Paper: “Revealing Pairs-trading opportunities with long short-term memory networks”

Authors: Andrea Flori and Daniele Regoli

Journal: European Journal of Operational Research

Publisher: Elsevier

Volume: 295(2)

Pages: 772-791

Year: 2021

Published Journal Article available at: https://doi.org/10.1016/j.ejor.2021.03.009

Revealing Pairs-Trading Opportunities with Long Short-Term Memory Networks

Andrea Flori*, Daniele Regoli†

Abstract

This work examines a deep learning approach to complement investors' practices for the identification of pairs-trading opportunities among cointegrated stocks. We refer to the reversal effect, consisting in the fact that temporarily market deviations are likely to correct and finally converge again, to generate valuable pairs-trading signals based on the application of Long Short-Term Memory networks (LSTM). Specifically, we propose to use the LSTM to estimate the probability of a stock to exhibit increasing market returns in the near future compared to its peers, and we compare and combine these predictions with trading practices based on sorting stocks according to either price or returns gaps. In so doing, we investigate the ability of our proposed approach to provide valuable signals under different perspectives including variations in the investment horizons, transaction costs and weighting schemes. Our analysis shows that strategies including such predictions can contribute to improve portfolio performances providing predictive signals whose information content goes above and beyond the one embedded in both price and returns gaps.

Finance; Machine learning; Pairs-trading; Statistical arbitrage; Neural networks

1 Introduction

The reversal effect documented by Fama (1965), Jegadeesh (1990), Jegadeesh and Titman (1993), and Lehmann (1990) states that stocks that have recently performed poorly will probably undergo a larger market reversal in the future. Past market performances may, in fact, influence the investment attitudes of the traders and losing stocks might experience higher reversals. For instance, this may occur because stocks become more volatile, thus impacting on the provision of liquidity, or because they might be affected by certain market behaviors such as fire sales, which cause excessive drops in their market prices but foster their subsequent rebounds (see, e.g., Lasfer et al. 2003, Huang et al. 2009, Da and Gao 2010, Cheng et al. 2017).

More generally, three main explanations, not necessarily mutually exclusive, have been proposed to motivate the emergence of the reversal effect: the influence of liquidity effects (Jegadeesh and Titman, 1995b, Chordia et al., 2002, Avramov et al., 2006), the non-synchronous trading of small vs. large capitalized stocks (Lo and MacKinlay, 1990, Boudoukh et al., 1994, Richards, 1997), and the investors' cognitive biases and reactions to new information or shocks (Lehmann, 1990, Jegadeesh and Titman, 1995a, Subrahmanyam, 2005). The first explanation refers to microstructural phenomena such as inventory imbalances by market makers which determine reversal profits as a compensation for bearing inventory risk. The second one relates to the market size and assumes that the reversal effect generates

*Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Via Lambruschini 4/B, 20156 Milano, Italy

†Big Data Lab, Intesa Sanpaolo S.p.A., Corso Inghilterra 3, 10138, Torino, Italy. The views and opinions expressed are those of the authors and do not necessarily reflect the views of Intesa Sanpaolo, its affiliates or its employees; do not constitute an offer, solicitation of an offer, or any advice or recommendation, to purchase any securities or other financial instruments, and may not be construed as such.

profits especially among small-cap stocks. Finally, the third explanation is based on a behavioral sentiment-based perspective, in which market overreaction by impatient traders temporarily affects stock demand. These aspects together provide support for the interpretation that past performances influence traders' behavior and liquidity provision, thus potentially contributing to the formation of profits from contrarian strategies. However, within the framework of efficient markets, returns are memoryless stochastic processes and prices react immediately as new information becomes available, making it impossible for investors to exploit past information to predict future returns (Fama, 1970). Therefore, the investigation of the reversal effect, as an analysis of detection of predictive signals, finds its place in the literature on market anomalies, in contrast with the efficient market hypothesis. As a consequence, the reversal effect has been investigated as a market anomaly that can be exploited to build several types of allocation strategies (as outlined, e.g., in Chan et al. 1996, Coval and Stafford 2007, Blitz et al. 2013, Da et al. 2013, Hameed and Mian 2015, Blackburn and Cakici 2017).

Against this background, scholars and practitioners have struggled to challenge the efficiency of financial markets applying different techniques in search of investment opportunities. In recent years, with the rise of computation resources and their widespread availability, machine learning techniques are gaining momentum within the finance community and several approaches have been proposed to extract valuable information from financial data both to leverage statistical arbitrage opportunities in financial markets (see, e.g. Andreou et al. 2008, Atsalakis and Valavanis 2009, Bekiros 2010, Huck 2009, 2010, Sermpinis et al. 2013, Patel et al. 2015, Heaton et al. 2017, Krauss et al. 2017, Fischer and Krauss 2018, Gu et al. 2018, Huck 2019, Schnaubelt et al. 2020) and for more general finance-related pursuits (see Kim et al. 2020, Kraus et al. 2020 and references therein).

Following a growing literature that exploits a large-scale use of deep learning concepts to spot patterns in financial markets, in this paper we attempt to study the complex non-linear relationships among and within financial time series by applying the technique of the *Long Short-term Memory Networks (LSTM)* to predict the market performances of a large sample of stocks. LSTM networks for financial applications are also considered, e.g., in Bao et al. (2017), Troiano et al. (2018), Fischer and Krauss (2018), and Borovkova and Tsiamas (2019). Here, in line with previous studies (see, e.g., Huck 2009, 2010, Krauss et al. 2017, Fischer and Krauss 2018), we refer to stocks composing the S&P 500 index as our perimeter of analysis and we test the predictions of the LSTM within the framework of investment strategies based on pairs-trading.

Pairs-trading relies, in fact, on the apparent profitability of a strategy in which stocks with similar past performances start to exhibit opposite, possibly temporarily, market patterns. Practically, once having identified stocks that tended to behave in a "similar" way in the past, traders try to exploit potential short-term relative mispricings by buying the relatively underperforming stocks and taking short positions in overperforming stocks. If the future performances resemble those of the past, then the relative mispricing is a temporary deviation, and market prices are likely to correct and finally converge again, thereby generating profits. Pairs-trading represents, therefore, a zero-cost framework to investigate reversal effects by means of a (possibly) market neutral strategy simultaneously establishing both a long and short position in two stocks with each position having the same dollar amount. When the underperforming stock regains value and the outperforming stock declines, then profits are generated while controlling for risks by maintaining a low exposure to market dynamics. Considerable effort has been spent, therefore, both by scholars and practitioners to detect and exploit this market anomaly to build return-based pairwise relative value trading strategies (see, e.g., Vidyamurthy 2004, Gatev et al. 2006, Chen et al. 2017, Do and Faff 2012, Blitz et al. 2013, Jacobs and Weber 2015).

In our work, to better define the perimeter of those stocks whose temporarily market deviations may signal pairs-trading opportunities, we first identify for each stock the existence of peers sharing *cointegrated* market patterns. In fact, the spread between two cointegrated series follows a stationary process, meaning that deviations from the mean are only temporary and eventually revert. Hence, since pairs-trading opportunities are perceived as deviations from the equilibrium due to market reactions, which are temporary and will be timely corrected, we rely only on those stocks having at least one cointegrated peer as candidate stocks for pairs-trading strategies. We describe the application of an LSTM architecture to quantify such emerging deviations. More specifically, we employ an LSTM based on information from returns and trading volumes to generate predictions on the probability of a stock exhibiting increasing market returns in the near future compared to its peers of cointegrated stocks. Then, we extensively compare such predictions with common trading practices based on both price and returns gaps. However, the goal of the study is not the design of a more performing indicator than those typically employed to build pairs-trading strategies. Instead, with the inclusion of the outcomes of the LSTM we aim to verify whether it is possible to extract valuable signals from financial time series containing information that can "complement" the one already embedded in price or returns gaps, thereby generating even better portfolio performances once jointly combined.

In so doing, we first construct strategies investing in portfolio of stocks grouped according to sorting criteria based on either price or returns gaps (see, e.g., Gatev et al. 2006, Rad et al. 2016, Chen et al. 2017, Krauss 2017), or our proposed outcomes from the LSTM. Effectively, pairs-trading strategies can be constructed by buying the top and selling the bottom stocks with respect to the given sorting criterion, i.e. ranked on either their price or returns gaps with respect to peers or their LSTM outcomes. We show that such *Top-Bottom* strategies based on gaps in either prices or returns generate over the whole period from January 2003 to June 2019 gross returns of about 15% and 19% (Sharpe ratios: 1.03 and 1.17), respectively. These results are in line with the 18.3% raw performance (Sharpe ratio: 1.23) of the *Top-Bottom* strategy based on the LSTM outcomes. Interestingly, we note that such performances are economically significant even when we control for the volatility of returns and that do not vanish once we take into account factor exposures.

Then, by applying double sorting procedures on such portfolios, we show that our proposed approach based on the LSTM outcomes is able to generate informative predictive signals that go above and beyond the ones embedded in both price and returns sorting criteria, and that, once these sorting criteria are jointly combined, the LSTM outcomes can contribute to improve portfolio performances. For instance, conditional sorts show that, controlling for price gap, the annualized alpha performances of the *Top-Bottom* strategies ranked by the LSTM outcomes range between about 7.7% and 22% and that the average annualized alpha constructed by equally investing in each *Top-Bottom* portfolio from the double sorting procedure is about 16%. These results thus support the use of the LSTM outcomes to complement the information provided by the other sorting criteria and build portfolios by jointly sorting stocks according to either price or returns gaps together with the LSTM outcomes ranking. Specifically, the *Top-Bottom* strategy, which goes long in the top stocks and short in the bottom ones according to the conditional sorting based on both price gaps and the LSTM outcomes ranking, produces an extra-performance of about 23%. Similar findings are observed when returns gaps and LSTM outcomes are jointly considered.

This double sorts analysis allows us to investigate our main claim about the importance of disentangling whether an observed gap in market patterns, which is a typical criterion used by practitioners to initialize a pairs-trading allocation, is going to increase or revert, for which therefore opposite buy

or sell market signals should be provided. We show that the information jointly provided by the LSTM with either price or returns gaps is able to better identify the buy or sell signals, thus improving portfolio performances. Hence, the outcomes of the LSTM contribute to reinforce the associated buy or sell signals by identifying for each stock the likelihood of an enlarging or reducing market gap with respect to its peers. For instance, our analysis reveals that when the price gap indicator signals to purchase recently underperformed stocks, the outcomes of the LSTM contribute to reinforce such signal by further indicating those stocks that have higher probability to increase in the near future their market returns with respect to peers. By contrast, the joint information tells us to sell those recently overperformed stocks that are also highly expected to show a substantial market reversal behaviour in the near future with respect to peers. The double sorts procedure thus helps us to assess such relationships by providing a synthetic representation of the joint distribution.

We then propose to relate these portfolio performances to factor exposures. We discuss how strategies based on the joint information relate to exposures to the momentum or the short-term reversal factors, which are basically at the ground level of an enlarging or reducing market gap. From an operational point of view, our findings support the purpose of identifying investment strategies specifically devoted to extract such market dynamics, suggesting that the outcomes of the LSTM contribute to improve portfolio performance by indicating whether the observed market gap of a stock with respect to its peers is going to increase or decrease, thereby generating opposite buy or sell signals.

Finally, we present and discuss how such approach can be generalized to various holding periods and investment settings as well as alternative machine learning methodologies. Although our analysis shows that the LSTM outperforms other techniques, in this work we do not advocate its advantages over other machine learning methodologies that can be applied for similar tasks. With this regard, the choice of the LSTM algorithm is here only instrumental to extract signals for deviating market patterns among peer stocks using time series information. This aspect is crucial in our framework, since we are not proposing a certain machine learning architecture to maximize portfolio performance, but instead we are using signals generated by the LSTM to perform a strategy in which the identification of deviation from peers is of fundamental importance. Our analysis reveals that our proposed *Top-Bottom* strategies, which in this context mimic the pairs-trading strategies, are able to generate valuable portfolio performances, especially once information from both the LSTM and traditional price or returns gaps indicators are jointly combined.

The rest of the paper is organized as follows. Section 2 describes the data sample and explains the methodological aspects related to the LSTM architecture and the portfolio construction. Section 3 shows the empirical results, comparing different approaches for pairs-trading strategies and discussing the contribution of the LSTM outcomes to portfolio performance and market factors exposures. Section 4 presents several robustness analysis including the role of different investment horizons, transaction costs, sectorial composition, LSTM formulation and alternative machine learning techniques. Section 5 is devoted to conclusions.

2 Data & Methodology

2.1 Data

We consider stocks approximately composing the S&P 500 index in the period from the beginning of January 2000 to the end of June 2019. We collect these components on a yearly basis at the beginning

of January, i.e. at the time of the year in which we are going to modify the pool of stocks we make analysis on. Given a 3-year train window rolling yearly (more details in the next sections), and a sample period from January 2000 to the end of June 2019, our out-of-sample analysis consists of 16.5 years. Finally, as factor exposures we refer to those reported in the Kenneth R. French website, while stocks profile information are collected from Orbis Bureau van Dijk.[1]

2.2 Computing Software

All data analysis are performed via the open source software `R` (R Core Team, 2019). In particular, the LSTM network is built and trained using the R interface to `Keras` (Chollet et al., 2015), while "standard" machine learning algorithms are trained via `caret` (Kuhn, 2020), an `R` wrapper library for predictive modeling. For all portfolio analysis, we employ the `PerformanceAnalytics R` package (Peterson and Carl, 2019).

The estimation time for the computation of cointegration groups (see subsection 2.3) is on average about 3 hours per training block, while the LSTM training (see subsection 2.5.4) takes about 2 hours per training block. This time refers to computations performed on common CPUs (Intel i7-7500U 2.70GHz×4, 16GB RAM).

2.3 Cointegration groups

Several competing approaches have been proposed in the literature to detect pairs-trading opportunities. Krauss (2017) reviews some attempts which relate the identification of the candidate pairs to, for instance, distance metric approaches, cointegration tests, the extent of mean-reverting spread in time series, and stochastic control frameworks.

In our work we opt for the cointegration approach as a theoretical framework for identifying pairs-trading opportunities as temporarily deviations of mean reverting processes between stocks that have shown to follow equilibrium relationships (see, e.g., Vidyamurthy 2004, Dunis and Ho 2005, Puspaningrum et al. 2010, Huck and Afawubo 2015, Rad et al. 2016, Krauss 2017). Our approach to find cointegrated time series thus proceeds with the following 4 steps: 1) at the beginning of each year t we collect the historical adjusted close price for the past 3 years, namely in the interval $[t-3, t-1]$, for each stock S_t^i present in the S&P 500 index at t; 2) for each possible couple $(S_t^i,\ S_t^j)$, we test for cointegration using two different testing approaches: the Engle-Granger two-step method (Engle and Granger, 1987), and the Johansen test (Johansen, 1991). 3) In order to avoid false positives, we deem cointegrated – in formula $S_t^i \overset{ci}{\sim} S_t^j$ – couples that are cointegrated according to both tests with a significance level of 1%. 4) For each stock i and year t, we define its *contegration group* CG_t^i as the set of all stocks cointegrated to i at that time:

$$CG_t^i = \left\{ S_t^j : S_t^i \overset{ci}{\sim} S_t^j \right\}. \tag{1}$$

Note that the symmetry of the cointegration relation $\overset{ci}{\sim}$ (which is in principle true) is imposed whenever the cointegration test outputs an asymmetric result: namely, if the couple (S_t^i, S_t^j) gives a positive result for both tests but (S_t^j, S_t^i) does not, then $S_t^i \overset{ci}{\sim} S_t^j$ is considered true anyway. Finally, once the cointegration groups CG_t^i are defined for each stock i at the beginning of year t, the set of groups $\{CG_t^i\}_i$ is held true for the whole year t, and all pair relations are computed with respect to

[1] See: `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/` and `https://orbis.bvdinfo.com/`.

this set.

Table 1 shows some summary statistics of the set of cointegration groups for each year t: on average, around 200 stocks (out of the ∼500 of S&P 500) results in at least one cointegration group and, typically, half of the cointegration groups has less than 5 stocks. In addition, we observe a quite erratic behaviour of the stocks belonging to the cointegration groups, with a consistent in-out dynamics of their respective members.[2] For instance, the "turnover" of stocks between two subsequent years, computed by the Jaccard similarity index among the sets of all stocks present in at least one cointegration group at year $t-1$ and at year t, indicates that on average only one third of stocks are persistent from one year to the next with an oscillating behavior, above 40% in 2004, 2005 and in 2011, 2012 and below 30% during the years of the global financial crisis, in 2013, 2014, and at the end of the sample period.

Table 1: Summary statistics of cointegration groups. The table shows, for each year under study and on average, the total number of stocks present in at least one cointegration group; the median of the group sizes; the Jaccard similarity index and the overlap coefficient between the stocks in the corresponding year and in the preceding one.

	average	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
total number of cointegrated stocks	196	191	219	234	143	135	126	190	211	332	214	146	225	190	214	167	158	238
median size	4.7	5	10	7	3	3	3	4	4	9.5	5	3	5	4	4	4	3	4
Jaccard similarity (%)	33	-	45	49	36	29	26	26	37	47	44	29	25	33	30	29	25	25
overlap similarity (%)	57	-	66	68	70	46	43	52	57	82	79	56	51	54	50	51	41	50

2.4 Pairs-trading approaches

Since the deviation of two cointegrated series should follow a stationary process, then deviations from its long-run mean should be absorbed and revert to the mean. Similarly, pairs-trading opportunities are perceived as deviations from the equilibrium due to market reaction on, for instance, some news or changes in the market structure, which are temporary and will be timely corrected. Hence, once candidate groups have been identified by cointegration, we introduce some criteria to distinguish observed deviations by relying on a very generalized framework which reflects, consistently with the cointegration stocks filtering, the most common approaches proposed by scholars and practitioners.

For instance, the idea of pairs-trading opportunities based on deviations in terms of prices goes back to the seminal paper of Gatev et al. (2006), where authors provide argumentation based on the description of how traders select pairs. Instead, gaps in terms of returns are investigated for instance in Chen et al. (2017) to assess to which extent stocks that significantly under-perform (over-perform) their peers are able to experience abnormally higher (lower) returns in the future. Indeed, our work tests pairs-trading opportunities by employing criteria based on these two main paradigms, referring to the identification of gaps in either prices or returns between pairs of stocks that are previously identified as being cointegrated (see subsection 2.3).

Specifically, we compute the following variables to be used as gap-ranking criteria:

[2] Given two sets A and B, the overlap similarity is $\frac{|A\cap B|}{\min(|A|,|B|)}$, namely the fraction of the smaller set contained in the larger. The Jaccard similarity is instead defined as $\frac{|A\cap B|}{|A\cup B|}$, and is much more sensitive to the size difference among the two sets.

$$\text{price gap} \quad \begin{cases} \Delta p_t^i = p_t^i - p_t^{CG,i}, \\ \Delta^\beta p_t^i = p_t^i - \beta^{p,i} p_t^{CG,i} \end{cases} \tag{2a}$$

$$\text{returns gap} \quad \begin{cases} \Delta r_t^i = r_t^i - r_t^{CG,i}, \\ \Delta^\beta r_t^i = r_t^i - \beta^{r,i} r_t^{CG,i}, \end{cases} \tag{2b}$$

where: p_t^i and r_t^i are the normalized[3] adjusted close price of stock i on day t and its (daily) simple return, respectively; $p_t^{CG,i}$ and $r_t^{CG,i}$ are the average normalized adjusted close price and return of the peers of stock i (i.e., its cointegration group); $\beta^{p,i}$ and $\beta^{r,i}$ are defined via the following OLS regressions for each stock i, estimated on the previous 3-year interval:

$$p_t^i = \beta^{p,i} p_t^{CG,i} + \varepsilon_t^i, \quad \varepsilon_t^i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2);$$
$$r_t^i = \beta^{r,i} r_t^{CG,i} + \eta_t^i, \quad \eta_t^i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2).$$

In words, we construct the price gap for each stock i first by normalizing all the prices in order for them to be on the same scale, and then by taking the difference between its normalized price at time t and the average normalized price of its peers. We label this criterion Δp. We also provide a variant of this indicator: following Rad et al. (2016), we consider the residuals of an OLS model where we regress the stock price versus its peers average price. This case is indicated as $\Delta^\beta p$. Similarly, for gaps in terms of returns we consider the difference of the returns between each stock i and its peers for each time t (namely, Δr). As for the case of price gaps, we also take the variant in which we extract the residuals from an OLS regression, namely $\Delta^\beta r$.

For all the four variables in (2), we build standardized variants as well, where we normalize each indicator with its standard deviation computed over the preceding 3-year training interval. We represent these variants with an overset tilde, namely $\widetilde{\Delta p}$, $\widetilde{\Delta^\beta p}$, $\widetilde{\Delta r}$, $\widetilde{\Delta^\beta r}$. According to Gutierrez and Prinsky (2007), standardization favors a representation of the residuals that better differentiates information versus noise.

Blitz et al. (2013) find that short-term reversal strategies based on residual stock returns that do not exhibit dynamic exposures to factors generate substantially higher performances than conventional short-term reversal strategies. Hence, this motivated our decision to include also this approach to provide a more comprehensive picture of the methodologies that have been proposed so far to investigate reversal effects. Besides the variables defined in (2) and their standardized variants, we thus rely on an indicator in line with the one proposed by Blitz et al. (2013), where a conditional factor model is applied to measure the residuals in the following sense: for each stock i we fit a factor model on the previous 3-year window, rolling daily; we then take the last value of the residuals ϵ of this model, thus having a residual value for each day. The resulting series of residuals for each stock is then used as indicator. We consider both the three factor model (Fama and French, 1993) and a seven factor model that includes the momentum (Carhart, 1997) and the short-term reversal (Jegadeesh and Titman, 1993) factors into the five factors of Fama and French (2015). These indicators are labeled as ϵ_3 and ϵ_7, respectively. These additional ranking criteria should therefore capture the presence of additional information not embedded in the typical market factors and that could be exploited to extract meaningful signals from stocks time series. It should be noted, however, that the core of our

[3]Each price is rescaled to be 1 in the first day of the 3-year training window to get values on the same scale.

analysis mainly relates to the contribution of our proposed indicator from the LSTM outcomes to signals provided by traditional gaps in terms of prices or returns.

Finally, we propose a new indicator to capture the gap between a stock and its peers: the probability that the return difference Δr will increase in the near future. We indicate this variable as $\mathrm{P}(\Delta r \nearrow)$ throughout next sections. We compute $\mathrm{P}(\Delta r \nearrow)$ via an LSTM network as described in subsection 2.5, where the target variable y_τ is the up-down movement of Δr within a h-day horizon:

$$\begin{cases} y_\tau = 1 & \text{when } \Delta r \text{ increases in } h \text{ days,} \\ y_\tau = 0 & \text{when } \Delta r \text{ decreases in } h \text{ days.} \end{cases} \tag{3}$$

Thus, the LSTM output $\hat{y}_\tau$ is precisely the probability of Δr increasing in a h-day horizon. Differently from the indicators Δp and Δr and their variants, the proposed indicator $\mathrm{P}(\Delta r \nearrow)$ is based not only on the information available at the day when the investment decision is taken, but it also takes into account the market history of the involved time series in order to enrich the prediction. Moreover, this indicator is not a mere snapshot or collection of past information, but learns from past (and present) data the useful patterns for making a prediction on future movements, weighting appropriately the different time contributions.

The idea inspiring this indicator is that, in determining the presence of a gap among series of peers, an important piece of information is whether the gap is in an expanding or in a closing phase. If the gap is going to disappear in the short term, then a reversal approach is to be taken, as in the pairs-trading framework discussed up to now; instead if the gap is expanding, then profit is to be searched in a momentum-like strategy. This issue will be specifically addressed in subsection 3.4.

2.5 LSTM networks

Long Short-Term Memory networks are an instance of the broader class of Recurrent Neural Networks (RNN), specifically designed to deal with sequential data. LSTM networks, first introduced in Hochreiter and Schmidhuber (1997), aim to solve some specific drawbacks of RNN, namely the vanishing/exploding gradient issue that prevents recurrent networks to learn long-term dependencies within sequences, as pointed out e.g. in Bengio et al. (1994). Below, we briefly introduce the structure and main ingredients of RNN and LSTM networks, while we refer to Graves (2012) and Goodfellow et al. (2016) for a more comprehensive review.

The structure of RNN, from which LSTM derives, is composed by an input layer, in which a sequence of data $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\tau)$ (in general a multivariate time series) enters the network one step at a time, one or more hidden layers for each time step $(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_\tau)$, and an output layer that is chosen appropriately with respect to the problem.[4] Our study considers a single classification output $\boldsymbol{y}_\tau$ at the end of the sequence, but in general RNN can deal with sequence-to-sequence predictions as well. Thus, for the present case, the RNN is trying to model the following relation:

$$\boldsymbol{y}_\tau = f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\tau). \tag{4}$$

The key feature of RNN, as opposed to others Artificial Neural Networks, is that the hidden layers have an autoregressive nature, namely

$$\boldsymbol{h}_t = \phi(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t; \boldsymbol{\theta}), \tag{5}$$

[4] We use bold letters to indicate vectors and uppercase bold for matrices.

where the vector $\boldsymbol{\theta}$ summarizes the parameters defining the family of functions ϕ. For the "vanilla" RNN with one hidden layer and classification output, the model can be summarized by the following system of equations:

$$\boldsymbol{h}_t = \tanh\left(\boldsymbol{W}\boldsymbol{h}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}^h\right), \quad t = 1, \dots, \tau \tag{6a}$$

$$\hat{\boldsymbol{y}}_\tau = \text{softmax}\left(\boldsymbol{V}\boldsymbol{h}_\tau + \boldsymbol{b}^y\right); \tag{6b}$$

where $(\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{b}^h, \boldsymbol{b}^y)$ are the parameters to be calibrated from data, "tanh" is the hidden layer activation function (that is intended to be applied, as usual, component-wise), and

$$\text{softmax}(\boldsymbol{z})_i = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{7}$$

is needed to output probabilities in the classification case, effectively reducing to a logit in the binary case.

The LSTM model is slightly more involved than (6), introducing three gated units: a forget gate $\boldsymbol{f}$, an external input gate $\boldsymbol{g}$ and an output gate $\boldsymbol{o}$, together with another internal state $\boldsymbol{s}$ (see Figure 1):

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}^f\boldsymbol{h}_{t-1} + \boldsymbol{U}^f\boldsymbol{x}_t + \boldsymbol{b}^f\right), \quad t = 1, \dots, \tau \tag{8a}$$

$$\boldsymbol{g}_t = \sigma\left(\boldsymbol{W}^g\boldsymbol{h}_{t-1} + \boldsymbol{U}^g\boldsymbol{x}_t + \boldsymbol{b}^g\right), \quad t = 1, \dots, \tau \tag{8b}$$

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}^o\boldsymbol{h}_{t-1} + \boldsymbol{U}^o\boldsymbol{x}_t + \boldsymbol{b}^o\right), \quad t = 1, \dots, \tau \tag{8c}$$

$$\boldsymbol{s}_t = \boldsymbol{f}_t \otimes \boldsymbol{s}_{t-1} + \boldsymbol{g}_t \otimes \tanh\left(\boldsymbol{W}^s\boldsymbol{h}_{t-1} + \boldsymbol{U}^s\boldsymbol{x}_t + \boldsymbol{b}^s\right), \quad t = 1, \dots, \tau \tag{8d}$$

$$\boldsymbol{h}_t = \tanh\left(\boldsymbol{s}_t\right) \otimes \boldsymbol{o}_t, \quad t = 1, \dots, \tau \tag{8e}$$

$$\hat{\boldsymbol{y}}_\tau = \text{softmax}\left(\boldsymbol{V}\boldsymbol{h}_\tau + \boldsymbol{b}^y\right); \tag{8f}$$

where $\otimes$ denotes the element-wise multiplication operator and σ the sigmoid activation function:

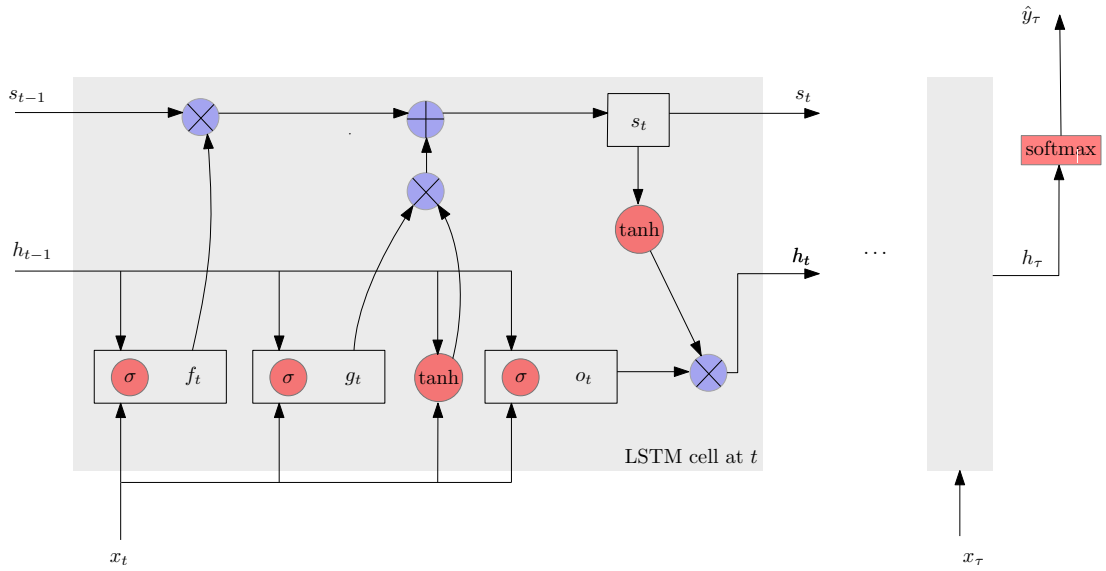
$$\sigma(x) = \frac{1}{1 + e^{-x}}\ . \tag{9}$$

The idea behind LSTM networks is to use gated units to control the flow of information coming from past states in order to avoid the well known problem of standard RNN in learning from long sequences due to the vanishing/exploding gradient problem (Bengio et al., 1994). The key point in this respect is the (conditional) self-loop in the internal state (8d), where the information from step $t-1$ is directly pushed to step t, conditionally only on the activation of the forget gate $\boldsymbol{f}$. Specifically, the forget gate $\boldsymbol{f}$ determines how much of the past state $\boldsymbol{s}_{t-1}$ is evolved directly to $\boldsymbol{s}_t$; the external input gate $\boldsymbol{g}$ determines the way in which the past hidden layer contributes to $\boldsymbol{s}$; while the output gate $\boldsymbol{o}$ accounts for the on/off state of the current hidden layer $\boldsymbol{h}_t$. Thus, the information is easily propagated back in time along the sequence from state $\boldsymbol{s}_t$ to $\boldsymbol{s}_{t-1}$ through gate $\boldsymbol{f}$, without the issue of "losing the memory" by successive multiplication of small gradients[5].

The number of parameters of LSTM is much greater than that of a "vanilla" RNN. Specifically, for a single LSTM layer with m input features and n hidden nodes, the number of parameters is $4(n(m+1)+n^2)$, since we have 4 "copies" of the triplet $(\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{b})$. The output layer accounts for other $K(n+1)$ parameters, where K is the number of classes. Notice that the length of the time

[5]Models with direct memory self loops and gated units controlling the flow of information through time steps are called *gated RNN*. Another state-of-the art gated RNN is the Gated Recurrent Unit network (Cho et al., 2014).

Figure 1: LSTM. Schematic visualization of an LSTM cell described by equation (8).



sequence τ does not impact the complexity of the model, since parameters are time-invariant in all RNN networks; the time dependence is entirely accounted for by the autoregressive nature of the hidden states.

2.5.1 Training and test sets

We organize the predictive framework in the following way: in line with what discussed for cointegration groups (see subsection 2.3) we take blocks of 4 years $[t-3, t]$ where we use the first 3 years $[t-3, t-1]$ as *training set* and the last year t as *test set* or *live set*, and we roll the block yearly, effectively producing a continuous time interval of non-overlapping live sets. The training set is used to calibrate parameters[6], while the live set is the out-of-sample period where we make predictions using the (trained) LSTM, actually computing the $\mathrm{P}(\Delta r \nearrow)$ indicator for each stock. Since our data span an interval from January 2000 to June 2019, we actually implement 17 blocks, with the last live set lasting for half year only. In each 3-year training window, we reserve the last 10% of data as *validation set* (see subsection 2.5.4 for more details). The entire backtest procedure is summarized by Algorithm 2.5.1.

[h!]

InputinputOutputoutput price time series in the period 2000-01-01//2019-06-31 for stocks in S&P500. $\mathrm{P}(\Delta r \nearrow)$ in all days in the (out-of-sample) period 2003-01-01//2019-06-30 for all cointegrated stocks.

loop over years year t in the period 2003-01-01//2019-06-30

yearly cointegration groups

stocks S^i present in S&P500 at t stocks S^j with $j \neq i$ present in S&P500 at t Engle-Granger two-step test and Johansen test for the (normalized) daily price series $\{p_s^i\}$ and $\{p_s^j\}$ with $s \in [t-3, t-1]$ build the cointegrated group $CG_t^i = \left\{ S_t^j : S_t^i \overset{ci}{\sim} S_t^j \right\}$ Cointegration groups for the year t for all stocks.

LSTM yearly training stocks $S^i \in \bigcup_j CG_t^j$ compute $\Delta r_s^i = r_s^i - r_s^{CG_t^i}$ for day $s \in [t-3, t-1]$ build the LSTM input τ-sequences as by eq. (10) employing Δr_s^i, r_s^i, V_s^i for day $s \in [t-3, t-1]$ build the LSTM output labels with h-day horizon $y_s^i = \mathrm{sign}(\Delta r_{s+h} - \Delta r_s)$

train the LSTM on all the computed τ-sequences of all stocks a model LSTM_t trained with data up to t with which make out-of-sample predictions for year t.

out-of-sample predictions days $\theta \in t$ stocks $S^i \in \bigcup_j CG_t^j$ build the LSTM input τ-sequence with $(\Delta r_{\theta-\tau+1}^i, \ldots, \Delta r_\theta^i)$,

[6] The parameters of the LSTM in this case, but in general all the needed parameters, are calibrated in the 3-year training set, such as the OLS betas in eq. (2), or the standard deviations for the standardized versions of the indicators.

$(r^i_{\theta-\tau+1}, \ldots, r^i_\theta)$, $(V^i_{\theta-\tau+1}, \ldots, V^i_\theta)$ use the model LSTM_t with the just computed input τ-sequence to get $\hat{y}^i_\theta = \text{P}\left(\Delta r^i \nearrow\right)$
Out-of-sample predictions for all days in year t for all stocks belonging to at least one cointegration group.

2.5.2 Input features

The input features for the LSTM are multivariate time series. Following Fischer and Krauss (2018), for each stock i we take sequences with a lookback period roughly 1-year long, $\tau = 240$ days, with: past values of the Δr^i gap with respect to its peers, past trading volume (V^i), past returns (r^i). We compute these features for every stock present in at least one cointegration group, and we slide each sequence daily. Hence, for each stock in the 3-year training window, we collect about $750 - 240$ input sequences, since for each training period of 3 solar years the first $\tau = 240$ days are used to form the first sequence. Then, by sliding forward of one day at a time, we get roughly $750 - 240$ sequences per stock. The input sequences are thus of the form:

$$\begin{pmatrix} \Delta r^i_\theta & \Delta r^i_{\theta-1} & \ldots & \Delta r^i_{\theta-\tau+1} \\ V^i_\theta & V^i_{\theta-1} & \ldots & V^i_{\theta-\tau+1} \\ r^i_\theta & r^i_{\theta-1} & \ldots & r^i_{\theta-\tau+1} \end{pmatrix}, \tag{10}$$

with day θ varying along the 3-year training window. More precisely, if we are in year t, thus in the training years $[t-3, t-1]$, indicating by t_0 the first day of year t, then $\theta \in [t_0 - 3\text{years} + \tau, t_0 - h]$.

Given that the number of considered stocks is, on average, roughly 200 (see Table 1), we get approximately 100,000 training sequences for each training round. The same structure of input sequences is then computed for the out-of-sample year as well.

Each of the three input variables is standardized with the overall mean and standard deviation computed on the training set, which is known to ensure a more rapid learning (see, e.g., Bishop 2006).

2.5.3 LSTM architecture

The architecture of the LSTM we train is the following: an input layer with 3 features and 240 lookback time dimension, a single hidden layer with 35 nodes, and a 2-node dense output layer with softmax activation.

Between the LSTM layer and the output layer Batch Normalization is applied, which rescales the LSTM layer outputs with 2 additional (trainable) parameters for each node: one scale and one shift parameter. Batch normalization, first introduced in Ioffe and Szegedy (2015) as a powerful tool for stabilizing the learning process, was discovered to provide regularization against overfitting similar to Dropout, even if the actual reasons for this are yet not fully understood (Goodfellow et al., 2016), but are likely due to extra-randomness injected during the training process.

This structure was decided by trial and error, and then held fixed throughout the entire time interval, without any automatic hyperparameter tuning. The total number of parameters amounts to 5602: 5460 for the LSTM hidden layer, $35 \times 2 + 2 = 72$ for the output layer and 35×2 for the batch normalization layer.

2.5.4 LSTM training

The fitting of the LSTM weights to the training sequences is a (highly non-convex) minimization problem, where we set the crossentropy as the target loss function. We employ *RMSprop* stochastic descent learning algorithm (introduced in Tieleman and Hinton 2012), with learning rate 0.001 and

decay factor 0.9. Moreover, we make use of *early stopping* in order to choose the optimal out-of-sample number of epochs: we retain the last 10% of each training sequence as a *validation set*, and we stop training when the validation crossentropy does not decrease for 10 consecutive epochs (known as *patience* interval), with a cap of 50.[7]

2.5.5 LSTM prediction performance

Here we discuss the out-of-sample performance of the LSTM network above introduced. As previously stated, the out-of-sample period consists of 16.5 years, each of which is the test set of a different model.

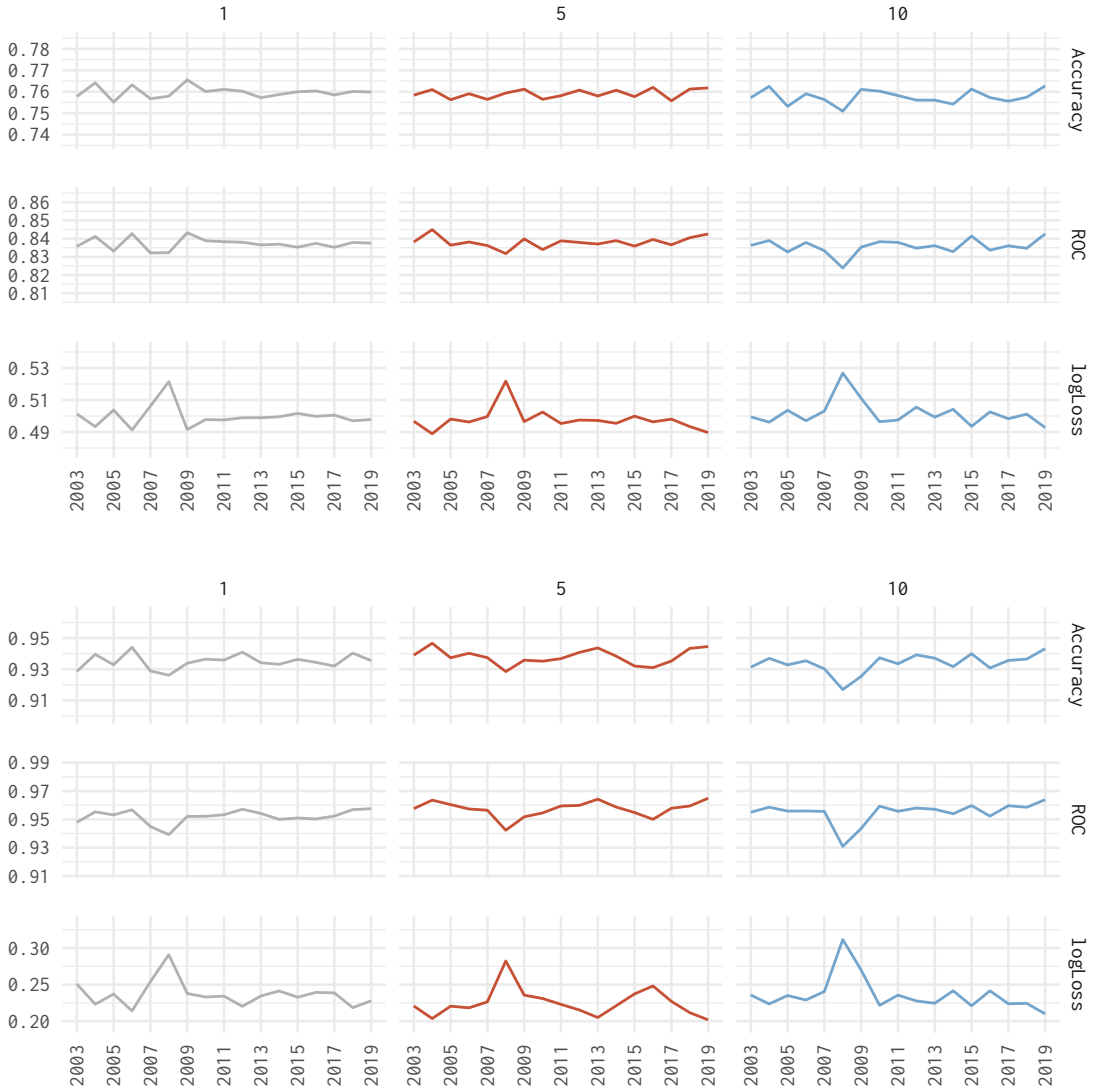


Figure 2: LSTM performance. Yearly out-of-sample prediction performance of the LSTM network discussed in subsection 2.5 for investment horizon $h = 1$ (left panel, in gray), $h = 5$ (middle panel, in red) and $h = 10$ (right panel, in blue). Top panel refers to the entire distribution of stocks, while bottom panel to the signals related to stocks composing the *Top-Bottom* strategies. We report the accuracy (with a 0.50 threshold), the area under the ROC curve and the logloss for each out-of-sample year over the period from January 2003 to June 2019.

Figure 2 reports three different classification metrics, namely: accuracy, area under the ROC curve, and logloss (or crossentropy) for three investment horizons ($h = 1, 5, 10$).[8] As a reference,

[7] In the case of stopping, the optimal parameters are the ones referring to the beginning of the patience period.

[8] Accuracy is the only threshold dependent metric. We display accuracy with a 0.50 probability threshold, namely we predict an up movement whenever the estimated probability $\mathrm{P}(\Delta r \nearrow)$ is greater or equal to 50%. Both ROC and logloss are based on the estimated probability $\mathrm{P}(\Delta r \nearrow)$.

recall that when the number of up and down movements is roughly the same, as in this case, the area under the ROC for a random classifier is about 0.5, while the accuracy and logloss are about 0.5 and $\log(2) \simeq 0.693$, respectively. Top panel of Figure 2 reports the performances related to the entire distribution of stocks. Note how the LSTM shows a good performance in terms of all the three displayed metrics. Performances result quite stable over the years, with some fluctuations in the years of the financial crisis of 2007-08, and show a slight deterioration and more erraticity for longer investment horizons h. More interestingly, the prediction performance is even higher (see bottom panel of Figure 2) if we consider those stocks related to the pairs-trading framework and constituting the top and bottom deciles of the $\mathrm{P}(\Delta r \nearrow)$ distribution.

2.6 Portfolio construction

To analyze the role of our proposed indicators, we consider illustrative investment strategies which equally invest in stocks grouped in percentiles according to a selected ranking measure among those discussed in subsection 2.4. In subsection 4.3, for the LSTM case we also propose a linear weighting scheme as a function of the value of the sorting variable for each stock.

More practically, we construct portfolios either separately for each ranking criterion or jointly by combining information of pairs of ranking measures. For each ranking indicator, we employ the same procedure of rebalancing portfolio on a daily basis using information available up to the previous day. We assume a holding period of one day (i.e., $h = 1$) and the corresponding returns are linked in time to form series for every portfolio. In subsection 4.1, we also discuss longer investment horizons (namely, $h = 5, 10$ business days). For each year t, the sample of stocks that is considered for the analysis is composed by the stocks included in the S&P 500 index at the beginning of the year and that have found to be cointegrated at least with one other series at the end of the previous year using a look back period of three years (i.e., on the interval $[t-3, t-1]$). This sample is maintained for the entire year t. In the main results of the manuscript we neglect transactions costs, which are instead introduced in subsection 4.1.

3 Empirical Analysis

3.1 Comparing different approaches for pairs-trading opportunities

We investigate the predictive capacity of the indicators introduced in subsection 2.4 by relying on one-way procedures which invest in decile portfolios over an investment horizon of 1 day. We also report in Table 2 the *Top-Bottom* strategies which buy the top and sell the bottom decile portfolios with respect to a given sorting criterion, thus constituting zero-cost investment strategies.

The ranking criterion based on Δp and its normalized variant $\widetilde{\Delta p}$ indicates an almost monotonic pattern of the annualized returns, with the first deciles portfolios that systematically overperform the last ones. As a consequence, the *Top-Bottom* strategy is able to generate a consistent and statistically significant performance (i.e., $11.09\% - 14.56\%$). These results are qualitatively similar to the ones obtained by the indicators $\Delta^{\beta} p$ and $\widetilde{\Delta^{\beta} p}$ (i.e., $8.08\% - 10.94\%$). Hence, sorting stocks by means of the price gap with respect to peers seems to support previous empirical findings that detected pairs-trading opportunities based on temporary deviations of the price dynamics.

We run the same comparisons but considering as sorting criterion, instead of price gaps, differences in returns. Note how also the indicators Δr and $\widetilde{\Delta r}$ present clear monotonic patterns which contribute

Table 2: One-Way Sorts. The table shows the raw returns (in percentage) obtained by equally investing in stocks composing decile portfolios defined using as sorting criteria those discussed in subsection 2.4. The holding period is $h = 1$ day. Newey-West t-statistics are reported in parenthesis. Data are annualized and refer to the period from January 2003 to June 2019.

	1	2	3	4	5	6	7	8	9	10	*Top-Bottom*
Δp	21.79	15.27	13.37	16.60	13.41	12.27	10.06	8.81	10.84	9.55	11.09
	(3.48)	(3.36)	(3.10)	(3.84)	(3.47)	(3.40)	(2.84)	(2.65)	(3.02)	(2.37)	(2.47)
$\widetilde{\Delta p}$	21.66	16.66	15.06	12.18	15.20	10.76	13.41	12.14	9.11	6.32	14.56
	(3.90)	(3.30)	(3.48)	(2.89)	(3.68)	(2.91)	(3.41)	(3.36)	(2.65)	(1.84)	(3.83)
$\Delta^{\beta} p$	17.70	17.19	19.41	15.15	14.41	12.18	12.49	8.20	7.74	8.75	8.08
	(3.49)	(3.68)	(4.20)	(3.62)	(3.61)	(3.17)	(3.34)	(2.45)	(2.12)	(2.27)	(2.48)
$\widetilde{\Delta^{\beta} p}$	20.15	16.80	16.64	14.82	16.63	12.25	12.18	9.96	5.61	8.25	10.94
	(4.08)	(3.52)	(3.74)	(3.50)	(4.03)	(2.90)	(3.23)	(2.77)	(1.71)	(2.28)	(3.38)
Δr	25.50	19.83	17.20	15.55	14.11	10.59	9.27	6.91	7.95	6.40	16.42
	(4.58)	(4.31)	(4.15)	(4.01)	(3.56)	(2.74)	(2.49)	(2.00)	(2.05)	(1.62)	(4.04)
$\widetilde{\Delta r}$	26.85	19.00	17.56	15.31	14.20	10.66	8.47	7.41	8.72	5.47	18.96
	(5.21)	(4.05)	(4.28)	(3.78)	(3.46)	(2.66)	(2.18)	(2.07)	(2.23)	(1.59)	(4.88)
$\Delta^{\beta} r$	26.62	18.08	21.72	13.94	13.61	10.91	8.89	5.88	10.80	3.27	20.86
	(4.57)	(4.21)	(5.10)	(3.69)	(3.28)	(2.89)	(2.55)	(1.72)	(2.54)	(1.16)	(4.53)
$\widetilde{\Delta^{\beta} r}$	24.15	22.14	20.14	13.50	16.13	11.68	9.05	6.79	7.43	3.14	18.86
	(4.54)	(5.07)	(4.57)	(3.27)	(3.75)	(2.93)	(2.32)	(1.90)	(2.09)	(1.13)	(4.54)
ϵ_3	22.45	23.07	22.10	14.92	12.93	14.43	8.14	6.26	4.86	4.89	15.26
	(4.53)	(4.94)	(4.82)	(3.64)	(3.11)	(3.45)	(2.05)	(1.82)	(1.44)	(1.44)	(3.75)
ϵ_7	23.26	19.12	18.94	15.56	14.97	11.49	8.84	8.41	6.65	6.36	14.74
	(4.73)	(4.20)	(4.40)	(3.68)	(3.72)	(2.82)	(2.19)	(2.16)	(1.87)	(1.69)	(3.74)
$\mathrm{P}(\Delta r \nearrow)$	4.04	7.12	8.58	9.25	10.91	16.24	13.32	21.85	18.31	24.57	18.32
	(1.27)	(1.91)	(2.24)	(2.37)	(2.83)	(3.98)	(3.31)	(4.81)	(4.04)	(5.14)	(4.98)

to determine even higher *Top-Bottom* strategy performances (i.e., $16.42\% - 18.96\%$) than the analog in terms of price gaps. Interestingly, these performances are also very similar to the ones obtained by ranking stocks according to the residuals of the fitted OLS models (namely, $\Delta^{\beta} r$ and $\widetilde{\Delta^{\beta} r}$, respectively). More generally, those stocks with larger positive differences in returns with respect to their peers are thus more likely to underperform the latter in the near future.

The one-way performances for the sorting criterion based on the residuals of the three or seven factor models appear largely in line with the previous ones and able to effectively distinguish across decile portfolios. More importantly, it is worth noticing that using a more or less parsimonious factor model does not really seem to matter in the evaluation, being the resulting performances generally quite similar across the decile portfolios of the two factor models. Hence, the inclusion of many drivers of factor risks does not seem to substantially vary the ranking performances, thus questioning the neutralization extent of controlling for factors such as short-term reversal and momentum. Blitz et al. (2013) find that ranking stocks according to residuals from a 3-factor model generates profitable results that do not exhibit dynamic exposures to the Fama and French (1993) factors. In our study, we extend the residuals computation by including a wider list of exposures to risks (namely, the 7-factor residuals ϵ_7) and, even controlling for factors such as short-term reversal and momentum, we still observe profitable opportunities in the *Top-Bottom* strategy. It seems, therefore, that there exist other reasons than these factor exposures behind the emergence of valuable investment opportunities, which can be thus captured by systematically investing in stocks according to the ranking positioning of the corresponding residuals.

Finally, we observe that our proposed metric $\mathrm{P}(\Delta r \nearrow)$ is able to sort the annualized returns of the decile portfolios and obtain an economically and statistically significant performance in the *Top-*

Bottom strategy[9] in line with the highest performing cases of Table 2. It is worth noting that the LSTM prediction does not refer to a simple quantification of market returns, but it represents instead the probability of a stock to get increasing market returns in the near future with respect to the average value of its peers only, thus not in relation to the overall sample. As an example, those stocks placed in the top decile portfolio in a given time period according to $\mathrm{P}(\Delta r \nearrow)$ do not necessarily refer to the set of stocks in the sample with the highest predicted market appreciation in the near future: they stand for the set of stocks with the highest probability of showing a market increase with respect to the average market returns of their corresponding cointegrated peers. Hence, the top decile portfolio is not indicating the predicted top performer stocks in the sample (and, similarly, the bottom decile portfolio is not referring to the worst stocks), but it stands for those stocks with the highest probability to deviate from their peers reaching higher market returns over an investment horizon h.

Then, in order to select a few appropriate alternatives for our proposed sorting indicator $\mathrm{P}(\Delta r \nearrow)$ to be used to further investigate to which extent the outcomes of the LSTM can complement the information already embedded in either price or returns gaps, we decide to include a risk-adjusted assessment of portfolio performances. We report different formulations of Sharpe ratios using as measures of risk the standard deviation, the modified Cornish-Fisher VaR and the Expected Shortfall. Table 3 reports for each sorting criterion the corresponding risk-adjusted indicators. Overall the monotonic patterns observed in Table 2 appear less pronounced, thus suggesting less clear relationships between the decile portfolios' performances and their corresponding levels of returns dispersion. We note that *Top-Bottom* strategies based on normalized variants provide in general better results, thus suggesting that controlling for historical dispersion can favor the out-of-sample stability of the risk-adjusted performances, and are also in line with performances including the residuals of OLS models.

To assess the statistical significance of the reported Sharpe ratios, we perform an analysis with two different benchmarks. The first benchmark is inspired by the well known Malkiel's monkeys strategy (Malkiel, 1973): we simulate 10,000 investors that each day in the period under study select randomly two tenths of the stocks present in at least one cointegration group and then go long in one tenth and short in the other. These investors thus pick stocks among those passing the "cointegration filter", but then they do not use any particular indicator to select the top and bottom deciles. The resulting Sharpe ratios for the whole sample period of all the 10,000 "monkeys" are summarized in Figure 3: the median value is slightly negative (-4.6%), the 99% percentile is 54.3%, i.e. lower than the Sharpe ratio reached by all *Top-Bottom* strategies here considered (Table 3), and even the best performing monkey, reaching 94.8%, is well below the value reached by the *Top-Bottom* decile strategy based on $\mathrm{P}(\Delta r \nearrow)$ ($\sim 123\%$), and also below the value reached by well-performing *Top-Bottom* strategies based on other ranking indicators (e.g., $\widetilde{\Delta r} \sim 117\%$ and $\widetilde{\Delta p} \sim 103\%$).

The second benchmark is the S&P 500 index, against which we perform a studentized bootstrap test following Ledoit and Wolf (2008). Namely, a circular block resampling with optimal block size is computed on the return time series of each *Top-Bottom* strategy and on the S&P 500 daily returns series. An asymmetric two-sample t-test is then performed with the alternative hypothesis that the pairs-trading strategy has a higher Sharpe ratio than a simple buy&hold strategy on the S&P 500 index. Results are as follows: the *Top-Bottom* strategy based on our proposed indicator $\mathrm{P}(\Delta r \nearrow)$ has the lowest p-value (0.030), showing a significant difference with respect to the S&P 500, while $\widetilde{\Delta r}$ and

[9]Note that while for price and returns gaps the pairs-trading approach suggests that high positive (negative) value of the gap corresponds to subsequent low (high) performance, for our indicator the natural interpretation suggests that high (low) value of the indicator corresponds to subsequent high (low) performance. This is indeed confirmed by results shown in Table 5.

Table 3: One-Way Sorts Risk-adjusted Performances. The table shows the Sharpe ratios obtained by equally investing in stocks composing decile portfolios defined using as sorting criteria those discussed in subsection 2.4. Ratios are computed using as measure of risk the standard deviation (SR), the Modified Cornish-Fisher VaR (Mod_SR) or the Expected Shortfall (ES). Values are expressed in percentage. The holding period is $h = 1$ day. Data are annualized and refer to the period from January 2003 to June 2019.

		1	2	3	4	5	6	7	8	9	10	*Top-Bottom*
	SR	76.82	64.85	62.10	77.72	63.13	59.26	48.99	44.39	54.65	43.22	61.13
Δp	Mod_SR	78.77	66.22	54.32	63.88	53.51	53.01	42.38	35.28	40.40	31.84	76.32
	ES	41.61	35.74	34.20	41.09	34.63	32.79	28.11	25.87	30.56	25.81	33.15
	SR	86.28	70.19	67.02	55.45	68.23	49.19	63.89	60.03	46.82	31.50	102.54
$\widetilde{\Delta p}$	Mod_SR	75.86	63.76	64.62	48.14	62.25	42.38	48.55	46.70	34.83	25.13	108.62
	ES	45.20	38.12	36.53	31.31	37.05	28.48	34.93	33.06	26.91	19.95	51.46
	SR	70.09	70.52	88.12	70.36	70.21	58.82	60.26	40.19	37.85	40.60	56.93
$\Delta^{\beta} p$	Mod_SR	65.21	69.25	71.05	58.53	57.17	49.74	48.74	32.17	30.04	31.20	69.03
	ES	38.32	38.36	45.67	37.87	37.66	32.59	33.25	24.06	22.98	24.49	30.54
	SR	85.81	72.89	72.19	67.09	76.93	56.53	56.28	48.61	27.92	41.10	87.33
$\widetilde{\Delta^{\beta} p}$	Mod_SR	80.69	63.44	68.20	53.76	64.81	46.74	43.49	38.98	23.30	31.56	132.64
	ES	44.82	39.21	38.89	36.53	40.77	31.74	31.63	27.93	18.28	24.40	44.41
	SR	97.01	87.21	80.91	77.19	69.07	53.04	46.36	33.34	37.10	25.07	94.98
Δr	Mod_SR	90.82	73.37	63.84	57.71	57.76	44.15	37.73	30.27	31.21	26.99	77.02
	ES	49.77	45.35	42.47	40.70	37.13	29.84	26.80	20.94	22.85	18.19	48.26
	SR	109.95	84.86	80.91	73.17	66.95	51.13	40.46	35.48	40.52	23.30	117.21
$\widetilde{\Delta r}$	Mod_SR	100.84	70.08	66.58	53.83	57.16	42.34	34.83	29.94	35.22	23.89	92.55
	ES	54.98	44.30	42.52	39.03	36.31	29.14	24.28	21.99	24.44	16.89	58.02
	SR	95.40	77.76	104.44	69.97	68.71	56.78	45.45	28.95	49.09	12.36	110.27
$\Delta^{\beta} r$	Mod_SR	89.04	78.04	78.96	55.57	54.08	41.94	37.35	25.94	41.02	18.09	88.64
	ES	49.24	41.33	52.57	37.46	36.89	31.41	26.29	18.80	28.45	12.43	54.99
	SR	96.11	98.12	93.07	64.05	77.63	56.60	43.60	32.48	34.47	13.18	106.86
$\widetilde{\Delta^{\beta} r}$	Mod_SR	89.65	90.77	80.05	50.32	58.49	43.03	35.25	28.05	30.66	18.42	86.16
	ES	49.31	49.97	47.75	35.01	40.98	31.59	25.69	20.59	21.66	12.17	53.50
	SR	91.88	104.70	101.94	69.42	61.20	68.65	38.66	29.82	22.25	20.74	88.71
ϵ_3	Mod_SR	83.24	86.88	85.63	54.60	51.13	54.36	31.42	28.71	21.95	23.70	71.87
	ES	47.49	52.72	51.54	37.46	33.73	37.04	23.50	19.35	16.01	15.70	45.47
	SR	93.82	85.87	88.12	73.08	71.89	55.05	42.47	39.57	30.70	27.29	90.49
ϵ_7	Mod_SR	108.10	76.71	69.53	54.66	54.98	44.78	33.85	34.78	27.84	27.99	79.48
	ES	48.31	44.72	45.62	39.05	38.46	30.93	25.19	23.94	19.92	18.71	46.18
	SR	17.33	33.07	40.26	43.87	52.23	76.25	61.91	99.99	81.90	106.59	122.74
$\mathrm{P}(\Delta r \nearrow)$	Mod_SR	20.50	28.57	35.16	36.49	44.38	64.80	51.32	78.85	69.71	83.48	92.61
	ES	14.03	21.00	24.28	25.89	29.65	40.42	34.12	50.72	43.02	53.56	60.50

$\widetilde{\Delta p}$ follow closely (p-values 0.048 and 0.058, respectively), and $\Delta^{\beta} r$ and $\widetilde{\Delta^{\beta} r}$ are weakly significant with p-values equal to 0.054 and 0.064, respectively. Δp and $\Delta^{\beta} p$ show the lowest significance levels (p-values 0.503 and 0.647, respectively).

In order to investigate whether and to which extent $\mathrm{P}(\Delta r \nearrow)$ contributes to provide additional predictive signals, from the combined information reported by Tables 2-3, summarized also in Figure 4, we thus propose to select two main ranking metrics alternative to $\mathrm{P}(\Delta r \nearrow)$, one referring to a sorting criterion based on a dynamic gap in prices and the other based on a dynamic gap in returns. For the sake of simplicity, we also decide to select for both price and returns gaps the same version along those proposed in subsection 2.4. For these reasons, the aforementioned evidences on the risk-adjusted performances suggest us to rely on $\widetilde{\Delta p}$ and $\widetilde{\Delta r}$ as illustrative indicators against which compare the strategies based on $\mathrm{P}(\Delta r \nearrow)$.

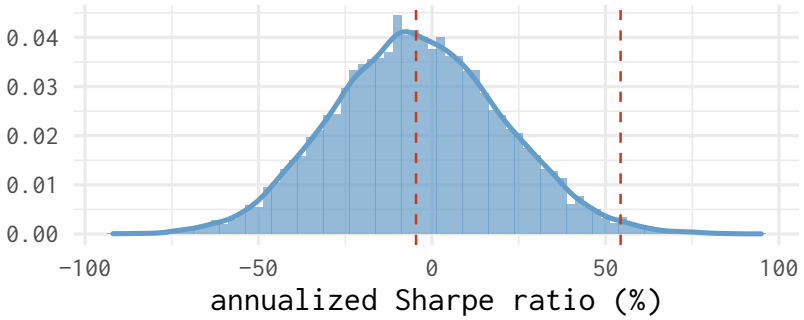


Figure 3: Sharpe ratios distribution (in percentage) of 10,000 investors who daily pick 2/10 of all stocks belonging to at least one cointegration group, going long in 1/10 and short in the remaining 1/10 in the period under study. Red vertical lines denotes the median (-4.6%) and the 99% percentile (54.3%). The highest reached Sharpe ratio is 94.8%.

Figure 4: One-Way Sort *Top-Bottom* Performances. Compounded gross returns (blu bars, right axis) and Sharpe ratios (red bars, left axis) for *Top-Bottom* decile portfolios as reported in Tables 2-3. The holding period is $h = 1$ day. Data are annualized and refer to the period from January 2003 to June 2019. Highlighted bars point out the 3 indicators chosen for detailed comparison.

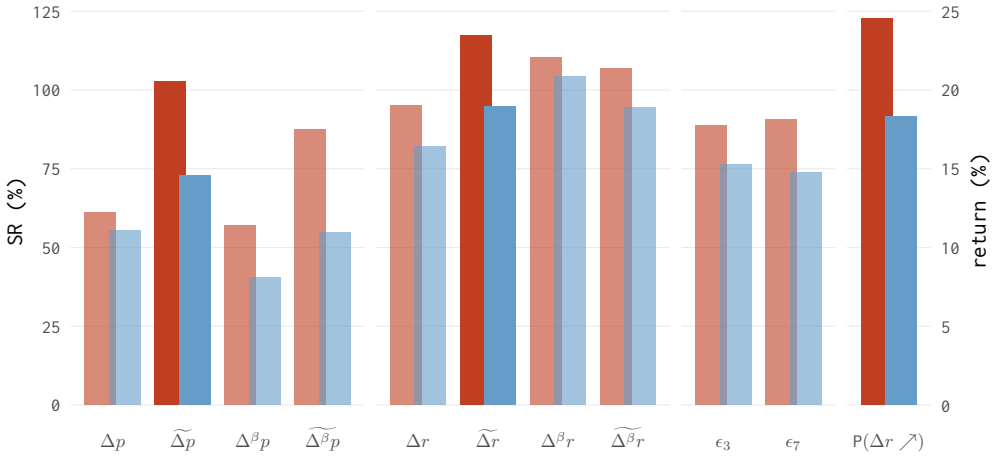


Table 4: Profit & Loss Analysis. Gross returns, standard deviation of returns (σ), Sharpe ratio (SR), Modified Cornish-Fisher VaR (Mod_SR), Expected Shortfall (ES), maximum draw-down (DD), Conditional Value at Risk at level 99% ($\text{CVaR}_{99\%}$), Omega ratio (Ω), and Pain Index (PI) for the *Top-Bottom* strategies based on $\widetilde{\Delta p}$, $\widetilde{\Delta r}$, and $\text{P}(\Delta r \nearrow)$. The holding period is $h = 1$. Newey-West t-statistics are reported in parentheses. The reference period is from January 2003 to June 2019.

		2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019-Q1-2
$\widetilde{\Delta p}$	return	10.76 (0.90)	8.69 (1.04)	6.70 (0.67)	7.45 (0.90)	-6.82 (-0.60)	120.28 (2.50)	108.45 (2.21)	20.37 (2.19)	1.25 (0.23)	6.87 (1.07)	-3.02 (-0.36)	1.28 (0.19)	6.50 (0.71)	5.31 (0.50)	-0.51 (0.00)	2.35 (0.32)	17.93 (1.77)
	σ	11.54	10.42	9.91	8.10	10.84	26.55	33.05	8.44	13.74	7.54	8.27	8.76	9.36	12.66	10.70	10.85	7.52
	SR	93.28	83.38	67.61	92.02	-62.95	453.01	328.18	241.22	9.07	91.04	-36.47	14.62	69.41	41.93	-4.78	21.69	238.45
	Mod_SR	60.37	55.94	47.86	57.23	-34.98	251.84	188.49	136.19	10.07	64.81	-20.29	13.97	44.88	30.16	0.36	16.24	171.43
	ES	46.34	43.92	38.82	42.46	-26.76	199.73	155.00	97.59	7.52	49.88	-16.61	13.16	33.55	22.74	0.30	12.49	139.27
	DD	8.44	6.70	11.50	5.21	13.66	12.24	20.83	3.52	13.15	5.75	7.05	9.05	7.03	8.90	12.89	5.81	2.58
	$\text{CVaR}_{99\%}$	-2.05	-1.73	-1.37	-1.57	-2.09	-5.37	-9.79	-1.82	-2.94	-2.17	-1.30	-3.18	-1.81	-2.52	-1.43	-1.99	-0.93
	Ω	1.16	1.15	1.12	1.16	0.91	1.81	1.63	1.44	1.03	1.17	0.95	1.03	1.13	1.08	1.00	1.04	1.45
	PI	0.02	0.02	0.03	0.02	0.03	0.02	0.03	0.01	0.04	0.01	0.04	0.02	0.02	0.03	0.07	0.03	0.01
$\widetilde{\Delta r}$	return	-0.16 (0.04)	-0.07 (0.05)	4.57 (0.60)	11.94 (1.19)	7.79 (0.76)	128.46 (3.15)	84.15 (2.16)	18.66 (1.85)	29.60 (2.09)	13.45 (1.45)	15.84 (1.86)	11.38 (0.90)	6.38 (0.79)	8.29 (0.57)	0.95 (0.12)	28.53 (1.97)	-4.26 (-0.23)
	σ	11.89	10.89	10.41	10.68	13.32	35.52	32.12	10.40	12.21	8.42	7.65	12.36	8.95	20.77	9.18	11.58	11.50
	SR	-1.34	-0.68	43.95	111.73	58.47	361.65	262.01	179.51	242.38	159.70	207.07	92.07	71.28	39.92	10.32	246.35	-37.05
	Mod_SR	2.68	2.70	28.59	74.71	43.41	160.35	157.71	115.98	164.70	100.05	132.84	58.05	45.27	30.75	9.01	163.20	-17.43
	ES	1.96	1.94	18.98	55.46	31.39	88.70	127.12	78.11	123.46	76.50	93.17	22.43	32.35	17.13	6.21	128.16	-11.72
	DD	14.15	10.94	8.48	8.23	14.82	12.48	14.19	7.03	8.96	6.95	4.07	8.32	8.53	13.35	13.24	6.53	13.33
	$\text{CVaR}_{99\%}$	-2.48	-2.42	-2.27	-2.38	-3.07	-7.37	-8.06	-2.47	-3.19	-1.47	-1.71	-4.32	-1.88	-4.93	-2.05	-1.85	-2.82
	Ω	1.01	1.01	1.08	1.21	1.12	1.55	1.45	1.34	1.47	1.28	1.38	1.19	1.13	1.10	1.03	1.46	0.95
	PI	0.05	0.05	0.03	0.02	0.04	0.03	0.05	0.02	0.02	0.02	0.01	0.03	0.02	0.04	0.06	0.02	0.07
$\text{P}(\Delta r \nearrow)$	return	6.93 (0.57)	7.42 (0.70)	4.72 (0.57)	24.62 (2.29)	12.40 (1.14)	85.51 (2.30)	61.67 (1.91)	20.79 (2.18)	26.54 (1.81)	15.00 (1.62)	15.75 (1.94)	14.89 (1.34)	6.57 (0.76)	6.50 (0.43)	-5.65 (-0.51)	19.49 (1.43)	11.80 (0.79)
	σ	12.54	11.16	9.84	10.14	12.36	29.37	30.39	10.24	12.19	8.81	7.32	10.38	9.07	19.61	9.41	11.64	10.89
	SR	55.30	66.45	47.93	242.84	100.32	291.19	202.94	202.96	217.75	170.36	215.09	143.43	72.47	33.13	-60.00	167.49	108.33
	Mod_SR	35.20	41.67	30.28	156.28	73.33	138.03	132.03	123.29	146.59	105.80	140.90	88.81	45.25	26.51	-35.04	109.57	63.71
	ES	25.81	29.22	19.72	113.54	58.64	80.63	108.72	77.33	106.06	78.14	103.12	37.63	32.00	14.84	-28.28	87.41	46.17
	DD	8.14	7.43	10.62	7.41	7.31	15.57	14.61	6.07	8.06	5.27	5.23	6.71	8.92	13.04	13.93	11.47	9.18
	$\text{CVaR}_{99\%}$	-2.59	-2.39	-2.14	-2.24	-3.25	-5.88	-9.47	-2.11	-3.04	-1.68	-1.55	-3.24	-1.92	-4.71	-1.53	-1.70	-2.32
	Ω	1.10	1.12	1.09	1.46	1.20	1.48	1.38	1.38	1.41	1.31	1.39	1.29	1.13	1.08	0.91	1.30	1.20
	PI	0.02	0.03	0.04	0.02	0.02	0.03	0.05	0.02	0.02	0.01	0.01	0.02	0.02	0.04	0.07	0.04	0.04

3.2 Risk-adjusted evaluation of the strategies

Before exploring how $\widetilde{\Delta p}$, $\widetilde{\Delta r}$ and $\mathrm{P}(\Delta r \nearrow)$ provide complementary predictive signals, in this subsection we further investigate the performance and risk characteristics of the *Top-Bottom* strategies based on such indicators. As performance measure, we follow previous subsections and we rely on the compounded returns gross of trading costs and fees. We then employ the standard deviation of the returns (σ), the Sharpe ratio (SR), the Modified Cornish-Fisher VaR (Mod_SR), the Expected Shortfall (ES), the maximum draw-down (DD), the Conditional Value at Risk at level 99% ($\mathrm{CVaR}_{99\%}$), the Omega ratio (Ω), and the Pain Index (PI) to assess the risk levels of the strategies. Table 4 shows the estimates, reported on annual basis, for the *Top-Bottom* strategies based on the reference criteria $\widetilde{\Delta p}$, $\widetilde{\Delta r}$ and $\mathrm{P}(\Delta r \nearrow)$, while Figure 5 shows the corresponding market patterns in time.

Figure 5: *Top-Bottom* Performances. Daily compounded returns gross of trading costs and fees. The holding period is $h = 1$. Color lines refer to the *Top-Bottom* strategies for: $\widetilde{\Delta p}$ (grey), $\widetilde{\Delta r}$ (blue) and $\mathrm{P}(\Delta r \nearrow)$ (red); the dark line stands for the market compounded returns in the same period.

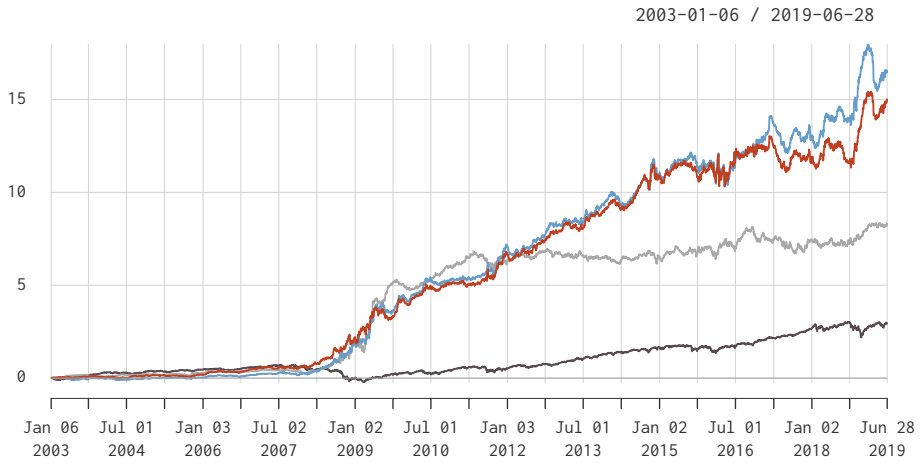


Table 4 reveals some interesting findings. Firstly, in the years of the global financial crisis each strategy generates consistent returns which are only in part nuanced by higher levels of volatility (see, e.g., rows for σ), as also supported by values of SR, Mod_SR and ES that are consistently higher in that interval. These results are also confirmed by the gain-loss ratio Ω (see Keating and Shadwick 2002, Kazemi et al. 2004), which helps enforcing the risk-adjusted performance assessment, when returns are not symmetrically distributed, by including higher moments of the distribution. It seems, therefore, that the extent of profitable investment decisions based on short-term reversal appears more pronounced during phases of market instability and that each of these indicators is able to promptly capture such temporarily deviations. Our findings thus indicate that long-short strategies perform particularly well during phases of market distress, as already detected in other similar empirical studies (see, e.g., Do and Faff 2010, Bogomolov 2013, Huck and Afawubo 2015 and Krauss et al. 2017).

Secondly, we observe remarkable differences in the stream of annual performances across strategies both prior and post the crisis; both $\widetilde{\Delta p}$ and $\widetilde{\Delta r}$ have three time observations with negative performances (namely, $\widetilde{\Delta p}$: 2007, 2013 and 2017; $\widetilde{\Delta r}$: 2003, 2004 and 2019-Q1-2), while for $\mathrm{P}(\Delta r \nearrow)$ is only one (2017). Excluding negative performances, the worst annual result for $\widetilde{\Delta p}$ corresponds to about 1.3% (2011 and 2014), for $\widetilde{\Delta r}$ is 0.95% (2017), while it is much higher for $\mathrm{P}(\Delta r \nearrow)$ (4.7% in 2005). By contrast, the latter has a maximum corresponding to about 85% (2008), while the other two strategies obtain performances of more than 120% in the same period. Strategy based on $\mathrm{P}(\Delta r \nearrow)$ seems

to provide, therefore, a more stable stream of positive annual performances, although in a narrower range. This is also confirmed in terms of the risk-adjusted performances measured by SR, its variants Mod_SR and ES, and the Omega ratio.

Thirdly, the volatility of returns is very similar across the strategies, ranging in general around $8-12\%$ while reaching more than 30% in the years 2008-2009. A significant exception is year 2016, when the volatility of the corresponding strategies increased substantially with respect to the neighborhood period (in particular for Δr and $\mathrm{P}(\Delta r \nearrow)$). More in general, these volatility patterns suggest that the three investment criteria do not seem to systematically differ between each other in the selection of more or less volatile stocks during this sample period.

Lastly, both DD and $\mathrm{CVaR}_{99\%}$ seem to follow the dynamics of the financial cycle, with a sharp rise (in absolute value) in the years of the crisis and lower values during tranquil phases. DD indicates the maximum loss accrued in each year with respect to the peak value recorded in that period, thus representing the maximum loss suffered due to the trading activity. Over the entire sample period, strategies based on $\widetilde{\Delta p}$ and $\widetilde{\Delta r}$ are respectively the less and most risky in such framework, which is also confirmed by the annual mean values of the drawdowns (see PI). Instead, $\mathrm{CVaR}_{99\%}$ quantifies the amount of tail risk a portfolio has by computing the average of the extreme losses beyond a cutoff equal to 99% of the distribution of returns. $\mathrm{CVaR}_{99\%}$ is thus the expected loss due to the returns that cross the threshold of 99%. From Table 4, $\mathrm{CVaR}_{99\%}$ is typically around $2-3\%$ of the investment, with significant peaks in the period 2008-2009. Notwithstanding maximum losses reported by the DD in the range around $7-13\%$ and conditional values at risk of about $2-3\%$, still every strategy is able to get consistent positive cumulative returns in most of the annual observations.

3.3 Detecting $\mathrm{P}(\Delta r \nearrow)$ contribution to portfolio performance

Table 5 reports the empirical results obtained by equally investing in decile portfolios formed by sorting stocks according to either $\widetilde{\Delta p}$ (in Panel A), $\widetilde{\Delta r}$ (in Panel B) or $\mathrm{P}(\Delta r \nearrow)$ (in Panel C). Raw annualized returns discussed in subsection 3.1 clearly show monotonic patterns for these sorting criteria. Portfolios formed by stocks with lower values of $\widetilde{\Delta p}$ or $\widetilde{\Delta r}$ and those with higher values of $\mathrm{P}(\Delta r \nearrow)$ are therefore typically associated with better performances in the near future, while the opposite occurs for those portfolios in the bottom side of these rankings. Interestingly, as exhibited in Table 5, these patterns are confirmed even when portfolios are evaluated in terms of the *alphas* from factor models. We consider two specifications: the 5-factor model proposed by Fama and French (2015) and a 7-factor model which includes the momentum (Carhart, 1997) and the short-term reversal (Jegadeesh and Titman, 1993) factors. Both measures of alphas are very coherent and support the interpretation that top decile portfolios are not only able to get on average better raw returns, but also that extra-market performances do not vanish when controlling for dynamic factor exposures. Indeed, both $\widetilde{\Delta p}$ and $\widetilde{\Delta r}$ seem to contain information regarding future returns. The persistence of these extra performances emerges even more clearly when we consider the *Top-Bottom* strategy investing long in the top performer decile and short in the worst decile portfolio. These findings are in line with those reported in Panel C, where the sorting criterion is based on $\mathrm{P}(\Delta r \nearrow)$. Even in this case, the *Top-Bottom* strategy is able to generate a profitable result, comparable with the $\widetilde{\Delta r}$ and higher than $\widetilde{\Delta p}$ (e.g., the 7-factor alphas are about 17.55% vs. 18.02% and 13.69% respectively). The performances of the *Top-Bottom* strategies based on $\mathrm{P}(\Delta r \nearrow)$ are significant not only economically but also statistically, with t-statistics equal to about 4.9 and 4.4 in the 5-factor and 7-factor models, respectively.

Table 5: One-Way Sort with respect to $\widetilde{\Delta p}$, $\widetilde{\Delta r}$ and P($\Delta r \nearrow$). Performances (in percentage) of decile portfolios composed by equally investing in stocks sorted according to $\widetilde{\Delta p}$ in Panel A, $\widetilde{\Delta r}$ in Panel B and P($\Delta r \nearrow$) in Panel C. Each panel reports in the last column the performances of the corresponding *Top-Bottom* strategy. We report as measure of performance the *alpha* generated by both the 5-factor model (Fama and French 2015) and the 7-factor model which includes the momentum (Carhart 1997) and the short-term reversal (Jegadeesh and Titman 1993) factors into the 5-factors model. The holding period is $h = 1$ day. Newey-West t-statistics are reported in parenthesis. Data are annualized and refer to the period from January 2003 to June 2019.

Panel A	deciles of $\widetilde{\Delta p}$										
	1	2	3	4	5	6	7	8	9	10	*Top-Bottom*
annualized 5-factor α	12.49	6.88	5.34	2.52	5.00	1.18	3.54	2.58	0.30	-2.05	14.84
	(3.49)	(2.73)	(2.39)	(1.43)	(3.20)	(0.69)	(1.93)	(1.61)	(0.17)	(-1.15)	(3.39)
annualized 7-factor α	12.16	6.74	4.96	2.33	4.53	1.08	3.85	2.94	0.82	-1.35	13.69
	(3.93)	(2.96)	(2.58)	(1.39)	(2.93)	(0.60)	(2.11)	(1.83)	(0.49)	(-0.78)	(3.82)
Panel B	deciles of $\widetilde{\Delta r}$										
	1	2	3	4	5	6	7	8	9	10	*Top-Bottom*
annualized 5-factor α	15.95	8.92	7.68	5.73	4.58	1.44	-0.60	-1.52	-0.29	-2.90	19.41
	(5.46)	(3.86)	(3.98)	(3.31)	(2.63)	(0.84)	(-0.35)	(-0.86)	(-0.15)	(-1.32)	(4.61)
annualized 7-factor α	15.32	8.67	7.59	5.88	4.48	1.42	-0.27	-1.51	-0.11	-2.29	18.02
	(5.37)	(3.79)	(4.88)	(3.38)	(2.56)	(0.85)	(-0.16)	(-0.85)	(-0.05)	(-0.94)	(4.26)
Panel C	deciles of P($\Delta r \nearrow$)										
	1	2	3	4	5	6	7	8	9	10	*Top-Bottom*
annualized 5-factor α	-4.21	-1.39	-0.54	-0.20	1.48	6.49	3.64	11.70	8.19	14.25	19.27
	(-1.90)	(-0.69)	(-0.28)	(-0.12)	(0.99)	(3.39)	(1.85)	(5.59)	(3.86)	(6.00)	(4.91)
annualized 7-factor α	-3.31	-0.97	-0.68	0.17	1.39	6.33	3.50	11.71	7.76	13.67	17.55
	(-1.39)	(-0.49)	(-0.35)	(0.11)	(0.91)	(3.49)	(1.81)	(5.06)	(3.78)	(5.45)	(4.45)

Since each of the three indicators is able to generate extra-profits not captured by standard market drivers of risk, it seems interesting to ask whether they contain portions of mutually exclusive information useful to forecast stock returns. In line with a wide extant literature in finance (see, e.g., Fama and French 1996, Cohen et al. 2005, Whited and Wu 2006, Adrian and Franzoni 2009, Flori et al. 2019, among others), we employ a joint sorting of the stocks in our sample to investigate the contribution of these indicators to the creation of extra performances. In particular, Table 6 shows the alphas from the 7-factor model related to double sorting procedures in which stocks are sorted in quintiles based on one indicator and then, conditionally on this sort, they are further sorted in quintiles according to a second indicator, thus generating 5×5 portfolios as well as five *Top-Bottom* portfolios in each double sort specification. In particular, in Panel A of Table 6 we use as the first sorting dimension either $\widetilde{\Delta p}$ (on the left side of the panel) or $\widetilde{\Delta r}$ (on the right side of the panel), while the second sorting dimension is P($\Delta r \nearrow$). Conversely, in Panel B the first sorting criterion is always P($\Delta r \nearrow$), while the second sorting dimension is either $\widetilde{\Delta p}$ (on the left of the panel) or $\widetilde{\Delta r}$ (on the right of the panel).

Conditional sorts reported in Panel A indicate that, controlling for $\widetilde{\Delta p}$, the annualized alpha performances of the *Top-Bottom* strategies ranked by P($\Delta r \nearrow$) range between about 7.7% and 22%. Moreover, the average annualized alpha, which is constructed by equally investing in the five *Top-Bottom* portfolios and that represents our cleanest measure of whether the second dimension of the double sorting provides information beyond the first sorting criterion (see *Avg*), is about 16%. These performances appear significant not only economically but also statistically, with the t-statistics of the average performance of the *Top-Bottom* portfolios equal to 6. Findings from this double sort procedure also suggest that stocks with a lower probability to increase their Δr in the near future tend to perform worse than those belonging to the highest quintiles of the P($\Delta r \nearrow$) distribution (see the average values by row), thus confirming the monotonic pattern already observed in Table 5.

Furthermore, the double sort procedure ranked by $\widetilde{\Delta r}$ and then by $\mathrm{P}(\Delta r \nearrow)$ indicates that the *Top-Bottom* strategies provide valuable extra-performances, although lower than those reported for the case which involves $\widetilde{\Delta p}$. Nevertheless, the average annualized alpha of the *Top-Bottom* portfolios is economically consistent and equal to 3.79% (t-statistics of 1.87). More generally, both double sorting procedures indicate that $\mathrm{P}(\Delta r \nearrow)$ provide additional information to both $\widetilde{\Delta p}$ and $\widetilde{\Delta r}$ that can be exploited to construct profitable investment strategies.

Table 6: Double Sorts by either $\widetilde{\Delta p}$ or $\widetilde{\Delta r}$, and P($\Delta r \nearrow$). Performances (in percentage) of 25 quintiles portfolios, as well as the performance of the *Top-Bottom* strategies in the last row of each block, obtained by firstly sorting stocks in quintiles according to the indicator reported by column and then, conditionally on this sort, further sorting in quintiles constructed in terms of the indicator reported by row. Panel A considers sorting stocks firstly by $\widetilde{\Delta p}$ (on the left) or $\widetilde{\Delta r}$ (on the right) and then, within these quintiles, by further sorting based on $\mathrm{P}(\Delta r \nearrow)$, while Panel B considers sorting stocks by $\mathrm{P}(\Delta r \nearrow)$ and then by $\widetilde{\Delta p}$ (on the left of the panel) or $\widetilde{\Delta r}$ (on the right of the panel). For each panel and double-sort specification, we report as measure of performance the alpha generated by a 7-factor model which includes the momentum (Carhart 1997) and the short-term reversal (Jegadeesh and Titman 1993) factors into the 5-factor model (Fama and French 2015). The holding period is $h = 1$. Newey-West t-statistics are reported in parenthesis. Data are annualized and refer to the period from January 2003 to June 2019.

Panel A: sorting stocks by $\widetilde{\Delta p}$ or $\widetilde{\Delta r}$ and then by $\mathrm{P}(\Delta r \nearrow)$

quintiles of $\mathrm{P}(\Delta r \nearrow)$	quintiles of $\widetilde{\Delta p}$						quintiles of $\widetilde{\Delta r}$						quintiles of $\mathrm{P}(\Delta r \nearrow)$
	1	2	3	4	5	Avg	1	2	3	4	5	Avg	
	annualized 7-factor α						annualized 7-factor α						
1	6.49	-5.41	-3.01	-4.38	-8.60	-3.11	7.71	7.16	2.88	-4.63	-0.04	2.51	1
	(1.92)	(-1.88)	(-1.03)	(-1.59)	(-3.07)	(-1.87)	(1.86)	(2.47)	(1.19)	(-1.55)	(-0.01)	(1.42)	
2	2.58	5.00	-3.85	-1.76	-4.90	-0.66	13.52	3.94	-1.67	-1.03	-5.66	1.62	2
	(0.82)	(1.71)	(-1.41)	(-0.69)	(-1.93)	(-0.55)	(4.23)	(1.75)	(-0.70)	(-0.41)	(-1.79)	(1.36)	
3	12.75	5.16	4.39	3.99	-3.02	4.54	9.37	6.83	2.42	0.46	-4.44	2.81	3
	(3.67)	(1.84)	(1.68)	(1.57)	(-1.25)	(3.20)	(2.90)	(2.69)	(0.99)	(0.20)	(-1.46)	(2.34)	
4	11.48	4.51	4.53	6.01	4.42	6.16	16.10	6.57	6.03	-1.39	1.25	5.55	4
	(2.64)	(1.56)	(1.65)	(2.54)	(1.57)	(4.07)	(5.05)	(2.39)	(1.88)	(-0.59)	(0.41)	(4.57)	
5	14.65	9.83	12.16	13.93	11.53	12.41	13.56	8.81	5.22	1.91	2.91	6.40	5
	(4.36)	(2.91)	(4.15)	(4.87)	(3.96)	(7.10)	(3.61)	(3.06)	(2.06)	(0.72)	(1.00)	(4.84)	
Top-Bottom	7.67	16.11	15.63	19.14	22.02	16.01	5.43	1.54	2.27	6.86	2.95	3.79	*Top-Bottom*
	(1.59)	(3.11)	(3.57)	(4.39)	(4.82)	(6.00)	(1.02)	(0.41)	(0.64)	(1.62)	(0.57)	(1.87)	

Panel B: sorting stocks by $\mathrm{P}(\Delta r \nearrow)$ and then by $\widetilde{\Delta p}$ or $\widetilde{\Delta r}$

quintiles of $\widetilde{\Delta p}$	quintiles of $\mathrm{P}(\Delta r \nearrow)$						quintiles of $\mathrm{P}(\Delta r \nearrow)$						quintiles of $\widetilde{\Delta r}$
	1	2	3	4	5	Avg	1	2	3	4	5	Avg	
	annualized 7-factor α						annualized 7-factor α						
1	6.47	0.10	10.27	9.24	12.61	7.65	-1.98	-0.87	6.46	14.44	14.09	6.20	1
	(1.77)	(0.03)	(3.31)	(2.66)	(3.56)	(3.88)	(-0.64)	(-0.31)	(2.28)	(3.22)	(3.50)	(3.83)	
2	-2.14	0.90	5.29	7.42	11.39	7.02	1.31	0.60	5.38	7.97	11.98	4.09	2
	(-0.65)	(0.29)	(1.91)	(2.64)	(2.98)	(2.92)	(0.48)	(0.22)	(2.17)	(2.72)	(3.80)	(2.00)	
3	-2.79	1.18	2.90	9.77	9.07	4.47	-4.54	-0.92	3.91	8.29	16.91	5.36	3
	(-0.86)	(0.43)	(1.14)	(3.46)	(2.89)	(3.12)	(-1.60)	(-0.34)	(1.50)	(3.38)	(5.38)	(3.84)	
4	-2.90	-0.70	1.91	5.65	8.15	3.92	-5.81	-1.45	0.02	4.01	9.24	4.47	4
	(-1.02)	(-0.31)	(0.78)	(2.11)	(2.80)	(2.85)	(-2.02)	(-0.62)	(0.01)	(1.60)	(3.30)	(3.42)	
5	-10.85	-2.97	-0.84	5.64	13.64	2.34	0.32	1.67	3.25	2.57	2.36	1.07	5
	(-4.12)	(-1.24)	(-0.31)	(2.09)	(4.61)	(1.74)	(0.07)	(0.60)	(1.48)	(1.02)	(0.87)	(0.88)	
Top-Bottom	19.42	3.16	11.20	3.41	-0.91	7.02	-2.30	-2.49	3.11	11.57	11.46	4.09	*Top-Bottom*
	(3.89)	(0.74)	(2.56)	(0.89)	(-0.23)	(2.92)	(-0.52)	(-0.70)	(0.86)	(2.32)	(2.43)	(2.00)	

We next examine how much information is provided by both $\widetilde{\Delta p}$ and $\widetilde{\Delta r}$ that is not already contained in $\mathrm{P}(\Delta r \nearrow)$. Hence, we sort stocks in reverse, firstly by $\mathrm{P}(\Delta r \nearrow)$ and then, separately, by $\widetilde{\Delta p}$ and $\widetilde{\Delta r}$ (see Panel B of Table 6). Notice how the annualized average alphas of the *Top-Bottom* portfolios are about 7% and 4% between the two specifications (t-statistics are around 3 and 2, respectively), thus significant both economically and statistically. However, when sorting by $\widetilde{\Delta p}$ within $\mathrm{P}(\Delta r \nearrow)$ quintiles, we observe that extra-performances decrease substantially with respect to the opposite double sorting procedure, while when double sorting by $\widetilde{\Delta r}$ it seems that incremental information about future performances over (and above) $\mathrm{P}(\Delta r \nearrow)$ is concentrated especially in the

highest quintiles of the latter indicator and that high heterogeneity is present in these *Top-Bottom* strategies. In general, we confirm that for both double sorting procedures, low quintiles of $\widetilde{\Delta p}$ and $\widetilde{\Delta r}$ are also in this setting typically associated with better average results (see column *Avg*).

Interestingly, from these double sorts procedures we can argue that the information provided by $\mathrm{P}(\Delta r \nearrow)$ is likely to enrich the one embedded in both $\widetilde{\Delta p}$ and $\widetilde{\Delta r}$. However, as expected, since $\mathrm{P}(\Delta r \nearrow)$ refers to a different layer of analysis than $\widetilde{\Delta p}$, it seems to better complement the information embedded in price gaps than the one present in $\widetilde{\Delta r}$. Specifically, the three indicators seem to provide complementary information that can be utilized to build profitable investment strategies. More importantly, the inclusion of $\mathrm{P}(\Delta r \nearrow)$ to build portfolios can thus contribute to improve the performances based on the other two criteria. For instance, as regards the contribution of $\mathrm{P}(\Delta r \nearrow)$ to $\widetilde{\Delta p}$, the strategy which buys the portfolio (1-5) and sells (5-1) in Panel A of Table 6 can produce an extra-performance of about 23%, much higher than the corresponding *Top-Bottom* one-way performances of Table 5, while the strategy that goes long in (5-1) and short in (1-5) in the reverse sorting of Panel B of Table 6 generates an extra-performance of about 23.5%, again much higher than the ones provided by the one-way *Top-Bottom* strategies. In a similar way, profitable illustrative portfolios can be constructed based on the combined information provided by $\mathrm{P}(\Delta r \nearrow)$ and $\widetilde{\Delta r}$. Notice in particular how better results than those reported in the analog one-way strategies of Table 5 can be obtained by strategies selecting those portfolios of Table 6 with performances statistically different from zero, for which therefore a much clearer contribution emerge from the double sorting procedure.

Overall, these results support the use of these sources of information to provide investment signals which more effectively select those stocks to be bought or sold depending not only on current gaps compared to their peers (as approximated by observing $\widetilde{\Delta p}$ or $\widetilde{\Delta r}$), but also on predicted deviations of the corresponding market performances (as estimated here by $\mathrm{P}(\Delta r \nearrow)$).

3.4 Extracting exposures to factor sources of risk

Findings of subsection 3.3 point to the presence of profitable investment opportunities based on integrated information provided by $\widetilde{\Delta p}$, $\widetilde{\Delta r}$ and $\mathrm{P}(\Delta r \nearrow)$. In this subsection, we address this aspect by focusing specifically on those pairs of portfolios that better provide indication for deviating or, alternatively, converging market trajectories. This information is relevant for constructing profitable investment strategies. In fact, at each trading day, practitioners may observe, for instance, a gap in price levels of stock i with respect to its peers and may wonder whether this divergence is likely to increase or decrease in the next days. From an operational perspective, this information is thus crucial for distinguishing whether to buy or sell stock i. If the trader believes that this gap is still widening, then he/she could be induced to buy stock i, while if he/she estimates that this gap has already reached its maximum and will therefore reverse in the near future, then he/she could opt to sell stock i. Basically, the trader may want to find a way to distinguish, therefore, between momentum and short-term reversal effects once a price gap is observed for stock i, on the basis of which an opposite signal of buy vs. sell should be provided.

Against this background, the double sort framework provides helpful indications for disentangling such scenarios which we relate to the provision of investment signals based on the joint contribution of $\widetilde{\Delta p}$ and $\mathrm{P}(\Delta r \nearrow)$. Once again, we consider as illustrative portfolios those placed in the extremes of the diagonals of the blocks reported in Panel A of Table 6 (left table). Hence, portfolio (5-1) is interpreted as composed by stocks with large price gaps with respect to their peers and for which the LSTM approach detects a strong contraction of the corresponding delta returns in the next trading

day. For these stocks, the suggested investment signal is to sell, since a reversal is highly expected. Conversely, traders should buy stocks belonging to portfolio (1-5). These stocks are, in fact, market under-valued with respect to their peers, thus having very low price gaps, but are expected to show a consistent appreciation in terms of returns compared to peers. Since these stocks are considered as those more likely to get high market performances in the next trading day, then they should enter in the strategy with a long position. Overall, the zero-cost investment strategy that goes long on portfolio (1-5) and short on portfolio (5-1) should contribute to emphasize the impact of the short-term reversal effect, since in so doing we sell those stocks that, besides having higher (absolute value of) gaps in price levels, they are also highly expected to show a substantial reversal of their market behaviour in the near future, while buying those stocks with opposite patterns.

In addition, we also consider the opposite scenario which, instead, favors momentum market dynamics. This is the case of portfolios placed in the extremes of the main diagonal of Panel A in Table 6 (left table). Portfolio (1-1) represents stocks with poor market performances compared to peers and for which LSTM expectations confirm a depressing market trend for the next trading day. For these stocks, the investment signal is therefore to be short. Conversely, traders could opt to buy stocks in portfolio (5-5) since not only these stocks are performing better than their peers, but they are expected to maintain such market dynamics in the near future. More generally, this indication refers to a market momentum dynamics that we aim to capture by combining the information provided by $\widetilde{\Delta p}$ and $\mathrm{P}(\Delta r \nearrow)$.

To investigate the market exposures of these competitive investment perspectives, Table 7 Panel A shows the application of the 7-factor model to the aforementioned zero-cost investment strategies for different investment horizons: 1 day ($h = 1$), 5 days ($h = 5$), and 10 days ($h = 10$). Notice how these strategies are in general neutral to factor exposures with the exception of momentum (Mom) and short-term reversal (ST_Rev) effects, for which instead consistent and statistically significant effects are observed. In particular, for the strategy involving portfolios (5-5) - (1-1) (referring to portfolios reported in the left table of Panel A of Table 6) we observe a relevant role for the Mom factor, while for the strategy investing in portfolios (1-5) - (5-1) the dominant effect refers to ST_Rev, at least for shorter investment horizons. These findings are thus coherent with our purpose to identify investment strategies specifically devoted to extract such market dynamics by combining the joint information from $\widetilde{\Delta p}$ and $\mathrm{P}(\Delta r \nearrow)$. In particular, by referring to stocks in the extreme portfolios of the main diagonal or the anti-diagonal we are basically isolating stronger signals with respect to market phases related to momentum or short-term reversal, thus exploiting current and past information to better identify those stocks that are more likely to persist in their market behaviour or those that instead are highly expected to revert. Importantly, these effects are largely confirmed even in parsimonious models restricted to only the momentum and short-term reversal factors (see Table 7 Panels B, C, D).

Finally, we observe that momentum and short-term reversal effects may coexist, which is a result in line with the literature that already detected the interplay between momentum and reversal in various trading settings (see, e.g., Bloomfield et al. 2009, Da et al. 2013, Cremers and Pareek 2014, Zhu and Yung 2016, among others). Nevertheless, from Table 7 it is clear that each strategy is capable to extract the targeted effect and that the corresponding magnitude overwhelms the other factor exposures. A relevant exception seems to be the coefficient of ST_Rev, strongly decreasing for longer horizons as $h = 10$, thus suggesting a more decisive impact in the short-term for such factor. However, different investment horizons seem to have only a marginal role, being the relevant exposures substantially confirmed across model specifications.

Table 7: Exposures to Factor Sources of Risk. Results of the factor regression models for the investment strategies referring to portfolios (1-5) - (5-1) and (5-5) - (1-1) in the anti-diagonal and main diagonal of the left table of Panel A of Table 6, respectively. The holding periods are $h = 1, 5, 10$ days. Newey-West t-statistics are reported in parenthesis. Panel A refers to the 7-factor model, Panel B displays coefficients for a 2-factor model with respect to momentum and short-term reversal, while Panels C and D show result for single factor models with respect to momentum and short-term reversal factor, respectively. Regression coefficients are in percentage and refer to the period from January 2003 to June 2019. *p-value < 0.1; **p-value < 0.05; ***p-value < 0.01.

	anti-diagonal: (1-5) - (5-1)			main diagonal: (5-5) - (1-1)		
holding days	$h = 1$	$h = 5$	$h = 10$	$h = 1$	$h = 5$	$h = 10$
			Panel A			
Intercept	0.09***	0.06***	0.05***	0.02	-0.00	−0.02*
	(5.00)	(4.32)	(3.93)	(1.09)	(-0.26)	(-1.86)
Mkt - R_f	4.02	0.65	0.39	-0.93	-0.48	-0.74
	(1.31)	(0.23)	(0.17)	(-0.34)	(-0.22)	(-0.33)
SMB	5.10	6.11	5.29*	-1.77	-6.20**	-4.15
	(1.08)	(1.38)	(1.68)	(-0.38)	(-2.11)	(-1.54)
HML	10.03	10.92	6.19	-7.34	-6.74	-3.12
	(1.37)	(1.36)	(1.00)	(-0.58)	(-0.73)	(-0.30)
RMW	-6.05	-6.89	-9.11*	-11.71	-1.63	-0.31
	(-0.70)	(-1.21)	(-1.79)	(-1.39)	(-0.25)	(-0.05)
CMA	-1.26	3.63	4.73	-13.53	-5.26	-10.37
	(-0.12)	(0.43)	(0.62)	(-1.06)	(-0.50)	(-1.02)
Mom	-28.49***	-37.83***	-41.04***	50.22***	48.86***	46.83***
	(-6.37)	(-8.49)	(-10.40)	(10.61)	(12.10)	(13.24)
ST_Rev	40.96***	39.78***	30.62***	-16.69***	-10.76**	-4.45
	(5.91)	(7.82)	(5.29)	(-2.57)	(-2.10)	(-0.83)
adjusted-R^2	0.16	0.36	0.36	0.16	0.30	0.31
			Panel B			
Intercept	0.09***	0.06***	0.05***	0.02	-0.00	−0.02*
	(4.64)	(4.22)	(3.68)	(0.97)	(-0.34)	(-1.92)
Mom	−34.00***	−42.60***	−44.26***	52.06***	51.38***	48.24***
	(-4.67)	(-5.94)	(-7.16)	(8.15)	(7.95)	(7.80)
ST_Rev	43.18***	40.38***	31.16***	−15.69***	−10.68**	-4.26
	(5.40)	(7.58)	(5.73)	(-2.65)	(-2.24)	(-0.88)
adjusted-R^2	0.16	0.35	0.35	0.16	0.30	0.31
			Panel C			
Intercept	0.10***	0.07***	0.06***	0.01	-0.01	−0.02**
	(5.41)	(4.84)	(4.51)	(0.72)	(-0.64)	(-2.17)
Mom	−38.55***	−46.85***	−47.54***	53.71***	52.50***	48.68***
	(-6.60)	(-6.39)	(-9.39)	(7.56)	(8.23)	(8.20)
adjusted-R^2	0.08	0.23	0.27	0.15	0.29	0.31
			Panel D			
Intercept	0.09***	0.06***	0.04***	0.02	0.00	-0.02
	(4.00)	(3.15)	(2.52)	(1.08)	(0.05)	(-1.05)
ST_Rev	47.87***	46.25***	37.27***	−22.87**	−17.77*	-10.91
	(4.86)	(5.09)	(3.80)	(-2.09)	(-1.65)	(-1.08)
adjusted-R^2	0.10	0.17	0.13	0.02	0.02	0.02

4 Robustness analysis

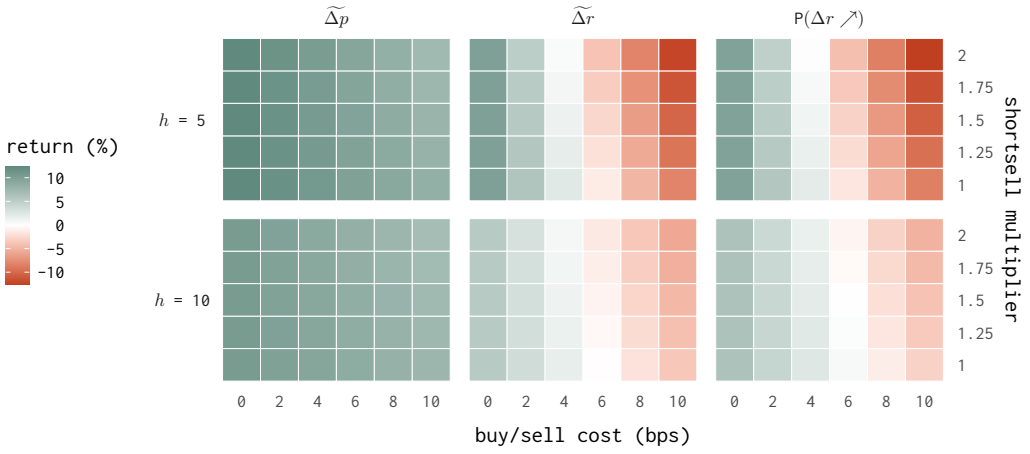
4.1 Transaction costs

Literature has investigated the profitability of reversal strategies net of transaction costs. For instance, Jegadeesh (1990) documents profits of about 2% per month over the interval 1934-1987 for a reversal strategy, buying and selling stocks on the basis of prior-month returns and with a monthly holding period, which are not explainable by direct trading costs. By contrast, several empirical studies have shown that profits from short-term reversal strategies are indistinguishable from zero or even negative once transaction costs, such as commissions, short-selling fees, bid-ask spreads, and price impact costs, are taken into account. Therefore, whether significant profits net of transaction costs are consistent with the notion of reversal effects appears still questioned (see, e.g., Conrad et al. 1997, Avramov et al. 2006, De Groot et al. 2012, Do and Faff 2012, Blitz et al. 2013 to name a few).

The dynamics of net performances relates to the turnover of the assets held in the portfolio, which can be consistent and deteriorate portfolio net results. Hence, in order to assess the impact of portfolio rebalancing, we compute daily turnover at t for each stock as the daily weight difference between the beginning of period weight at t and the end of period weight at $t-1$, namely as the weight difference due to an actual rebalancing, net with respect to market movements; the portfolio turnover is then computed as the sum of stock turnover and it goes from 0 (no rebalancing) to 200% (complete sell of stocks and buy of a new set of stocks). We find that, over the entire sample period, portfolio turnover of $\mathrm{P}(\Delta r \nearrow)$ for $h = 1$ is consistent, reaching a value on average of 175% on a daily basis. This, in turn, translates into negative net performances by assuming a typical level of trading costs equal to 5 basis points (see, e.g., Avellaneda and Lee 2010, Fischer and Krauss 2018). Similar levels of turnover are reached also by signals provided by $\widetilde{\Delta r}$. By contrast, the average turnover of $\widetilde{\Delta p}$ is only 17% on a daily basis, thus portfolios based on this ranking criterion are much more able to contain the impact of trading costs on their net performances.

In fact, like any long-short strategy, pairs-trading involves trading the same stocks twice, i.e., ideally when the initial divergence is detected and when a subsequent convergence is identified. This means that two roundtrips of costs should be considered. To analyze the impact of trading costs on the net performances of the *Top-Bottom* strategies, we thus provide in Figure 6 a scenario analysis employing a grid of values ranging from zero to 10 basis point per half-turn operation. Hence, we avoid to rely on a single arbitrary trading cost level, but we opt for a sufficiently large range of values that reasonably represent breakeven transaction costs for several statistical arbitrage mean-reverting strategies encompassing stocks belonging to the S&P 500 (Focardi et al., 2016). In particular, we refer to such costs as those required to both buy and sell the *Top* side of the strategy or for closing the *Bottom* position. Moreover, we take into account possible higher costs for the short selling initialization of the *Bottom* position by assuming a multiplier factor reaching value 2, with steps equal to 0.25. We assume, therefore, that short-selling can reach a level of costs double to that of buy/sell operations. Finally, Figure 6 takes into account the role of portfolio turnover by considering investment horizons $h = 5, 10$ days, for which therefore a lower portfolio rebalancing is required.

Figure 6: *Top-Bottom* Net Performance for Different Trading Costs. Returns (in percentage) of *Top-Bottom* decile strategies defined using as sorting criteria $\widetilde{\Delta p}$, $\widetilde{\Delta r}$ and $\mathrm{P}(\Delta r \nearrow)$. The holding periods are $h = 5, 10$ days. Trading costs are $0, 2, 4, 6, 8, 10$ basis points. Short-sell costs are taken to be $1, 1.25, 1.5, 1.75, 2$ times the buy/sell cost. Returns are annualized and refer to the period from January 2003 to June 2019.



Note how by enlarging the investment horizon to $h = 5, 10$ days, the *Top-Bottom* strategies based on $\mathrm{P}(\Delta r \nearrow)$ (similarly to $\widetilde{\Delta r}$) can generate positive net performances if low values of trading costs are

assumed, while regardless the investment horizon we confirm that the *Top-Bottom* strategies based on $\widetilde{\Delta p}$ are much less affected by costs thanks to low levels of portfolio rebalancing. Furthermore, our findings for $h = 5, 10$ indicate that the turning point between positive and negative net performances of *Top-Bottom* strategies based on $\mathrm{P}(\Delta r \nearrow)$ occur for trading costs around 4-6 basis points, hence for values in line with the half-turn costs applied in several studies in pairs-trading research (see, e.g., Liu et al. 2017, Clegg and Krauss 2018, Stübinger et al. 2018, Stübinger and Endres 2018). From an operational point of view, it seems that for the *Top-Bottom* strategies based on $\mathrm{P}(\Delta r \nearrow)$, an investment horizon of about 5 days is therefore sufficient to limit portfolio rebalances and determine positive net performances once reasonable trading costs are considered, while a longer investment horizon confirms similar patterns but at lower yields (in absolute terms). More in general, it is worth noticing that we present results that extend the original investment horizon $h = 1$ to slightly longer periods $h = 5, 10$ as illustrative cases to show how by enlarging the investment horizon it is possible to balance the accuracy of the model with the impact of operational costs.

4.2 Investment horizons

When working with holding period h longer than 1 day, following Jegadeesh and Titman (1993) and Fischer and Krauss (2018) we perform the following procedure in order to avoid specific weekday effects: we compute h overlapping portfolios with holding period h, each one starting in subsequent days in the first $1, \ldots, h$ investment days; we then compute daily returns for each of the h portfolios and consider their daily average as the representative h-horizon portfolio.

For the three benchmark investing criteria $\widetilde{\Delta p}$, $\widetilde{\Delta r}$ and $\mathrm{P}(\Delta r \nearrow)$, Table 8 reports statistically and economically significant *Top-Bottom* strategies even for investment horizons equal to 5 or 10 days (namely, $h = 5, 10$). In addition, the table supports previous findings discussed in subsection 3.1 about the presence of almost monotonic patterns in the raw performances of decile portfolios.

Table 8: One-Way Sorts for Different Investment Horizons. Raw returns (in percentage) obtained by equally investing in stocks composing decile portfolios defined using as sorting criteria $\widetilde{\Delta p}$, $\widetilde{\Delta r}$ and $\mathrm{P}(\Delta r \nearrow)$. The holding periods are $h = 5, 10$ days. Newey-West t-statistics are reported in parenthesis. Data are annualized and refer to the period from January 2003 to June 2019.

	1	2	3	4	5	6	7	8	9	10	*Top-Bottom*
$\widetilde{\Delta p}$; $h = 5$	20.84	15.62	12.54	11.61	14.57	12.16	13.50	12.51	9.57	7.77	12.31
	(3.72)	(3.17)	(3.04)	(2.90)	(3.57)	(3.27)	(3.58)	(3.41)	(2.76)	(2.24)	(3.30)
$\widetilde{\Delta p}$; $h = 10$	19.46	15.34	11.75	12.65	14.29	11.71	13.60	13.19	9.07	8.40	10.39
	(3.57)	(3.16)	(2.86)	(3.15)	(3.54)	(3.14)	(3.67)	(3.62)	(2.59)	(2.40)	(2.93)
$\widetilde{\Delta r}$; $h = 5$	19.75	16.60	14.24	14.09	12.95	13.90	11.46	10.70	10.16	8.99	9.82
	(4.52)	(3.87)	(3.45)	(3.58)	(3.31)	(3.44)	(3.04)	(2.92)	(2.76)	(2.44)	(5.44)
$\widetilde{\Delta r}$; $h = 10$	16.69	14.79	13.65	13.65	13.40	13.42	12.12	11.82	11.05	10.83	5.36
	(3.98)	(3.62)	(3.41)	(3.50)	(3.43)	(3.41)	(3.18)	(3.14)	(2.87)	(2.81)	(4.57)
$\mathrm{P}(\Delta r \nearrow)$; $h = 5$	8.95	10.16	11.23	11.36	12.97	13.82	13.57	13.74	17.90	19.44	9.59
	(2.38)	(2.69)	(3.06)	(2.97)	(3.42)	(3.51)	(3.48)	(3.41)	(4.31)	(4.18)	(5.22)
$\mathrm{P}(\Delta r \nearrow)$; $h = 10$	10.20	11.40	12.14	12.11	12.88	13.39	13.44	13.43	15.49	17.05	6.24
	(2.65)	(2.96)	(3.21)	(3.18)	(3.31)	(3.43)	(3.41)	(3.35)	(3.83)	(4.05)	(4.87)

Subsection 4.1 indicates how widening the investment horizon limits the portfolio turnover costs and enables positive net returns. Hence, there is a trade-off between the financial viability of these investment strategies and the capacity of generating profitable investment opportunities. Such variants for longer values of h thus make explicit the need for balance between the accuracy of the model, which deteriorates once we move towards longer investment horizons, and the associated transaction costs due to portfolio turnover, which instead are substantial in shorter investment horizons. Our findings indicate that the proposed approach is not heavily affected by the choice of the investment horizons,

with longer h confirming similar patterns but at lower yields, while slightly enlarging h appears sufficient to limit portfolio rebalancing and facilitate the formation of positive net performances.

4.3 Different weighting scheme

As an alternative to the decile portfolio scheme described in subsection 3.1, we build portfolios with a weighting linear in the pairs-trading indicator of interest. Namely, calling x the sorting variable, we set the weight:

$$w_t^i \propto x_t^i - \bar{x}_t, \tag{11}$$

where $\bar{x}_t$ stands for the cross-median of that indicator on day t. Effectively, we build a long/short portfolio analogous in spirit to the decile *Top-Bottom*, but investing in all stocks available that day, and with weights proportional to the variable of interest. As in subsection 3.1, we stick to a daily investment horizon and we do not consider transaction costs.

Taking the case $x = P(\Delta r \nearrow)$, the resulting portfolio has an annualized Sharpe ratio of 146.34% and an annualized return of 12.58%, which can be compared to 122.74% and 18.32% of the decile *Top-Bottom* portfolio discussed in subsection 3.1 (see Tables 2-3), respectively. Thus, the linear weighting scheme has a lower annualized return but an even lower volatility, resulting in a higher Sharpe ratio. The case $x = \widetilde{\Delta r}$, which is the one with performances in line with the indicator $\mathrm{P}(\Delta r \nearrow)$, when used to build a linear weighting scheme[10] results in a Sharpe ratio of 101.89% and an annualized return of 13.18%, to be confronted with 117.21% and 18.96%, respectively. Thus, in this case, the linear weighting scheme has a worse risk-adjusted performance with respect to the decile portfolio scheme. This is true also for the other competing indicator: $\widetilde{\Delta p}$. These considerations do not impact the results discussed insofar, and moreover seem to further indicate that the indicator $\mathrm{P}(\Delta r \nearrow)$ is better in capturing useful information for pairs-trading portfolio investing.

4.4 The role of sectors

In order to investigate the role of sectors we compute the entropy of portfolios with respect to sectors. Namely, indicating with s_t^i the sector[11] of stock i at time t, and with the sequence $(s_t^1, \ldots, s_t^n)$ the collection of sectors at time t in the portfolio, we compute its normalized entropy as:

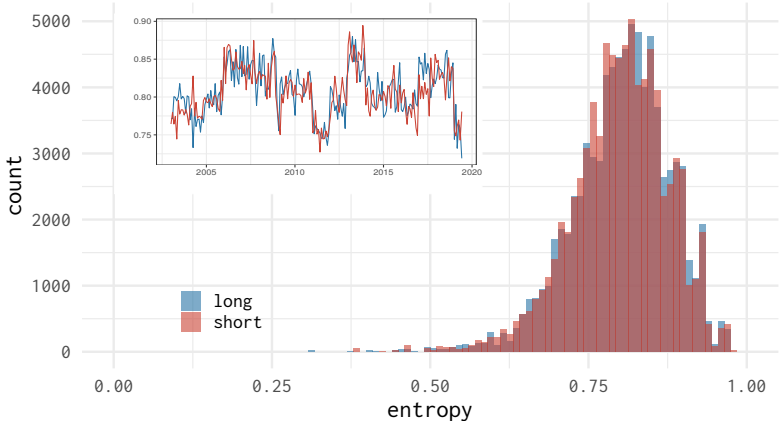
$$H(t) = \frac{-\sum_{\alpha=1}^{S} p_\alpha(t) \log_2 p_\alpha(t)}{\log_2 \min(n, S)}; \tag{12}$$

where S is the total number of possible sectors and $p_\alpha(t) = \#\{s_t^i = \alpha\}/n$ is the frequency of stocks belonging to sector α in the portfolio at time t. Normalized entropy of a sequence measures the concentration of that sequence with respect to a certain variable, in this case sector membership: when all the stocks in the portfolio belong to the same sector then $H(t) = 0$, while when each stock belongs to a different sector $H(t) = 1$.

Figure 7 shows the concentration of sectors in the *Top-Bottom* portfolios built with deciles of $\mathrm{P}(\Delta r \nearrow)$: both the daily distribution and the (average) monthly evolution of the entropy suggest a very small concentration both in the long and in the short part of the strategy, with the largest contribution coming from $H \sim 0.8$. Consider, for reference, that the median number of stocks in the strategy (long or short) is 20 and that the median number of sectors represented in the portfolios is 13:

[10] Note that in this case, as for all other indicators except $\mathrm{P}(\Delta r \nearrow)$, one actually has $w_t \propto -(x_t - \bar{x}_t)$.
[11] We employ the NACE Rev.2 classification at level 2.

Figure 7: Sector concentration. Distribution of sectorial daily entropy as by equation (12) for the *Top-Bottom* decile portfolio with respect to the indicator $\mathrm{P}(\Delta r \nearrow)$. The time evolution of the monthly average of the entropy is displayed in the inset.



this, by itself, indicates that the concentration cannot be too high. Moreover, the maximum possible (normalized) entropy given $n = 20$ and a number of sectors equal to 13 is $H = \log_2(13)/\log_2(20) \simeq 0.86$. This corresponds to the case in which $p_\alpha = 1/13$, i.e. all represented sectors are equally likely, without any source of concentration. Thus, sectors in the portfolios are, on average, only slightly more concentrated than uniform probability. Hence, our proposed approach does not seem to be confounded by any specific sectorial portfolio exposure, being able to spread the allocation evenly over a sufficiently large number of sectors in both the long and short sides of the strategy.

4.5 A different LSTM framework

We also investigate a different LSTM framework whose output variable simply refers to the prediction of an increase of stocks' returns at $t+1$ without any cointegration peer group assignment. This scenario is in line with the one proposed by Fischer and Krauss (2018) for predicting out-of-sample directional market performances for the stocks composing the S&P 500. Hence, under this framework, we are omitting the role of the cointegration filtering step. Table 9 shows that the *Top-Bottom* strategy is able to generate a positive and sizeable result in terms of alpha performances and raw returns even once adjusted for risk. However, these results are considerably lower than those generated by $\mathrm{P}(\Delta r \nearrow)$, thus supporting the use of our proposed framework to exploit pairs-trading opportunities.

Table 9: Performance measures for $\mathbf{P}(r_{t+1} > 0)$. Different performance measures (in percentage) obtained by equally investing in stocks composing decile portfolios defined using as sorting criteria $\mathrm{P}(r_{t+1} > 0)$. The measures reported by row are: the alphas from a 5-factor and a 7-factor models, the raw returns, the Sharpe ratios (SR). The holding period is $h = 1$ day. Newey-West t-statistics are reported in parenthesis. Data are annualized and refer to the period from January 2003 to June 2019.

	1	2	3	4	5	6	7	8	9	10	*Top-Bottom*
annualized 5-factor α	-3.26	-2.26	-0.43	2.50	2.47	3.99	3.76	6.13	8.03	11.78	15.55
	(-1.31)	(-1.26)	(-0.27)	(1.98)	(2.18)	(3.30)	(3.29)	(4.65)	(5.14)	(3.23)	(3.17)
annualized 7-factor α	-2.01	-1.27	0.07	3.01	2.79	4.02	3.51	5.43	7.09	9.95	12.21
	(-0.83)	(-0.69)	(0.05)	(2.44)	(2.51)	(3.48)	(3.21)	(4.22)	(4.65)	(2.93)	(2.38)
raw returns	5.13	6.99	8.78	11.89	11.74	13.46	13.25	15.63	17.77	20.86	12.40
	(1.59)	(1.91)	(2.06)	(2.45)	(3.11)	(3.23)	(3.55)	(3.51)	(3.94)	(4.23)	(3.27)
SR	18.92	30.90	42.06	59.99	62.10	71.08	70.23	82.16	89.55	83.69	68.31

4.6 A comparison against other machine learning approaches

In this subsection we compare the predictions of our proposed LSTM framework against other common machine learning approaches, namely the Logistic regression, the Gradient boosting machine and the Feedforward Neural Network.

The crucial difference between models with temporal structure (such as RNN and thus LSTM networks) and "standard" machine learning models lays in the fact that the former naturally deals with time series data, while the latter do not. In order to summarize the temporal nature of the inputs to be fed into standard machine learning methods, we employ the following strategy: we take as predictors the same 3 variables as in the LSTM case, namely Δr, returns and volumes, and we compute an Exponentially Weighted Moving Average of each of them with 4 different time windows in the past: 1 month, 2 months, 6 months, 1 year. This is done on a daily basis for the entire training period of 3 years, resulting in approximately 750 - 240 training observations per stock, as in the LSTM case. Finally, data are standardized as in the LSTM case.

Logistic regression is used as the simplest approach. An L_1 regularization is included (LASSO), with hyperparameter fixed by 10-fold cross-validation maximizing the Area Under the ROC. As in Fischer and Krauss (2018), we use the Logistic regression to gauge the incremental contribution of a much more complex approach such as the LSTM.

Gradient boosting machine generates predictions by means of an ensemble of weak prediction models, usually decision trees, as in random forests. Differently from random forests, Gradient boosting machines train the ensemble of trees sequentially, each time fitting the new tree to the residuals of the previous step. The number of trees and the interaction depth are chosen by 5-fold cross-validation maximizing the Area Under the ROC: the average value of the optimal number of trees is 150, the average value of interaction depth is 3. The minimum number of observations per node is instead fixed at 10 and the value of shrinkage at 0.1.

Finally, we employ Feedforward Neural Network as a form of artificial neural network in which inputs are processed irrespective of their (time) order. A Neural Network with at most 3 hidden layers is employed, whose number of nodes per layer is fixed by 5-fold cross-validation maximizing the Area Under the ROC, resulting on average in 13 nodes for layer 1, 10 in layer 2 and 9 in layer 3. In 6 out of 17 years a 2-layer Network is chosen, while a 3-layer one results optimal for the other 11 cases.

Table 10 shows the market performances of decile portfolios selected on the basis of $\mathrm{P}(\Delta r \nearrow)$ computed via these alternative machine learning approaches compared to those of the LSTM approach. Interestingly, we observe that in general performances of *Top-Bottom* strategies are significant not only economically but also statistically. However, the *Top-Bottom* strategies associated with the Logistic regression presents weaker performances both in terms of raw returns and alphas, while on the contrary Gradient boosting machine and Feedforward Neural Network show very similar market results, with Gradient boosting machine more able to generate valuable extra-performances. Although this work is not intended to compare and select the most performing algorithm, still we can notice that the *Top-Bottom* strategy based on LSTM reaches the highest levels of market performance with respect to the other approaches, especially once we control for dynamic exposures to factors which represents a pillar of our investigation strategy. Finally, by employing the Diebold and Mariano (2002)'s comparison test, we find strong supporting evidence for the H_0 that the LSTM has superior forecasting performances than the other alternative techniques (p-values ≈ 1): indeed, these 3 alternative machine learning models reach very similar prediction performances of about 60% accuracy on average, to be compared to $\sim$ 75% of LSTM. This comparison analysis seems to support the fact that choosing a

recurrent model, which is inherently suitable for time series data, allows to outperform very simple statistical models, as the Logistic regression, as well as more refined machine learning techniques, such as Feedforward Neural Networks, that could be employed for tasks similar to those that in this work are performed by the LSTM.

Table 10: Comparison with other ML approaches. The table shows the performances of our proposed LSTM framework along with those related to the Logistic regression, the Gradient Boosting Machine and the FeedForward Neural Network. The holding period is $h = 1$ day. Newey-West t-statistics are reported in parenthesis. Data are annualized and refer to the period from January 2003 to June 2019.

		1	2	3	4	5	6	7	8	9	10	*Top-Bottom*
LSTM	raw returns	4.04	7.12	8.58	9.25	10.91	16.24	13.32	21.85	18.31	24.57	18.32
		(1.27)	(1.91)	(2.24)	(2.37)	(2.83)	(3.98)	(3.31)	(4.81)	(4.04)	(5.14)	(4.98)
	annualized 5-factor α	-4.21	-1.39	-0.54	-0.20	1.48	6.49	3.64	11.70	8.19	14.25	19.27
		(-1.90)	(-0.69)	(-0.28)	(-0.12)	(0.99)	(3.39)	(1.85)	(5.59)	(3.86)	(6.00)	(4.91)
	annualized 7-factor α	-3.31	-0.97	-0.68	0.17	1.39	6.33	3.50	11.71	7.76	13.67	17.55
		(-1.39)	(-0.49)	(-0.35)	(0.11)	(0.91)	(3.49)	(1.81)	(5.06)	(3.78)	(5.45)	(4.45)
Logistic regression	raw returns	4.52	8.66	9.09	11.72	11.85	12.35	15.33	17.95	20.76	21.15	14.74
		(1.34)	(2.25)	(2.34)	(3.04)	(3.27)	(3.24)	(4.06)	(4.63)	(4.93)	(4.02)	(4.09)
	annualized 5-factor α	-3.75	-0.45	0.13	2.58	2.41	2.75	5.62	8.04	10.56	11.16	15.49
		(-1.40)	(-0.24)	(0.08)	(1.71)	(1.55)	(1.71)	(3.30)	(4.76)	(5.24)	(4.05)	(3.82)
	annualized 7-factor α	-1.78	0.65	0.86	2.86	2.57	2.64	5.32	7.41	9.50	8.95	10.93
		(-0.68)	(0.35)	(0.53)	(1.91)	(1.73)	(1.64)	(3.06)	(4.54)	(4.69)	(3.42)	(2.76)
Gradient Boosting	raw returns	2.51	8.98	8.75	13.35	11.43	14.65	14.24	17.69	21.12	21.05	17.02
		(1.05)	(2.36)	(2.23)	(3.31)	(3.13)	(3.83)	(3.38)	(4.36)	(4.66)	(4.44)	(5.11)
	annualized 5-factor α	-5.92	0.00	-0.14	3.89	2.20	4.70	4.36	7.64	10.86	11.66	18.68
		(-2.67)	(0.00)	(-0.08)	(2.55)	(1.39)	(2.88)	(2.39)	(4.29)	(4.77)	(4.62)	(4.92)
	annualized 7-factor α	-4.46	1.45	1.14	4.71	2.51	4.52	3.90	6.74	9.00	9.58	14.70
		(-1.96)	(0.77)	(0.65)	(3.06)	(1.46)	(2.75)	(2.17)	(3.90)	(4.11)	(3.98)	(4.00)
FeedForward NN	raw returns	2.67	8.66	10.25	11.41	12.79	16.11	14.45	18.56	17.30	21.31	17.10
		(1.01)	(2.31)	(2.65)	(2.94)	(3.10)	(4.11)	(3.51)	(4.39)	(3.97)	(4.54)	(4.32)
	annualized 5-factor α	-5.61	-0.38	1.14	2.28	3.37	6.10	4.73	8.73	7.25	11.33	17.95
		(-2.01)	(-0.21)	(0.67)	(1.30)	(1.77)	(3.52)	(2.51)	(4.91)	(3.78)	(4.33)	(3.99)
	annualized 7-factor α	-3.84	0.77	1.99	3.10	3.63	5.76	4.09	7.74	6.15	9.45	13.82
		(-1.47)	(0.42)	(1.19)	(1.84)	(1.90)	(3.39)	(2.23)	(4.52)	(3.14)	(3.95)	(3.49)

5 Conclusions

Reversal effect consists in the fact that temporarily market deviations are likely to correct and finally converge again. We investigate such effects through the lens of the detection of market anomalies. Deep learning techniques have been widely applied to detect patterns in financial markets. Here, we aim at investigating these complex non-linear relationships among financial time series by proposing a Long Short-Term Memory (LSTM) network framework to study the market patterns of a large sample of stocks and test its predictive performance within the context of pairs-trading strategies.

Pairs-trading opportunities are, in fact, perceived as deviations from the equilibrium due to market reactions, which are temporary and will be timely corrected by reverting their market patterns. To highlight the role of reversal effect, we also decide to rely on cointegrated stocks, thus focusing on financial time series for which deviations from long-run patterns should be temporary until they revert to the mean. Differently from typical machine learning applications, we do not propose to use an LSTM framework to simply produce buy or sell signals. We exploit it to generate an outcome representative of the likelihood of a stock to present in the near future an increase in its market return with respect to its cointegrated group of peers. This aspect is crucial since we want to combine such predictions from LSTM with common trading practices based on sorting stocks according to either price or returns gaps. Hence, our proposed approach is not intended to design a more performing indicator than those typically employed to construct pairs-trading strategies. The inclusion of the outcomes of the LSTM aims to show in fact whether valuable signals extracted from financial time

series contain information that can complement the one already embedded in price or returns gaps, thereby generating even better portfolio performances once jointly combined. Our analysis shows that pairs-trading strategies based on price or returns gaps can reach even better performances once the probability of stocks future market deviations with respect to peers are included as an additional criterion for constructing portfolios. We reveal, therefore, that the LSTM outcomes contribute to provide predictive signals whose information content go above and beyond the one typically embedded in both price and returns gaps. More specifically, we find that LSTM outcomes can be utilized to practically guide the construction of strategies based on short-term reversal or momentum.

The deep learning approach we have here pursued still has many possible refinements and details to be further analyzed, such as a more careful and systematic study of the hyperparameters (e.g., network architecture, temporal sequence length, optimization algorithm, among others), an analysis of the temporal stability of the models (i.e., sensitivity to train and test window lengths), a calibration of different models for different groups of stocks rather than one model for all stocks. Not to mention the possibility of adopting a different deep learning framework, e.g. a Gradient Recurrent Unit network (Cho et al., 2014), or even the more recent and promising Transformer approach to sequence learning introduced by Vaswani et al. (2017) for Natural Language Processing with already several applications to time series analysis even if not, to our knowledge, to financial time series. More in general, although the capacity of our proposed approach to provide valuable signals is investigated under different perspectives (e.g., variations in the investment horizons, impact of transaction costs, sectorial compositions and portfolio weighting scheme), a detailed analysis of what patterns in past time series data are leveraged by the model to produce informative signals on future returns would be extremely useful to further investigate the nature of market anomalies in this setting.

References

Adrian, T. and F. Franzoni (2009). Learning about beta: Time-varying factor loadings, expected returns, and the conditional capm. *Journal of Empirical Finance 16*(4), 537–556.

Andreou, P. C., C. Charalambous, and S. H. Martzoukos (2008). Pricing and trading european options by combining artificial neural networks and parametric models with implied parameters. *European Journal of Operational Research 185*(3), 1415–1433.

Atsalakis, G. S. and K. P. Valavanis (2009). Surveying stock market forecasting techniques–Part II: Soft computing methods. *Expert Systems with Applications 36*(3), 5932–5941.

Avellaneda, M. and J.-H. Lee (2010). Statistical arbitrage in the us equities market. *Quantitative Finance 10*(7), 761–782.

Avramov, D., T. Chordia, and A. Goyal (2006). Liquidity and autocorrelations in individual stock returns. *The Journal of Finance 61*(5), 2365–2394.

Bao, W., J. Yue, and Y. Rao (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PloS one 12*(7).

Bekiros, S. D. (2010). Fuzzy adaptive decision-making for boundedly rational traders in speculative stock markets. *European Journal of Operational Research 202*(1), 285–293.

Bengio, Y., P. Simard, P. Frasconi, et al. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks 5*(2), 157–166.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blackburn, D. W. and N. Cakici (2017). Overreaction and the cross-section of returns: International evidence. *Journal of Empirical Finance 42*, 1–14.

Blitz, D., J. Huij, S. Lansdorp, and M. Verbeek (2013). Short-term residual reversal. *Journal of Financial Markets 16*(3), 477–504.

Bloomfield, R. J., W. B. Tayler, and F. Zhou (2009). Momentum, reversal, and uninformed traders in laboratory markets. *The Journal of Finance 64*(6), 2535–2558.

Bogomolov, T. (2013). Pairs trading based on statistical variability of the spread process. *Quantitative Finance 13*(9), 1411–1430.

Borovkova, S. and I. Tsiamas (2019). An ensemble of lstm neural networks for high-frequency stock market classification. *Journal of Forecasting 38*(6), 600–619.

Boudoukh, J., M. P. Richardson, and R. Whitelaw (1994). A tale of three schools: Insights on autocorrelations of short-horizon stock returns. *Review of Financial Studies 7*(3), 539–573.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance 52*(1), 57–82.

Chan, L. K., N. Jegadeesh, and J. Lakonishok (1996). Momentum strategies. *The Journal of Finance 51*(5), 1681–1713.

Chen, H., S. Chen, Z. Chen, and F. Li (2017). Empirical investigation of an equity pairs trading strategy. *Management Science 65*(1), 370–389.

Cheng, S., A. Hameed, A. Subrahmanyam, and S. Titman (2017). Short-term reversals: The effects of past returns and institutional exits. *Journal of Financial and Quantitative Analysis 52*(1), 143–173.

Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chollet, F. et al. (2015). Keras. `https://keras.io`.

Chordia, T., R. Roll, and A. Subrahmanyam (2002). Order imbalance, liquidity, and market returns. *Journal of Financial Economics 65*(1), 111–130.

Clegg, M. and C. Krauss (2018). Pairs trading with partial cointegration. *Quantitative Finance 18*(1), 121–138.

Cohen, R. B., J. D. Coval, and L. Pástor (2005). Judging fund managers by the company they keep. *The Journal of Finance 60*(3), 1057–1096.

Conrad, J., M. N. Gultekin, and G. Kaul (1997). Profitability of short-term contrarian strategies: Implications for market efficiency. *Journal of Business & Economic Statistics 15*(3), 379–386.

Coval, J. and E. Stafford (2007). Asset fire sales (and purchases) in equity markets. *Journal of Financial Economics 86*(2), 479–512.

Cremers, M. and A. Pareek (2014). Short-term trading and stock return anomalies: Momentum, reversal, and share issuance. *Review of Finance 19*(4), 1649–1701.

Da, Z. and P. Gao (2010). Clientele change, liquidity shock, and the return on financially distressed stocks. *Journal of Financial and Quantitative Analysis 45*(1), 27–48.

Da, Z., Q. Liu, and E. Schaumburg (2013). A closer look at the short-term return reversal. *Management Science 60*(3), 658–674.

De Groot, W., J. Huij, and W. Zhou (2012). Another look at trading costs and short-term reversal profits. *Journal of Banking & Finance 36*(2), 371–382.

Diebold, F. X. and R. S. Mariano (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics 20*(1), 134–144.

Do, B. and R. Faff (2010). Does simple pairs trading still work? *Financial Analysts Journal 66*(4), 83–95.

Do, B. and R. Faff (2012). Are pairs trading profits robust to trading costs? *Journal of Financial Research 35*(2), 261–287.

Dunis, C. L. and R. Ho (2005). Cointegration portfolios of european equities for index tracking and market neutral strategies. *Journal of Asset Management 6*(1), 33–52.

Engle, R. F. and C. W. Granger (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251–276.

Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business 38*(1), 34–105.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance 25*(2), 383–417.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Fama, E. F. and K. R. French (1996). Multifactor explanations of asset pricing anomalies. *The Journal of Finance 51*(1), 55–84.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(1), 1–22.

Fischer, T. and C. Krauss (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research 270*(2), 654–669.

Flori, A., F. Lillo, F. Pammolli, and A. Spelta (2019). Better to stay apart: asset commonality, bipartite network centrality, and investment strategies. *Annals of Operations Research*, 1–37.

Focardi, S. M., F. J. Fabozzi, and I. K. Mitov (2016). A new approach to statistical arbitrage: Strategies based on dynamic factor models of prices and their performance. *Journal of Banking & Finance 65*, 134–155.

Gatev, E., W. N. Goetzmann, and K. G. Rouwenhorst (2006). Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies 19*(3), 797–827.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning*. MIT press.

Graves, A. (2012). *Supervised sequence labelling with recurrent neural networks*. Springer.

Gu, S., B. Kelly, and D. Xiu (2018). Empirical asset pricing via machine learning. Technical report, National Bureau of Economic Research.

Gutierrez, R. C. and C. A. Prinsky (2007). Momentum, reversal, and the trading behaviors of institutions. *Journal of Financial Markets 10*(1), 48–75.

Hameed, A. and G. M. Mian (2015). Industries and stock return reversals. *Journal of Financial and Quantitative Analysis 50*(1-2), 89–117.

Heaton, J., N. Polson, and J. H. Witte (2017). Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry 33*(1), 3–12.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 1735–1780.

Huang, W., Q. Liu, S. G. Rhee, and L. Zhang (2009). Return reversals, idiosyncratic risk, and expected returns. *The Review of Financial Studies 23*(1), 147–168.

Huck, N. (2009). Pairs selection and outranking: An application to the S&P 100 index. *European Journal of Operational Research 196*(2), 819–825.

Huck, N. (2010). Pairs trading and outranking: The multi-step-ahead forecasting case. *European Journal of Operational Research 207*(3), 1702–1716.

Huck, N. (2019). Large data sets and machine learning: Applications to statistical arbitrage. *European Journal of Operational Research 278*(1), 330–342.

Huck, N. and K. Afawubo (2015). Pairs trading and selection methods: is cointegration superior? *Applied Economics 47*(6), 599–613.

Ioffe, S. and C. Szegedy (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jacobs, H. and M. Weber (2015). On the determinants of pairs trading profitability. *Journal of Financial Markets 23*, 75–97.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of Finance 45*(3), 881–898.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance 48*(1), 65–91.

Jegadeesh, N. and S. Titman (1995a). Overreaction, delayed reaction, and contrarian profits. *The Review of Financial Studies 8*(4), 973–993.

Jegadeesh, N. and S. Titman (1995b). Short-horizon return reversals and the bid-ask spread. *Journal of Financial Intermediation 4*(2), 116–132.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica: journal of the Econometric Society*, 1551–1580.

Kazemi, H., T. Schneeweis, and B. Gupta (2004). Omega as a performance measure. *Journal of Performance Measurement 8*, 16–25.

Keating, C. and W. F. Shadwick (2002). A universal performance measure. *Journal of Performance Measurement 6*(3), 59–84.

Kim, A., Y. Yang, S. Lessmann, T. Ma, M.-C. Sung, and J. E. Johnson (2020). Can deep learning predict risky retail investors? a case study in financial risk behavior forecasting. *European Journal of Operational Research 283*(1), 217–234.

Kraus, M., S. Feuerriegel, and A. Oztekin (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research 281*(3), 628–641.

Krauss, C. (2017). Statistical arbitrage pairs trading strategies: Review and outlook. *Journal of Economic Surveys 31*(2), 513–545.

Krauss, C., X. A. Do, and N. Huck (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research 259*(2), 689–702.

Kuhn, M. (2020). *caret: Classification and Regression Training*. R package version 6.0-86.

Lasfer, M. A., A. Melnik, and D. C. Thomas (2003). Short-term reaction of stock markets in stressful circumstances. *Journal of Banking & Finance 27*(10), 1959–1977.

Ledoit, O. and M. Wolf (2008). Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance 15*(5), 850–859.

Lehmann, B. N. (1990). Fads, martingales, and market efficiency. *The Quarterly Journal of Economics 105*(1), 1–28.

Liu, B., L.-B. Chang, and H. Geman (2017). Intraday pairs trading strategies on high frequency data: The case of oil companies. *Quantitative Finance 17*(1), 87–100.

Lo, A. W. and A. C. MacKinlay (1990). When are contrarian profits due to stock market overreaction? *The Review of Financial Studies 3*(2), 175–205.

Malkiel, B. G. (1973). *A random walk down Wall Street*. WW Norton & Company.

Patel, J., S. Shah, P. Thakkar, and K. Kotecha (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications 42*(1), 259–268.

Peterson, B. G. and P. Carl (2019). *PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis*. R package version 1.5.3.

Puspaningrum, H., Y.-X. Lin, and C. M. Gulati (2010). Finding the optimal pre-set boundaries for pairs trading strategy based on cointegration technique. *Journal of Statistical Theory and Practice 4*(3), 391–419.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rad, H., R. K. Y. Low, and R. Faff (2016). The profitability of pairs trading strategies: distance, cointegration and copula methods. *Quantitative Finance 16*(10), 1541–1558.

Richards, A. J. (1997). Winner-loser reversals in national stock market indices: Can they be explained? *The Journal of Finance 52*(5), 2129–2144.

Schnaubelt, M., T. Fischer, and C. Krauss (2020). Separating the signal from the noise–financial machine learning for twitter. *Journal of Economic Dynamics and Control*, 103895.

Sermpinis, G., K. Theofilatos, A. Karathanasopoulos, E. F. Georgopoulos, and C. Dunis (2013). Forecasting foreign exchange rates with adaptive neural networks using radial-basis functions and particle swarm optimization. *European Journal of Operational Research 225*(3), 528–540.

Stübinger, J. and S. Endres (2018). Pairs trading with a mean-reverting jump–diffusion model on high-frequency data. *Quantitative Finance 18*(10), 1735–1751.

Stübinger, J., B. Mangold, and C. Krauss (2018). Statistical arbitrage with vine copulas. *Quantitative Finance 18*(11), 1831–1849.

Subrahmanyam, A. (2005). Distinguishing between rationales for short-horizon predictability of stock returns. *Financial Review 40*(1), 11–35.

Tieleman, T. and G. Hinton (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning 4*(2), 26–31.

Troiano, L., E. M. Villa, and V. Loia (2018). Replicating a trading strategy by means of lstm for financial industry applications. *IEEE transactions on industrial informatics 14*(7), 3226–3234.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.

Vidyamurthy, G. (2004). *Pairs Trading: quantitative methods and analysis*, Volume 217. John Wiley & Sons.

Whited, T. M. and G. Wu (2006). Financial constraints risk. *The Review of Financial Studies 19*(2), 531–559.

Zhu, Z. and K. Yung (2016). The interaction of short-term reversal and momentum strategies. *The Journal of Portfolio Management 42*(4), 96–107.