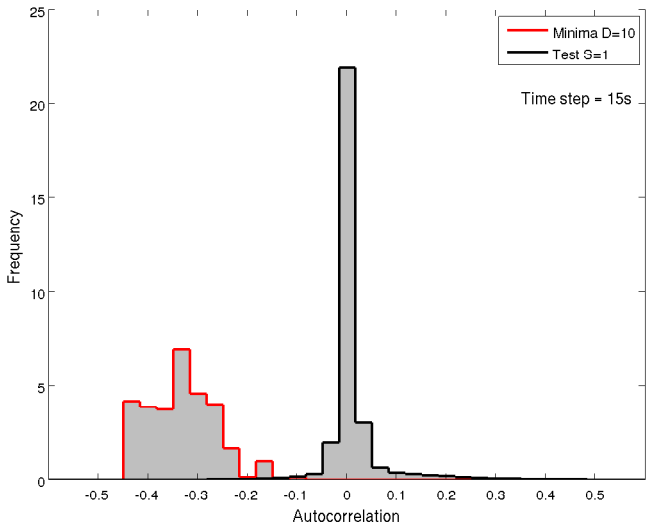

Figure 10: Empirical distribution of minimal values obtained during $D = 10$ consecutive days (red) and distribution of single entries the corresponding matrix $\hat{\mathbf{C}}$ (black). The depicted case corresponds to a time scale of $15s$; the remaining ones exhibit the same qualitative behaviour.

(see Figure 2). There is not much deviation between the various time scales, if only in the variance of the distributions. The mean value does not experience large deviations. The only time scale where we observe a smaller effect is $5s$. The desired property appears very strikingly for time scales of 10 and 15s.

The autocorrelations computed during the test periods deviate from the minimized values, with differences that become stronger for higher time scales. In this case too, the time scales of 10 and 15s display the most significant signal. Figure 6 shows the fraction of values that violate the null hypothesis, i.e., those values not compatible with independence as defined in Section 3. For the longer test and for the time scale equal to $15s$, this fraction is about 0.9.

In the same vein as in Figure 5, Figure 7 reports the mean values of the estimated Hurst exponent. In order to provide a comparison, Figure 8 presents the distribution of the Hurst exponents estimated by randomly extracting 10 consecutive days of DJIA data. The mean value of this distribution is slightly smaller than 0.5 and it is equal to 0.494. In general, the values of H we obtain in Figure 7 are significantly different from 0.5, with the stronger effects for time scales in the range of 20 to 35s.

A study of the weight dynamics is beyond the purpose of this paper. We content ourselves with the study of their empirical distribution, see Figure 9. Although quite peaked around 0, the distribution indeed suggests that a large portion of the stocks play a significant role in the optimized basket, so that we are effectively observing a collective dynamics rather than handling only a small subset of the assets.

Finally, Figure 10 reports the distribution of the entries of $\hat{\mathbf{C}}$, i.e. the autocorrelations and the lead-lag cross-correlations, together with the found minima, showing that the "distance" between

those two samples is strong, and thereby examplifying one more time the collective nature of the measured effect. In both Figures 9 and 10 the time scale is 15s, and the qualitative result does not change in the other cases.

6 Conclusion

A clear answer to the main research question of this work has been given. At short time scales, it is possible to build a basket that does not abide with the efficient market paradigm. Since we have analyzed mid-price time series, it is not completely clear, at this stage, to ascertain whether this effect is strong enough to be exploited for trading purposes. Preliminary back-testing of simple trading strategies using the "signal" we exhibited suggest that this practice, although not easy, is indeed possible. Other approaches using stochastic optimization algorithm are likely to provide stronger results. Moreover, in this study, we have focused on the raw autocorrelation values, but again, it is likely that suitably penalizing the criterion could lead to more favorable trading situations - for example, one can try to reproduce a similar study, considering baskets which minimize the autocorrelation only among those with a minimal given value of volatility. Moreover, the value of D for which we have shown the results can be fine-tuned, and so can the time scales.

Finally, we must also point out the fact that we have considered all the 30 DJIA stocks, without using any smart way of picking them. Since we already have clear results using the full index, we are keen to think that a smarter choice of the companies would lead to stronger effects.

Acknowledgement

F.A. would like to acknowledge that the idea of optimizing a basket of stocks with respect to statistical properties arose from a collaboration with Laurent Jaillet while they were colleagues at CAI Cheuvreux.