New Economic Windows

Frédéric Abergel
Hideaki Aoyama
Bikas K. Chakrabarti
Anirban Chakraborti
Asim Ghosh *Editors*

Econophysics and Data Driven Modelling of Market Dynamics

Springer

Econophysics and Data Driven Modelling of Market Dynamics

New Economic Windows

More information about this series at http://www.springer.com/series/6901

Frédéric Abergel · Hideaki Aoyama
Bikas K. Chakrabarti · Anirban Chakraborti
Asim Ghosh
Editors

Econophysics and Data Driven Modelling of Market Dynamics

Springer

Editors

Frédéric Abergel
Laboratory of Mathematics Applied to System
CentraleSupélec
Châtenay-Malabry
France

Hideaki Aoyama
Department of Physics
Kyoto University
Kyoto
Japan

Bikas K. Chakrabarti
Saha Institute of Nuclear Physics
Kolkata
India

Anirban Chakraborti
School of Computational and Integrative Sciences
Jawaharlal Nehru University
New Delhi
India

Asim Ghosh
Saha Institute of Nuclear Physics
Kolkata
India

ISSN 2039-411X ISSN 2039-4128 (electronic)
New Economic Windows
ISBN 978-3-319-08472-5 ISBN 978-3-319-08473-2 (eBook)
DOI 10.1007/978-3-319-08473-2

Library of Congress Control Number: 2014957704

Springer Cham Heidelberg New York Dordrecht London

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

This proceedings volume is based on the conference entitled ‘Econophysics and Data Driven Modelling of Market Dynamics’ that was held at Saha Institute of Nuclear Physics, Kolkata during 14–17 March 2014. This was the eighth event of the ‘Econophys-Kolkata’ series of conferences, and was organized jointly by Saha Institute of Nuclear Physics, École Centrale Paris, Jawaharlal Nehru University and Kyoto University.

During the past decades, the financial market landscape has been dramatically changing: deregulation of markets, growing complexity of products, etc. The ever-rising speed and decreasing costs of computational power and networks have led to the emergence of huge databases. We chose this particular theme for the conference, as we thought that it would be most appropriate with the availability of these data. Econophysicists, along with many others, have been relying primarily on empirical observations in order to construct models and validate them, or study models that are better empirically founded. Thus, a major part of the efforts of econophysicists have been the study of empirical data and financial time series analyses. Often, the empirics have guided researchers to design more realistic and practical models. The recent turmoil on financial markets and the 2008 crash seem to plead for new models or approaches, and the econophysics community indeed has an important role to play in market modelling in the future years to come.

This proceedings volume contains papers by distinguished experts from all over the world, mostly based on the talks and seminars delivered at the meeting and accepted after refereeing. For completeness, a few articles by experts who could not participate in the meeting due to unavoidable reasons were also invited and these too have been incorporated in this volume. This volume is organized as follows: A first part dedicated to ‘Market Analysis and Modelling’. A second part entitled ‘Miscellaneous’ presents other ongoing studies in related areas on econophysics and sociophysics. We have included in the third part, ‘Reviews’, two reviews which address recent developments in econophysics and sociophysics. We have included in the fourth part, ‘Discussions and Commentary’, an extensive note on the impact of econophysics researches (obtained from responses of leading researchers to questionnaire). Another write-up in this part discusses the influence of econophysics research on contemporary researches in social sciences.

We are grateful to all the participants at the meeting and for their contributions. We are also grateful to Mauro Gallegati and the Editorial Board of the 'New Economic Windows' series of Springer-Verlag (Italy) for their continuing support in getting this proceedings volume published in their esteemed series.

The conveners (editors) also express their thanks to Saha Institute of Nuclear Physics, École Centrale Paris, Jawaharlal Nehru University and Kyoto University for their support in organizing this conference. The support from J.C. Bose project fund (DST, India) of Bikas K. Chakrabarti is gratefully acknowledged.

Châtenay-Malabry, France	Frédéric Abergel
Kyoto, Japan	Hideaki Aoyama
Kolkata, India	Bikas K. Chakrabarti
New Delhi, India	Anirban Chakraborti
Kolkata, India	Asim Ghosh
October 2014	

Contents

Part II Miscellaneous

Part III Reviews

Part IV Discussions and Commentary

Part I
Market Analysis and Modelling

Empirical Evidence of Market Inefficiency: Predicting Single-Stock Returns

Marouane Anane and Frédéric Abergel

Abstract Although it is widely assumed that the stock market is efficient, some empirical studies have already tried to address the issue of forecasting stock returns. As far as is known, it is hard to find a paper involving not only the forecasting statistics but also the forecasting profitability. This paper aims to provide an empirical evidence of the market inefficiency and to present some simple realistic strategies based on forecasting stocks returns. In order to achieve this study, some linear and non linear algorithms are used to prove the predictability of returns. Many regularization methods are introduced to enhance the linear regression model. In particular, the RIDGE method is used to address the colinearity problem and the LASSO method is used to perform variable selection. The different obtained results show that the stock market is inefficient and that profitable strategies can be computed based on forecasting returns. Empirical tests also show that simple forecasting methods perform almost as well as more complicated methods.

1 Introduction

Forecasting the market has been one of the most exciting financial subjects for over a century. In 1900, Bachelier [1] admitted, “Undoubtedly, the Theory of Probability will never be applicable to the movements of quoted prices and the dynamics of the Stock Exchange will never be an exact science. However, it is possible to study mathematically the static state of the market at a given instant to establish the probability law for the price fluctuations that the market admits at this instant”. Seventy years later, Fama [2] proposed some formal definitions of the market efficiency; “A market in which prices always fully reflect available information is called efficient”. Opinions

M. Anane
Chair of Quantitative Finance, MAS Laboratory, Ecole Centrale Paris,
BNP Paribas 20 Boulevard des Italiens, 75009 Paris, France
e-mail: marouane.anane@gmail.com

F. Abergel (✉)
Chair of Quantitative Finance, MAS Laboratory, Ecole Centrale Paris,
Grande voie des vignes, 92290 Chatenay Malabry, France
e-mail: frederic.abergel@ecp.fr

F. Abergel et al. (eds.), *Econophysics and Data Driven Modelling of Market Dynamics*,
New Economic Windows, DOI 10.1007/978-3-319-08473-2_1

have been always divergent about the market efficiency. Malkiel [3] concluded that most investors trying to predict stocks' returns always ended up with profits inferior to passive strategies. In his famous book, *Fooled by Randomness*, Taleb [4] argued that even the best performances can be explained by luck and randomness. On the other hand, finance professionals demonstrated, in real life, that they can always make money beating the market; see Warren Buffett's response to efficient market claims [5].

The recent rise in electronic markets lead to big available financial data. The attempt to discover some predictable, and hopefully profitable, signal in the middle of those millions of numbers has never been as high as today.

In the academic world, Chakraborti et al. [6] studied in detail the statistical properties of the intraday returns, and came to the conclusion that there is no evidence of correlation between successive returns. Similarly, Lillo and Farmer [7] concluded that stock returns contain negligible temporal autocorrelation. Fortunately, Zheng and Abergel [8] found some promising results, in particular the liquidity imbalance on the best bid/ask seems to be informative to predict the next trade sign.

In the professional world, many books present hundred of strategies predicting the market and always earning money; see [9, 10] for example. When testing those strategies in other samples, results are so different and the strategies are no longer profitable. It is possible that the overfit of such methods played a key role in the good performances published in those books.

This study was performed from both an academic and a professional perspective. For each prediction method, not only are statistical results presented, but also presented are the performances of the correspondent strategies. The aim is to give another point of view of a good prediction and of an efficient market.

This work is organized as follows: In the first section, the data and the test methodology are presented. In the second section a non linear method, based on conditional probability matrices, is used to test the predictive power of each indicator. In the last section, the linear regression is introduced to combine the different indicators and many regularization ideas are tested in order to enhance the performances of the strategies.

2 Data, Methodology and Performance Measures

2.1 Data

This paper focuses on the EURO STOXX 50 European liquid stocks. One year (2013) of full daily order book data provided by BNP Paribas are used to achieve the study. For a stock with a mid price S_t at time t, the return to be predicted over a period dt is $Ln(\frac{S_{t+dt}}{S_t})$. At the time t, one can use all the available data for any time $s \leq t$ to perform the prediction.

In Sects. 2 and 3, the focus is on predicting the stocks' returns over a fixed period dt using some order book indicators. Once the returns and the indicators are computed, the data are sampled on a fixed time grid from 10 to 17 h with a resolution dt. Three different resolutions are tested; 1, 5 and 30 min.

Below are the definitions of the studied indicators and the rationale behind using them to predict the returns:

Past return: The past return is defined as $Ln(\frac{S_t}{S_{t-dt}})$. Two effects justify the use of the past return indicator to predict the next return; the mean-reversion effect and the momentum effect. If a stock, suddenly, shows an abnormal return that, significantly, deviates the stock's price from its historical mean value, the mean reversion effect is observed when an opposite return occurs rapidly to put the stock back in its usual average price range. On the other hand, if the stock shows, progressively, an important and continuous deviation; the momentum effect occurs when more market participants are convinced of the move and trade in the same sense increasing the deviation even more.

Order book imbalance: The liquidity on the bid (respectively ask) side is defined as $Liq_{bid} = \sum_{i=1}^{5} w_i b_i bq_i$ (respectively $Liq_{ask} = \sum_{i=1}^{5} w_i a_i aq_i$), where b_i (respectively a_i) is the price at the limit i on the bid (respectively ask) side, bq_i (respectively aq_i) is the corresponding available quantity, and w_i is a decreasing function on i used to give more importance to the best limits. Those indicators give an idea about the instantaneous money available for trading on each side of the order book. Finally, the order book imbalance is defined as $Ln(\frac{Liq_{bid}}{Liq_{ask}})$. This indicator summarizes the order book static state and gives an idea about the buy-sell instantaneous equilibrium. When this indicator is significantly higher (respectively lower) than 0, the available quantity at the bid side is significantly higher (respectively lower) than the one at the ask side; only few participants are willing to sell (respectively buy) the stock, which might reflect a market consensus that the stock will move up (respectively down).

Flow quantity: This indicator summarizes the order book dynamic over the last period dt. Q_b (respectively Q_s) is denoted as the sum of the bought (respectively sold) quantities, over the last period dt and the flow quantity is defined as $Ln(\frac{Q_b}{Q_s})$. This indicator is close to the order flow and shows a high positive autocorrelation. The rationale behind using the flow quantity is to verify if the persistence of the flow is informative about the next return.

EMA: For a process $(X)_{t_i}$ observed on discrete times (t_i), the Exponential Moving Average $EMA(d, X)$ of delay d is defined as $EMA(d, X)_{t_0} = X_{t_0}$ and for $t_{1 \leq i}$, $EMA(d, X)_{t_i} = \omega X_{t_i} + (1 - \omega) EMA(d, X)_{t_{i-1}}$, where $\omega = min(1, \frac{t_i - t_{i-1}}{d})$. The EMA is a weighted average of the process with an exponential decay. The smaller d is, the shorter the EMA memory is.

2.2 Methodology

The aim of this study is to prove, empirically, the market inefficiency by predicting the stocks' returns for three different periods: 1, 5 and 30 min. In Sect. 2, the used indicators are the past returns, the order book imbalance and the flow quantity. A simple method based on historical conditional probabilities is used to prove, separately, the informative effect of each indicator. In Sect. 3, the three indicators and their $EMA(X, d)$ for $d \in (1, 2, 4, 8, 16, 32, 64, 128, 256)$ are combined in order to perform a better prediction than the mono indicator case. Different methods, based on the linear regression, are tested. In particular, the statistical and the numerical stability problems of the linear regression are addressed. In the different sections, the predictions are tested statistically, then used to design a simple trading strategy. The goal is to verify, whether or not one can find a profitable strategy covering 0.5 basis point trading costs. This trading cost is realistic and corresponds to many funds, brokers, and banks trading costs. The possibility of computing, if it exists, a strategy, profitable, after paying the costs, would be an empirical argument of the market inefficiency. Notice that, in all the sections, the learning samples are sliding windows containing sufficient number of days, and the testing samples are the next days. The models parameters are fitted on the learning sample (called in-sample) and the strategies are tested on the testing sample (called out of sample). The sliding training avoids any overfit problem since performances are only computed out of sample.

2.3 Performance Measures

In the most of the studies addressing the market efficiency, the results are summarized in the linear correlation. However, this measure is not enough to conclude about the returns predictability or the market efficiency. Results interpretation should depend on the predicted signal and the trading strategy. A 1 % correlation is high if the signal is supposed to be totally random, and 99 % correlation is insufficient if the signal is supposed to be perfectly predictable. Moreover, a trader making 1 euro each time trading a stock with 50.01 % probability and losing 1 euro with 49.99 % probability, might be considered as a noise trader. However, if this strategy can be run, over 500 stocks, one time a second, for 8 hours a day, at the end of the day the gain will be the sum S_n of $n = 14.4$ million realisations. Using the central limit theorem, $\frac{S_n}{n}$ has a normal law $N(E, \frac{\sigma}{\sqrt{n}})$. Thus the probability of having a negative trading day is $\Phi(\frac{-E\sqrt{n}}{\sigma}) = \Phi(-0.62) = 26.5\,\%$, so much lower than the one of a noise trader. In this paper, returns are considered predictable and thus the market is considered inefficient, if one can run a profitable strategy covering the trading costs.

3 Conditional Probability Matrices

The conditional probability matrices method uses observed frequencies as an estimation of the conditional probability law. To apply this method, data need to be descritized in a small number of classes. Denote the explanatory variable as X, the return as Y and the frequencies matrix as M. Denote the classes of X (respectively Y) as $C^X = \{C_i^X : i \in \mathbb{N}_+ \cap \{i \leq S_X\}\}$ (respectively $C^Y = \{C_i^Y : i \in \mathbb{N}_+ \cap \{i \leq S_Y\}\}$).

Table 1 Historical frequencies matrix for Deutsh Telecom over 2013

	$A =$ "$Y < 0$"	$B =$ "$Y > 0$"
$A =$ "$X < 0$"	19,950	21,597
$B =$ "$X > 0$"	21,597	20,448

S_X (respectively S_Y) denotes the total number of classes for X (respectively Y). For a given learning period $[0, T]$ containing N observations, the frequencies matrix at the time T is constructed as:

$$M_T^{i,j} = card(\{(X_{t_n} \in C_i^X, Y_{t_n} \in C_j^Y)\})$$

where $n \in \mathbb{N}_+ \cap \{n \leq N\}$, and X_{t_n} (respectively Y_{t_n}) is the n^{th} observed value of X (respectively Y), observed at the time t_n. Note that the return Y_{t_n} is backshifted for one instant (namely $Y_{t_n} = Ln(\frac{S_{t_{n+1}}}{S_{t_n}})$). Finally, the prediction of the next Y conditional to the last observed X_T can be computed using the matrix M_T.

The idea of this method is a simple application of the statistical independence test. If some events $A =$ "$X_{t_n} \in C_i^X$" and $B =$ "$Y_{t_n} \in C_j^Y$" are statistically independent then $P(A|B) = P(A)$. For example, to check if the past returns (denoted X in this example) can help predicting the future returns (denoted Y in this example), the returns are classified into two classes, then the empirical historical frequencies matrix is computed. Table 1 shows the results for the 1-min returns of Deutsh Telecom over the year 2013.

In probabilistic terms, the historical probability to observe a negative return is $P(A) = 49.70\,\%$ and to observe a positive return is $P(B) = 50.30\,\%$. Thus a trader always buying the stock would have a success rate of 50.30 %. Notice that: $P(A/A) = 48.02\,\%, P(B/A) = 51.98\,\%, P(A/B) = 51.37\,\%, P(B/B) = 48.63\,\%$. Thus, a trader playing the mean-reversion (buy when the past return is negative and sell when the past return is positive), would have a success rate of 51.67 %. Notice that the same approach as **1.3** gives a success rate, when trading the strategy over 500 stocks, of 54.38 % for the buy strategy and of 72.91 % for the mean reversion strategy.

This simple test shows that the smallest statistical bias can be profitable and useful for designing a trading strategy. However the previous strategy is not realistic; the conditional probabilities are computed in sample and the full sample data of Deutsh Telecom was used for the computation. In reality, predictions have to be computed using only the past data. It is, thus, important to have stationary probabilities. Table 2 shows that the monthly observed frequencies are quite stable, and thus can be used

Table 2 Monthly historical conditional probabilities: in the most cases, $P(A/A)$ and $P(B/B)$ are lower than 50 % where $P(B/A)$ and $P(A/B)$ are higher than 50 %

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
P(A/A)	0.49	0.48	0.47	0.48	0.47	0.50	0.48	0.51	0.51	0.47	0.46	0.44
P(B/A)	0.51	0.52	0.53	0.52	0.53	0.50	0.52	0.49	0.49	0.53	0.54	0.56
P(A/B)	0.50	0.52	0.51	0.53	0.52	0.49	0.51	0.50	0.51	0.50	0.51	0.55
P(B/B)	0.50	0.48	0.49	0.47	0.48	0.51	0.49	0.50	0.49	0.50	0.49	0.45

to estimate out of sample probabilities. Each month, one can use the observed frequencies of the previous month as an estimator of current month probabilities.

In the following paragraphs, frequencies matrices are computed on sliding windows for the different indicators. Several classification and prediction methods are presented.

3.1 Binary Method

In the binary case, returns are classified into positive and negative as the previous example and explanatory variables are classified relatively to their historical mean. A typical constructed matrix is shown in Table 1. Denote, in the Table 1 example, $C_1^X = \{X < \overline{X} = 0\}$, $C_2^X = \{X > \overline{X} = 0\}$, $C_1^Y = \{Y < 0\}$, $C_2^Y = \{Y > 0\}$. Y can be predicted using different formula based on the frequency matrix. Below some estimators examples:

$\widehat{Y_1}$: The sign of the most likely next return conditionally to the current state.
$\widehat{Y_2}$: The expectation of the most likely next return conditionally to the current state.
$\widehat{Y_3}$: The expectation of next return conditionally to the current state.

$$\widehat{Y_1} = \begin{cases} +1 & \text{if } X_T \in C_1^X \\ -1 & \text{if } X_T \in C_2^X \end{cases}$$

$$\widehat{Y_2} = \begin{cases} E(Y|Y \in C_2^Y \cap X \in C_1^X) & \text{if } X_T \in C_1^X \\ E(Y|Y \in C_1^Y \cap X \in C_2^X) & \text{if } X_T \in C_2^X \end{cases}$$

$$\widehat{Y_3} = \begin{cases} E(Y|X \in C_1^X) & \text{if } X_T \in C_1^X \\ E(Y|X \in C_2^X) & \text{if } X_T \in C_2^X \end{cases}$$

In this study, only results based on the estimator $\widehat{Y_3}$ (denoted $\widehat{Y}$ in the rest of the paper) are presented. Results computed using different other estimators are equivalent and the differences do not impact the conclusions. To measure the quality of the prediction, four tests are applied:

AUC: (Area under the curve) combines the true positive rate and the false positive rate to give an idea about the classification quality.

Accuracy: defined as the ratio of the correct predictions (Y and $\widehat{Y}$ have the same sign).

Gain: computed on a simple strategy to measure the prediction performance. Predictions are used to run a strategy that buys when the predicted return is positive and sells when it is negative. At each time, for each stock the strategy's position is in $\{-100, 000, 0 , +100, 000\}$.

Profitability: defined as the gain divided by the traded notional of the strategy presented above. This measure is useful to estimate the gain with different transaction costs. Figure 1 summarizes the statistical results of predicting the 1-min returns

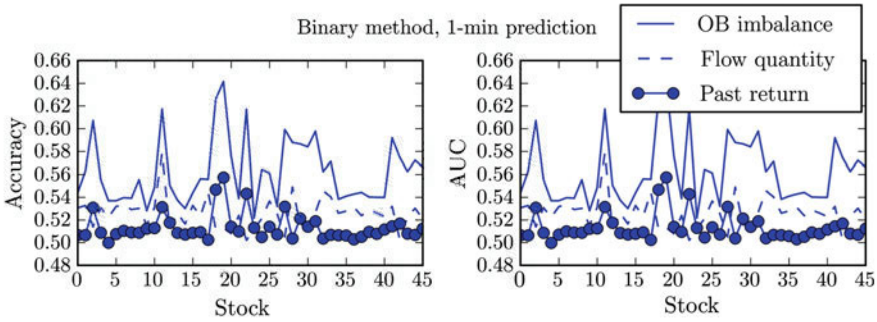

Fig. 1 The quality of the binary prediction: the AUC and the accuracy are higher than 50 %. The three predictors are better than random guessing and are significantly informative

using the three indicators. For each predictor, the AUC and the accuracy are computed over all the stocks. Notice that for each stock, results are computed over more than 100,000 observations and the amplitude of the 95 % confidence interval is around 0.6 %. For the three indicators, the accuracy and the AUC are significantly higher than the 50 % random guessing threshold. The graph shows also that the order book imbalance gives the best results and that the past resturn is the least successful predictor. Detailed results per stock are given in Table 3 of Appendix 1.

In Fig. 2, the performances of the trading strategies based on the prediction of the 1-min returns are presented. The strategies are profitable and the results confirm the predictability of the returns (see the details in Table 4 of Appendix 1).

In Fig. 3, the cumulative gains of the strategies based on the 3 indicators over 2013 are represented. When trading without costs, predicting the 1-min return using the past return and betting 100,000 euros at each time, would make a 5-million Euro profit. Even better, predicting using the order book imbalance would make more than 20 million Euros profit. The results confirm the predictability of the returns,

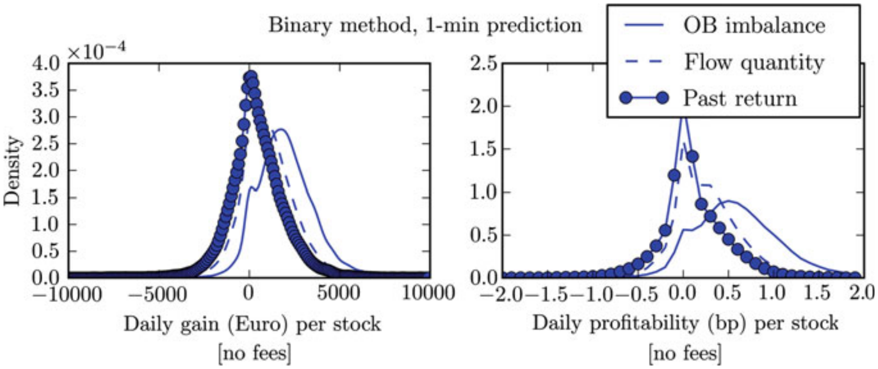

Fig. 2 The quality of the binary prediction: for the 3 predictors, the densities of the gain and the profitability are positively biased, confirming the predictability of the returns

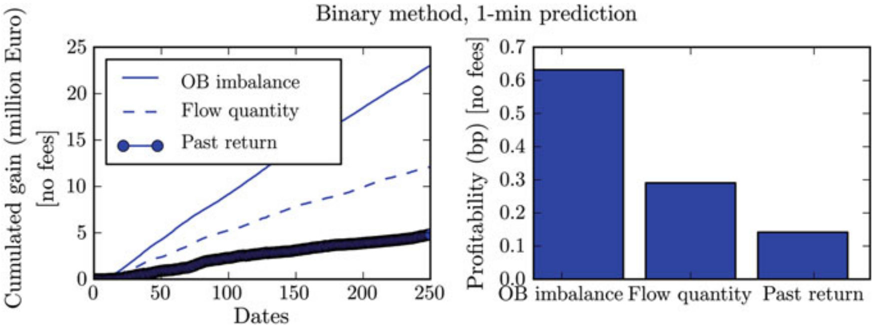

Fig. 3 The quality of the binary prediction: the graphs confirm that the 3 indicators are informative and that the order book imbalance indicator is the most profitable

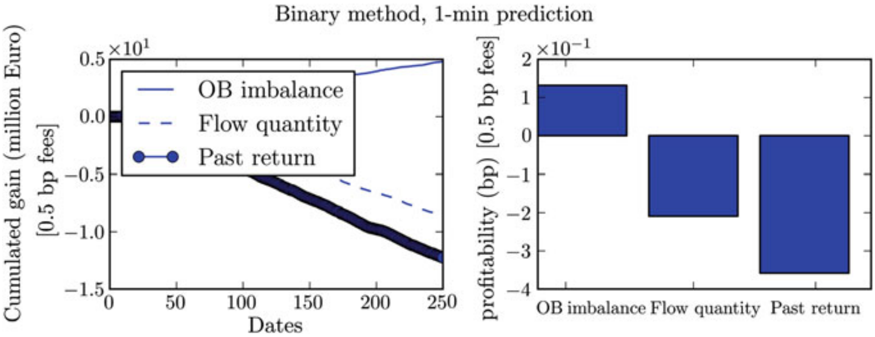

Fig. 4 The quality of the binary prediction: when adding the 0.5 bp trading costs, the strategies are no longer very profitable

but not the inefficiency of the market. In fact, Fig. 4 shows that, when adding the 0.5 bp trading costs, only the strategy based on the order book imbalance remains (marginally) positive. Thus, no conclusion, about the market efficiency, can be made (see more details in Table 5 of Appendix 1).

Figure 5 represents the cumulative gain and the profitability for the 5- and the 30-min strategies (with the trading costs). The strategies are not profitable. Moreover, the predictive power decreases with an increasing horizon.Similar as the 1-min prediction, the detailed results of the 5-min prediction can be found in Tables 6, 7 and 8 of Appendix 1. Those of the 30-min prediction can be found in Tables 9, 10 and 11 of the same Appendix.

The results of the binary method show that the returns are significantly predictable. Nevertheless, the strategies based on those predictions are not sufficiently profitable to cover the trading costs. In order to enhance the predictions, the same idea is applied to the four-class case. Moreover, a new strategy based on a minimum threshold of the expected return is tested.

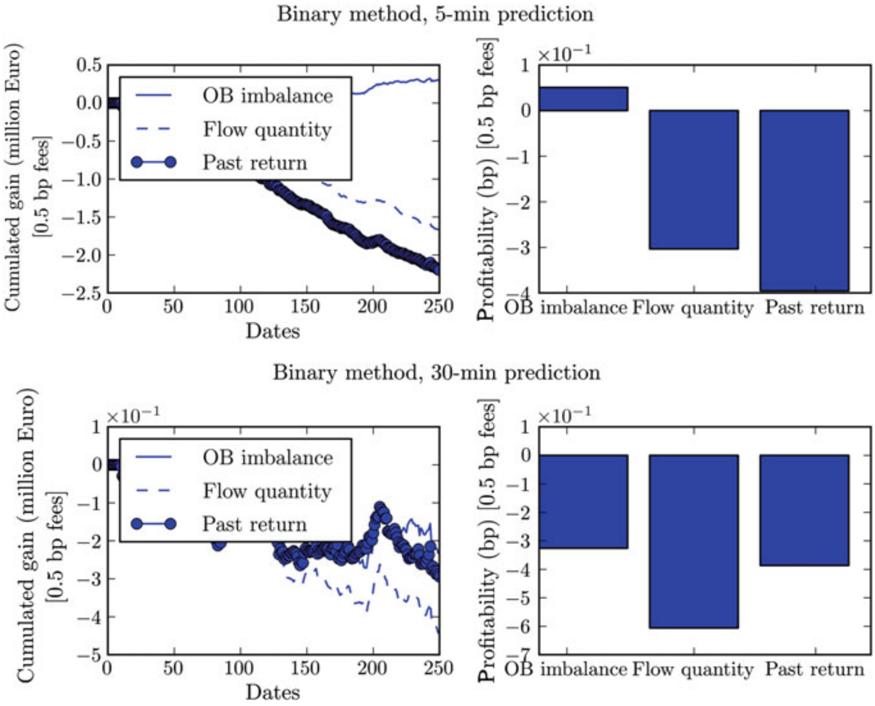

Fig. 5 The quality of the binary prediction: the strategies are not profitable. Moreover, the performances decreases significantly compared to the 1-min horizon

3.2 Four-Class Method

The indicator X is now classified into four classes; "very low values" C_1^X, "low values" C_2^X, "high values" C_3^X and "very high values" C_4^X. At each time t_n, Y is predicted as $\widehat{Y} = E(Y|X \in C_i^X)$, where C_i^X is the class of the current observation X_{t_n}. As the previous case, the expectation is estimated from the historical frequencies matrix.

Finally, a new trading strategy is tested. The strategy is to buy (respectively sell) 100,000 euros when $\widehat{Y}$ is positive (respectively negative) and $|\widehat{Y}| > \theta$, where θ is a minimum threshold (1 bp in this paper). Notice that the case $\theta = 0$ corresponds to the strategy tested in the binary case.

The idea of choosing $\theta > 0$ aims to avoid trading the stock when the signal is noisy. In particular, when analyzing the expectations of Y relative to the different classes of X, it is always observed that the absolute value of the expectation is high when X is in one of its extreme classes (C_1^X or C_4^X). On the other hand, when X is in one of the intermediary classes (C_2^X or C_3^X) the expectation of Y is close to 0 reflecting a noisy signal.

For each indicator X, the classes are defined as $C_1^X =]-\infty, X_a[$, $C_2^X =]X_a, X_b[$, $C_3^X =]X_b, X_c[$ and $C_4^X =]X_c, +\infty[$. To compute X_a, X_b and X_c, the three following classifications were tested:

Quartile classification: In the in-sample period, the quartile Q_1, Q_2 and Q_3 are computed for each day then averaged over the days. X_a, X_b and X_c corresponds, respectively, to $\overline{Q_1}$, $\overline{Q_2}$ and $\overline{Q_3}$.

K-means classification: The K-means algorithm [11], applied to the in-sample data with $k = 4$, gives the centers G_1, $G2$, G_3 and G_4 of the optimal (in the sense of the minimum within-cluster sum of squares) clusters. X_a, X_b and X_c are given respectively by $\frac{G_1+G_2}{2}$, $\frac{G_2+G_3}{2}$ and $\frac{G_3+G_4}{2}$.

Mean-variance classification: The average $\overline{X}$ and the standard deviation $\sigma(X)$ are computed in the learning period. Then, X_a, X_b and X_c correspond, respectively, to $\overline{X} - \sigma(X)$, $\overline{X}$ and $\overline{X} + \sigma(X)$.

In this paper, only the results based on the mean-variance classification are presented. The results computed using the two other classifications are equivalent and the differences do not impact the conclusions.

Figure 6 compares the profitabilities of the binary and the 4-class methods. For the 1-min prediction, the results of the 4-class method are significantly better. For the longer horizons, the results of the both methods are equivalent. Notice also that, using the best indicator, in the 4-class case, one could obtain a significantly positive performance after paying the trading costs. The detailed results per stock are given in Tables 12, 13, 14, 15, 16, 17, 18, 19 and 20, of Appendix 2.

The interesting result of this first section is that even when using the simplest statistical learning method, the used indicators are informative and provide a better prediction than random guessing. However, in most cases, the obtained performances are too low to conclude about the market inefficiency. In order to enhance the performances, the 3 indicators and their exponential moving average are combined using some classic linear methods in the next section.

4 Linear Regression

In this section, the matrix X denotes a 30-column matrix containing the 3 indicators and their $EMA(d)$ for $d \in (1, 2, 4, 8, 16, 32, 64, 128, 256)$. The vector Y denotes the target to be predicted. Results of the previous section proved that the used indicators are informative and thus can be used to predict the target. In general, one can calibrate, on the learning sample, a function f such that $f(X)$ is "the closet possible" to Y and hope that, for some period after the learning sample, the relation between X and Y is still close enough to the function f. Hence $f(X)$ would be a "good" estimator of Y. Due to the finite number of observations in the learning sample, one can always find $f(X)$ arbitrary close to Y by increasing the number of the freedom degree. However,

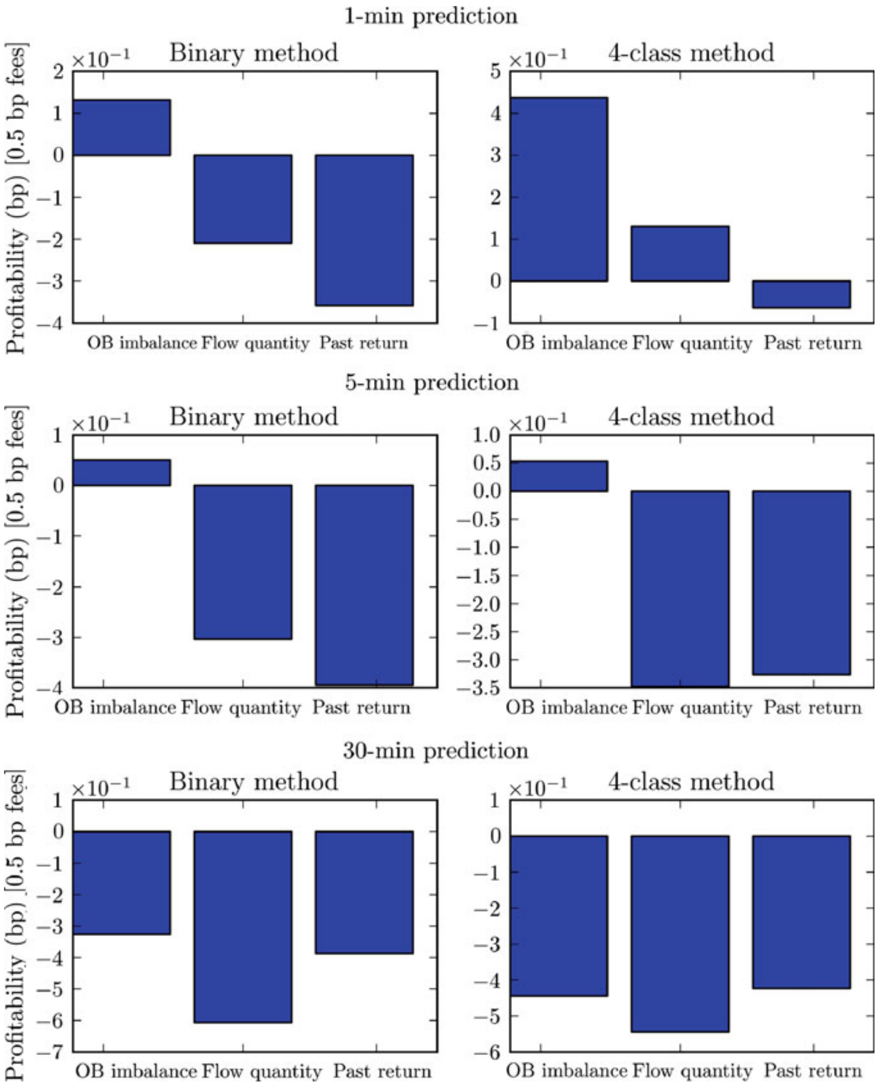

Fig. 6 The quality of the 4-class prediction: for the 1-min prediction, the results of the 4-class method are significantly better than the results of the binary one. For longer horizons, both strategies are not profitable when adding the trading costs

such perfect in-sample calibration overfits the data and the out of sample results are always irrelevant.

In the linear case, f is supposed to be linear and the model errors are supposed to be independent and identically distributed [12] (Gaussian in the standard textbook model). A more mathematical view of linear regression is that it is a probabilistic model of Y given X that assumes:

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

For technical reasons, the computations are done with z-scored data (use $\frac{X_i - \overline{X_i}}{\sigma(X_i)}$ in stead of X_i).

4.1 Ordinary Least Squares (OLS)

OLS method consists of estimating the unknown parameter β by minimizing the sum of squares of the residuals between the observed variable Y and the linear approximation $X\beta$. The estimator is denoted $\widehat{\beta}$ and is defined as

$$\widehat{\beta} = argmin_{\beta}(J_{\beta} = ||Y - X\beta||_2^2)$$

This criterion is reasonable if at each time i the row X_i of the matrix X and the observation Y_i of the vector Y represent independent random sample from their populations.

The cost function J_{β} is quadratic on β and differentiating with respect to β gives:

$$\frac{\delta J_{\beta}}{\delta\beta} = 2t_X X\beta - 2t_X Y$$
$$\frac{\delta^2 J_{\beta}}{\delta\beta\delta\beta} = 2t_X X$$

When $t_X X$ is invertible, setting the first derivative to 0, gives the unique solution $\widehat{\beta} = (t_X X)^{-1} t_X Y$. The statistical properties of this estimator can be calculated straightforward as follows:

$$\begin{aligned}
E(\widehat{\beta}|X) &= (t_X X)^{-1} t_X E(Y|X) = (t_X X)^{-1} t_X X\beta = \beta \\
Var(\widehat{\beta}|X) &= (t_X X)^{-1} t_X Var(Y|X) X (t_X X)^{-1} \\
&= (t_X X)^{-1} t_X \sigma^2 I X (t_X X)^{-1} = \sigma^2 (t_X X)^{-1} \\
E(||\widehat{\beta}||_2^2|X) &= E(t_Y X(tXX)^{-2} tXY|X)) = Trace(X(tXX)^{-2} tX\sigma^2 I) + ||\beta||_2^2 \\
&= \sigma^2 Trace((tXX)^{-1}) + ||\beta||_2^2 \\
MSE(\widehat{\beta}) &= E(||\widehat{\beta} - \beta||_2^2|X) = E(||\widehat{\beta}||_2^2|X) - ||\beta||_2^2 \\
&= \sigma^2 Trace((tXX)^{-1}) = \sigma^2 \sum \frac{1}{\lambda_i}
\end{aligned}$$

where MSE denotes the mean squared error and $(\lambda)_i$ denote the eigen values of $t_X X$. Notice that the OLS estimator is unbiased, but can show an arbitrary high MSE when the matrix $t_X X$ has close to 0 eigen values.In the out of sample period, $\widehat{Y} = X\widehat{\beta}$ is used to predict the target. As seen in Sect. 2, the corresponding trading strategy

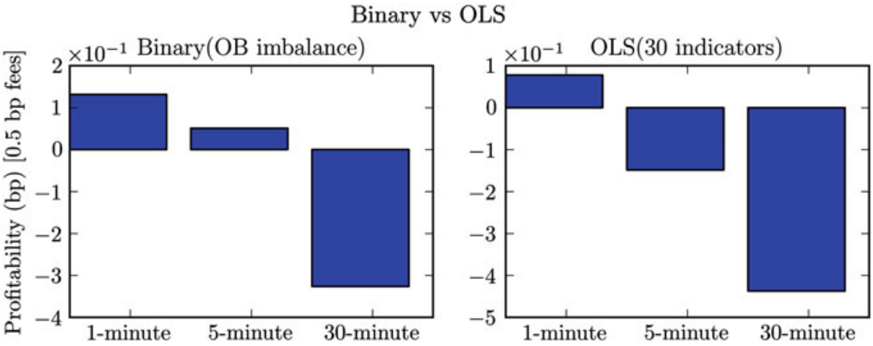

Fig. 7 The quality of the OLS prediction: the results of the OLS method are not better than those of the binary one

is to buy (respectively sell) 100,000 euros when $\widehat{Y} > 0$ (respectively $\widehat{Y} < 0$). To measure the quality of the predictions, the binary method based on the order book imbalance indicator is taken as a benchmark. The linear regression is computed using 30 indicators, including the order book imbalance, thus it should perform at least as well as the binary method. Figure 7 compares the profitabilities of the two strategies. The detailed statistics per stock are given in Tables 21, 22 and 23 of Appendix 3. Similar to the binary method, the performance of the OLS method decrease with an increasing horizon. Moreover, the surprising result is that when combining all the 30 indicators, the results are not better than just applying the binary method to the order book imbalance indicator. This leads to questioning the quality of the regression.

Figure 8 gives some example of the OLS regression coefficients. It is observed that the coefficients are not stable over the time. For example, for some period, the regression coefficient of the order book imbalance indicator is negative. This does not make any financial sense. In fact, when the imbalance is high, the order book shows more liquidity on the bid side (participants willing to buy) than the ask side (participants willing to sell). This state of the order book is observed on average before an up move—i.e. a positive return. The regression coefficient should, thus, be always positive. It is also observed that, for highly correlated indicators, the regression coefficients might be so different. This result also does not make sense, since one would expect to have close coefficients for similar indicators.

From a statistical view, this is explained by the high MSE caused by the high colinearity between the variables. In the following paragraphs, the numerical view is also addressed and some popular solutions to the OLS estimation problems are tested.

4.2 Ridge Regression

When solving a linear system $AX = B$, with A invertible, if a small change in the coefficient matrix (A) or a small change in the right hand side (B) results in a

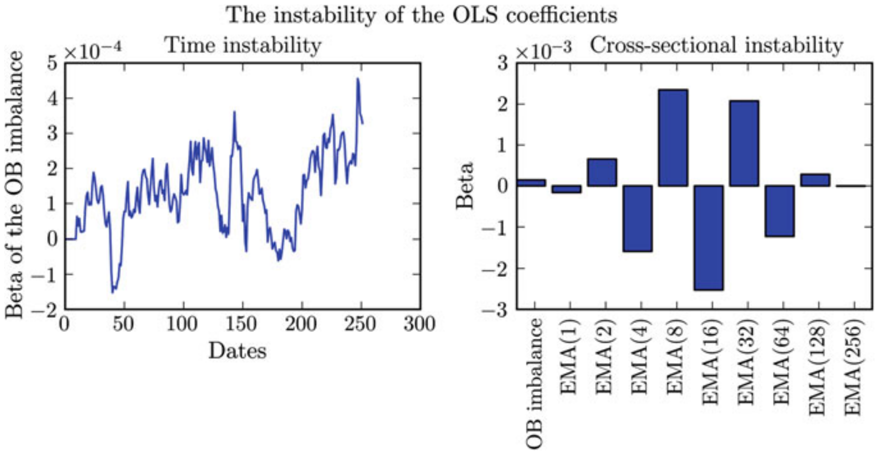

Fig. 8 The quality of the OLS prediction: the graph on the *left* shows the instability of the regression coefficient of the order book imbalance indicator over the year 2013 for the stock Deutsh Telecom. The graph on the *right* shows, for a random day, a very different coefficients for similar indicators; the order book imbalance and its exponential moving averages

large change in the solution vector (X) the system is considered ill-conditioned. The resolution of the system might give a non reliable solution which seems to satisfy the system very well.

An example of an ill-conditioned system is given bellow:

$$\begin{bmatrix} 1.000 & 2.000 \\ 3.000 & 5.999 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.000 \\ 11.999 \end{bmatrix} => \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2.000 \\ 1.000 \end{bmatrix}$$

When making a small change in the matrix A:

$$\begin{bmatrix} 1.001 & 2.000 \\ 3.000 & 5.999 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.000 \\ 11.999 \end{bmatrix} => \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -0.400 \\ 2.200 \end{bmatrix}$$

When making a small change in the vector B:

$$\begin{bmatrix} 1.000 & 2.000 \\ 3.000 & 5.999 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 11.999 \end{bmatrix} => \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -3.999 \\ 4.000 \end{bmatrix}$$

When dealing with experimental data, it is not reliable to have a completely different calibration because of a small change in the observations. Hence, it is mandatory to take into consideration such effects before achieving any computation.

In literature, various measures of the ill-conditioning of a matrix have been proposed [13], perhaps the most popular one [14] is $K(A) = ||A||_2||A^{-1}||_2$, where $||.||_2$ denotes the l_2-norm defined for a vector X as $||X||_2 = \sqrt{{}^tXX}$ and for a matrix A as $||A||_2 = \max\limits_{||X||_2 \neq 0} \frac{||AX||_2}{||X||_2}$. The larger is $K(A)$, the more ill-conditioned is A.

The rationale behind defining the condition number $K(A)$ is to measure the sensitivity of the solution X relative to a perturbation of the matrix A or the vector B. More precisely, it is proved (see the Appendix 4) that:

- If $AX = B$ and $A(X + \delta X) = B + \delta B$ then $\frac{||\delta X||_2}{||X||_2} \leq K(A)\frac{||\delta B||_2}{||B||_2}$
- If $AX = B$ and $(A + \delta A)(X + \delta X) = B$ then $\frac{||\delta X||_2}{||X+\delta X||_2} \leq K(A)\frac{||\delta A||_2}{||A||_2}$

Notice that $K(A)$ can be easily computed as the maximum singular value of A. For example, in the system above, $K(A) = 49,988$. The small perturbations can, thus, be amplified by almost 50,000, causing the previous observations.

Figure 9 represents the singular values of $t_X X$ used to compute the regression of the right graph of Fig. 8. The graph shows a hard decreasing singular values. In particular, the condition number is higher than 80,000.

This finding explains the instability observed on the previous section. Moreover that the OLS estimator is statistically not satisfactory, the numerical problems due to the ill-conditioning of the matrix makes the result numerically unreliable.

One popular solution to enhance the stability of the estimation of the regression coefficients is the Ridge method. This method was introduced independently by A. Tikhonov, in the context of solving ill-posed problems, around the middle of the 20th century, and by A.E. Hoerl in the context of addressing the linear regression problems by the sixteeth. The Ridge regression consists of adding a regularization term to the original OLS problem:

$$\widehat{\beta_\Gamma} = argmin_\beta(||Y - X\beta||_2^2 + ||\Gamma\beta||_2^2)$$

The new term gives preference to a particular solution with desirable properties. Γ is called the Tikhonov matrix and chosen usually as a multiple of the identity matrix; $\lambda_R I$, where $\lambda_R \geq 0$. The new estimator of the linear regression coefficients is called the Ridge estimator, denoted $\widehat{\beta}_R$, and defined as follows:

$$\widehat{\beta}_R = argmin_\beta(||Y - X\beta||_2^2 + \lambda_R||\beta||_2^2)$$

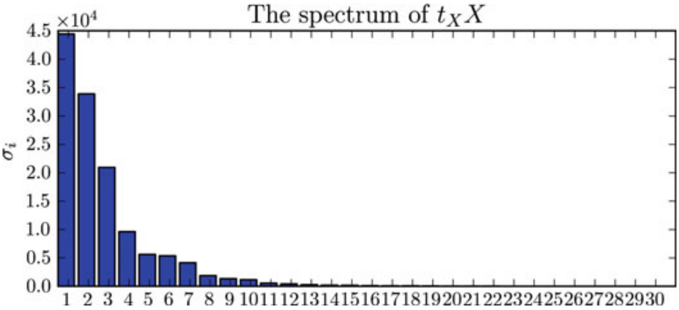

Fig. 9 The quality of the OLS prediction: the graph shows that the matrix inverted when computing the OLS coefficient is ill-conditioned