The Real-Life Performance of Market Timing with Moving Average and Time-Series Momentum Rules*

Valeriy Zakamulin†

First draft: April 1, 2013
This revision: November 8, 2013

Abstract

In this paper we revisit the myths about the superior performance of the market timing strategies with moving average and time-series momentum rules. These active timing strategies are very appealing to investors because of their extraordinary simplicity and because they promise substantial advantages over their passive counterparts (see, for example, the paper by M. Faber (2007) "A Quantitative Approach to Tactical Asset Allocation" published in the Journal of Wealth Management). However, "too good to be true" reported performance of these market timing rules raises a legitimate concern whether this performance is realistic and whether the investors can hope that the expected future performance will be the same as the documented historical performance. We argue that the reported performance of market timing strategies usually contains a considerable data-mining bias and ignores important market frictions. In order to deal with these issues, we perform out-of-sample tests of these two timing models where we account for realistic transaction costs. Our findings reveal that at best the real-life performance of the market timing strategies is only marginally better than that of the passive counterparts.

Key words: technical analysis, market timing, simple moving average, time-series momentum, out-of-sample testing, bootstrap simulation

JEL classification: G11, G17.

*The author is grateful to the following individuals for their helpful comments and suggestions regarding earlier drafts of this paper: Kamal Doshi; Henry Stern; Steen Koekebakker; the participants at the faculty seminar at the University of Agder (Kristiansand, Norway), the 2013 Forecasting Financial Markets Conference (Hannover, Germany), and the 2013 International Conference on Computational and Financial Econometrics (London, UK). Any remaining errors in the manuscript are the author's responsibility.

†a.k.a. Valeri Zakamouline, Faculty of Economics and Social Sciences, University of Agder, Service Box 422, 4604 Kristiansand, Norway, Tel.: (+47) 38 14 10 39, E-mail: Valeri.Zakamouline@uia.no

1 Introduction

Market timing is an active strategy that attempts to outperform the passive buy-and-hold strategy by anticipating the future direction of a financial market. At the heart of every market timing strategy lies a belief that the future security prices are predictable, typically through the use of technical indicators computed from the past prices. One of the most common beliefs is that security prices move in trends. Consequently, trend following is the most widespread market timing strategy that tries to jump on a trend and ride it. However, whereas technical analysis has been extensively employed by practitioners over the last century, academics had long been skeptical about its usefulness. The study by Brock, Lakonishok, and LeBaron (1992) probably marked the onset of change in the academics' attitude towards technical analysis of financial markets. In particular, Brock et al. (1992) perform the tests of the most popular trading rules and find significant trading profits (exclusive of transaction costs) generated by some trading rules. Still, before the decade of 2000s, on the academic side there was common agreement that market timing does not work. For example, Sullivan, Timmermann, and White (1999) find that the in-sample profitable technical trading rules perform poorly out-of-sample. The inferiority of the market timing strategies is also confirmed by Bauer and Dahlquist (2001) among others.

The decade of 2000s was marked by an explosion in the academic literature on technical analysis of financial markets (see, for example, Park and Irwin (2007)). This is because it turned out that the use of some trend following trading rules would help investors to avoid massive losses during the severe bear markets occurred throughout the decade of 2000s. As a rule, these trend following market timing strategies exploit various technical indicators based on moving averages. Most often one uses the simple moving average (SMA) rule which prescribes buying securities (moving to cash) when the security price is above (below) the k-month moving average. For example, Faber (2007) reports the superior performance of the 10-month SMA rule. Subsequently the profitability of the 10-month SMA rule has been confirmed in several published papers (see, among others, Gwilym, Clare, Seaton, and Thomas (2010) and Kilgallen (2012)) and numerous unpublished working papers. Another popular market timing strategy employs the time-series momentum[1] (MOM) rule which prescribes buying securities (moving

[1]The asset price anomaly known as "momentum" in the academic finance literature is documented for the first time by Jegadeesh and Titman (1993). This "momentum" effect focuses on the relative performance of

to cash) when the k-month moving average increases (decreases). Moskowitz et al. (2012) are probably the first to report the profitability of the 12-month MOM rule. This paper, in the same manner as the paper by Faber (2007), is followed now by numerous studies that confirm the superior performance of the MOM rule in many financial markets.

A typical study of a trend following strategy reports that market timing allows investors both to enhance returns and greatly reduce risk as compared to the buy-and-hold strategy, producing very attractive Sharpe ratios that are from 50% to 100% higher than that of the passive counterpart. As a natural result of these studies, market timing strategies has dramatically increased in popularity in recent years, and the total assets managed in accordance with market timing rules constantly rise. However, "too good to be true" performance raises a legitimate concern whether the reported performance is realistic and whether the investors can hope that the expected future performance of market timing rules will be the same as the documented historical performance. In this paper we argue that the reported performance of market timing strategies usually contains a substantial data-mining bias. In addition, in a study of a technical trading rule one usually ignores important market frictions such as, for example, trading costs.

The data-mining fallacy that has been repeated over and over again in many studies consists in the following. Using a full historical data sample one tests many k-month moving average trading rules and picks up a rule that performs best. One then reports the performance of the best trading rule in a back test and either explicitly or implicitly assumes that the expected future performance of this rule will be the same as the past performance. Yet, as a matter of fact, one should expect a much poorer performance in the future than the reported performance in the past. Out-of-sample performance deterioration is a very well-known problem in technical analysis (see Aronson (2006), Chapter 6). Even if there is no obvious data snooping in some studies, the data-mining issue may be relevant nevertheless. For example, in the studies by Brock et al. (1992), Siegel (2002) (see Chapter 17), and Faber (2007) the authors acknowledge that the 10-month (200-day) SMA rule is the most popular trading rule among practitioners. Consequently, the reported performance of this rule in the above mentioned studies might be contaminated by data-mining. More specifically, it is quite natural to suppose that prior to

securities in the cross-section. The term "time-series momentum" is introduced by Moskowitz, Ooi, and Pedersen (2012). The "time-series momentum" focuses purely on a security's own past performance. Throughout the paper when we use the term "momentum", we always mean the "time-series momentum".

these studies practitioners back-tested many k-month SMA rules and the 10-month SMA rule was selected as the rule with the best observed performance. Using (almost) the same data as in the study by Faber (2007), we also find that the 10-month SMA rule is the best trading rule in the back test. Thus, one can reasonably suspect that the reported performance of this rule might be highly overstated as compared with the real-life performance.

Conventional wisdom says that the out-of-sample performance of a trading strategy provides an unbiased estimate of it's real-life performance (see Sullivan et al. (1999), White (2000), and Aronson (2006)). In traditional out-of-sample testing the historical data are partitioned into in-sample and out-of-sample subsets. Then the best rule discovered in the mined data (in-sample) is evaluated on the out-of-sample data. In our paper we challenge this conventional wisdom and argue that one should also be skeptical to the trading rule's out-of-sample performance. The main concern is that no guidance exists on how to choose the split point between the initial in-sample and out-of-sample subsets. The choice seems can be done completely arbitrary without affecting the results of out-of-sample tests. Yet, as a matter of fact, it turns out that the out-of-sample performance is very sensitive to the choice of the split point.[2] The reason for this is that market timing strategies do not consistently deliver superior performance relative to a benchmark. On the contrary, the superior performance is confined to relatively short historical episodes. We demonstrate in this paper that, depending on the choice of a split point, the out-of-sample performance of a market timing rule might by either superior or inferior as compared to that of the passive counterpart.

The goal of this paper is to obtain unbiased estimates of the real-life historical performance of market timing strategies with moving average and time-series momentum rules. These estimates are supposed to provide investors with reliable assessments of the rules' expected future performance. We consider several alternative passive benchmarks and account for realistic transaction costs. In order to overcome the deficiencies of the traditional out-of-sample performance measurement procedure, we propose and implement a robust method of performance measurement in out-of-sample tests. In this method the out-of-sample performance is insensitive to how we apportion data between the in-sample and out-of-sample subsets. The method

[2]In the context of out-of-sample forecast evaluation, recently Rossi and Inoue (2012) and Hansen and Timmermann (2013) have also made exactly the same observation. Namely, the results of out-of-sample forecast tests depend crucially on how the sample split point is determined. These authors propose new methodologies for evaluating out-of-sample forecasting performance that are robust to the choice of the split point.

is based on simulating many alternative historical realizations of the underlying data series. We use a semi-parametric stationary block-bootstrap simulation method which preserves all relevant statistical features of the historical data series. After each simulation, the traditional out-of-sample performance measurement procedure is repeated. In the end, we compute the average out-of-sample performance which is supposed to be a trustworthy estimate of the real-life historical performance.

The rest of the paper is organized as follows. Section 2 describes our data and methodology, while Section 3 outlines the empirical results. Section 4 presents a discussion of the empirical results where we revisit the myths about the superior performance of the market timing strategies. Section 5 concludes the paper.

2 Data and Methodology

2.1 Data

In our study we use data on two stock market indices, two bond market indices, and the risk-free rate of return. The two stock market indices are the Standard and Poor's Composite stock price index and the Dow Jones Industrial Average index. The two bond market indices are the long-term and intermediate-term US government bond indices. Our sample period begins in January 1926 and ends in December 2012 (87 full years), giving a total of 1044 monthly observations. This particular choice for the sample period is motivated by the fact that high quality monthly stock price and dividend data, as well as the bond price and return data, are available only from 1926 (from the Center for Research in Security Prices).

We use the monthly Standard and Poor's Composite stock price index data and corresponding dividend data provided by Amit Goyal.[3] The Standard and Poor's Composite index is a value-weighted stock index. From 1926 to 1956, the index data come from various reports of the Standard and Poor's. From 1957 this index is identical to the Standard and Poor's 500 index which is intended to be a representative sample of leading companies in leading industries within the US economy. Stocks in the index are chosen for market size, liquidity, and industry group representation. For more details about the construction of the index and its dividend series see Goyal and Welch (2008).

[3]See `http://www.hec.unil.ch/agoyal/`.

The Dow Jones Industrial Average index is a price-weighted stock index. Specifically, the DJIA is an index of the prices of 30 large US corporations selected to represent a cross section of US industry. As for today, the components of the DJIA have changed 48 times in its 117 year history. Changes in the composition of the DJIA are made to reflect changes in the companies and in the economy. The DJIA index values for the total sample period and dividends for the period 1988 to 2012 are provided by S&P Dow Jones Indices LLC, a subsidiary of the McGraw-Hill Companies.[4] The dividends for the period 1926 to 1987 are obtained from Barron's.[5]

The bond data are from the Ibbotson SBBI 2013 Classic Yearbook. We use both the capital appreciation returns and total returns on the long-term and intermediate-term government bonds. The risk-free rate of return is also provided by Amit Goyal. In particular, the risk-free rate of return for our sample period is the Treasury bill rate.

Table 1 summarizes the descriptive statistics for the aforementioned data. Specifically, for each index the table reports the means, standard deviations, skewness, kurtosis, and autocorrelation coefficients for the capital appreciation return, dividend yield, and total return over the overall sample period 1926 to 2012. For the Treasury bill rate the table reports the descriptive statistics for the total return. From the table it is evident that all the dividend yields and the total return on the Treasury bills are highly persistent stochastic variables. The table also shows that the capital appreciation returns and total returns for all the indices, but the long-term bonds, exhibit positive and statistically significant autocorrelations. The highest degree of autocorrelation is observed for the intermediate-term bond returns, the next highest for the returns on the Standard and Poor's Composite stock price index.

2.2 Technical trading rules and returns to the market timing strategy

In our study we examine two technical trading rules: the simple moving average rule and the (time-series) momentum rule. Every technical trading rule prescribes investing in the stocks/bonds when a Buy signal is generated and moving to the risk-free asset when a Sell signal is generated.

In the simple moving average rule a Buy signal is generated when the closing monthly price

[4]See `http://www.djaverages.com`.
[5]See `http://online.barrons.com`.

Market index	Return	Mean	Std	Skew	Kurt	AC_1
Standard and Poor Composite	CAR	0.61	5.51	0.30	12.22	**0.09** (0.00)
	DIV	0.33	0.14	1.16	6.77	**0.98** (0.00)
	TOR	0.93	5.52	0.37	12.48	**0.09** (0.00)
Dow Jones Industrial Average	CAR	0.57	5.36	0.02	10.67	**0.07** (0.03)
	DIV	0.34	0.13	1.10	5.22	**0.98** (0.00)
	TOR	0.90	5.34	0.06	10.72	**0.06** (0.05)
Long-term bonds	CAR	0.06	2.38	0.43	7.94	0.03 (0.28)
	DIV	0.43	0.23	1.02	3.48	**0.98** (0.00)
	TOR	0.49	2.40	0.62	8.14	0.04 (0.17)
Intermediate-term bonds	CAR	0.06	1.24	0.43	11.28	**0.12** (0.00)
	DIV	0.39	0.25	0.95	3.67	**0.98** (0.00)
	TOR	0.45	1.27	0.91	11.90	**0.15** (0.00)
Treasury bills	TOR	0.30	0.26	1.04	4.29	**0.99** (0.00)

Table 1: Descriptive statistics for the study data, including means, standard deviations, skewness, kurtosis, and first-order autocorrelations (denoted by AC_1). The variables CAR, DIV, and TOR represent the capital appreciation return, dividend yield, and total return respectively. The descriptive statistics are computed using monthly data. The means and standard deviations are given in percents. For each AC_1 we test the hypothesis $H_0 : AC_1 = 0$. The p-values are given in brackets. Bold text indicates values that are statistically significant at the 5% level.

is above a k-month simple moving average. Otherwise, if the closing monthly price is below a k-month simple moving average, a Sell signal is generated. More formally, let $(P_1, P_2, \ldots, P_T)$ be the observations of the monthly closing prices[6] of a stock/bond price index. A k-month simple moving average at month-end t is computed as

$$SMA_t(k) = \frac{1}{k}\sum_{j=0}^{k-1} P_{t-j}.$$

[6]It is worth emphasizing that in the computation of trading signals we use the prices not adjusted for dividends. In other words, we use the capital appreciation return to generate Buy and Sell signals. In contrast, Faber (2007) uses the total return whereas Moskowitz et al. (2012) use the total return in excess of the risk-free rate of return. However, this is highly non-standard in technical analysis. In particular, we have studied many handbooks on technical analysis of financial markets, beginning from the book by Gartley (1935), and in every handbook a technical indicator is supposed to be computed using the prices which are easily observable in the market, in contrast to, for example, dividend-adjusted prices. Therefore in the paper we stick to the standard computation of trading signals. We also used the total return and excess return for the computations of trading signal. Yet, our conclusions about the real-life performance of market timing strategies remain intact regardless of the type of return used to generate trading signals.

The trading signal for month $t+1$ is generated according to the following rule

$$\text{Signal}_{t+1} = \begin{cases} \text{Buy} & \text{if } P_t > SMA_t(k), \\ \text{Sell} & \text{if } P_t \leq SMA_t(k). \end{cases}$$

In the momentum rule a Buy signal is generated when a k-month momentum is positive. Otherwise, a Sell signal is generated. A k-month momentum at month-end t is computed as

$$MOM_t(k) = P_t - P_{t-k}.$$

The trading signal for month $t+1$ is generated according to the following rule

$$\text{Signal}_{t+1} = \begin{cases} \text{Buy} & \text{if } MOM_t(k) > 0, \\ \text{Sell} & \text{if } MOM_t(k) \leq 0. \end{cases}$$

Observe that the capital appreciation return (CAR) over the period $t-k$ to t is computed as

$$\text{CAR}_{t-k,t} = \frac{P_t - P_{t-k}}{P_{t-k}}.$$

Therefore the momentum rule prescribes investing in the stocks/bonds when the CAR over the last k months is positive.

Even though the two technical trading rules seem to be quite different at the first sight, there is a relationship between these two rules. Indeed, note that

$$SMA_t(k+1) - SMA_{t-1}(k+1) = \frac{P_t - P_{t-k}}{k+1}.$$

Therefore

$$\frac{MOM_t(k)}{k+1} = SMA_t(k+1) - SMA_{t-1}(k+1).$$

Thus, if $MOM_t(k) > 0$ then $SMA_t(k+1) > SMA_{t-1}(k+1)$. Putting it into words, the momentum rule prescribes investing in the stocks/bonds when the simple moving average over $k+1$ months increases.

Let $(R_1, R_2, \ldots, R_T)$ be the monthly total returns on a stock/bond price index, and let

$(RF_1, RF_2, \ldots, RF_T)$ be the monthly risk-free rates of return over the same sample period $[1, T]$. We suppose that buying and selling stocks/bonds is costly, whereas buying and selling Treasury bills is costless. Denoting by λ the one-way transaction costs, the return to the market timing strategy over month t is given by

$$r_t = \begin{cases} R_t & \text{if } (\text{Signal}_t = \text{Buy}) \text{ and } (\text{Signal}_{t-1} = \text{Buy}), \\ R_t - \lambda & \text{if } (\text{Signal}_t = \text{Buy}) \text{ and } (\text{Signal}_{t-1} = \text{Sell}), \\ RF_t & \text{if } (\text{Signal}_t = \text{Sell}) \text{ and } (\text{Signal}_{t-1} = \text{Sell}), \\ RF_t - \lambda & \text{if } (\text{Signal}_t = \text{Sell}) \text{ and } (\text{Signal}_{t-1} = \text{Buy}). \end{cases}$$

2.3 Amount of transaction costs

Since the goal of our study is to estimate the real-life performance of market timing strategies, we need to account for the fact that the rebalancing an active portfolio incurs transaction costs. Transaction costs in capital markets consist of the following three main components: half-size of the quoted bid-ask spread, brokerage fees (commissions), and market impact costs. In addition there are various taxes, delay costs, opportunity costs, etc. (see, for example, Freyre-Sanders, Guobuzaite, and Byrne (2004)). All investors face the same bid-ask spreads and market-impact costs for a trade of any given size and security at any given moment. In contrast, the commissions (on purchase and sale) are negotiated and depend on the annual volume of trading, as well as on the investor's other trading practices. In order to model realistic transaction costs one usually distinguishes between two classes of investors: institutional and individual (see, for example, Dermody and Prisman (1993)). Individual investors pay substantially larger commissions than institutional investors. Whereas institutional investors usually pay very low commissions of about 0.1% (or even less) of the volume of trade, individual investors pay commissions of about 0.5-1.5% (see, for example, Hudson, Dempsey, and Keasey (1996)).

Market impact costs are closely related to liquidity: a relatively big order exerts pressure on price. Market impact costs become a problem if an investor places an order to buy or sell a quantity of securities that is large relative to a market average daily security volume. Market impact costs are less significant with liquid securities. Liquidity refers to the ease with which a securities can be bought or sold without disturbing its price.

We assume that a market timing strategy is implemented by an institutional investor who pays very low commissions that can be neglected. In addition, we do not consider taxes. Finally, we assume that the buy and sell orders are not big so that we can disregard the market impact costs. That is, we consider the average bid-ask half-spread as the only determinant of the one-way transaction costs. Still, there are different estimates of average one-way transaction costs in the stock market. Whereas Berkowitz, Logue, and Noser (1988), Chan and Lakonishok (1993), and Knez and Ready (1996) estimate the average one-way transaction costs for institutional investors to be in the range of 0.23% to 0.25%, Bessembinder (2003) reports that the average one-way transaction costs amounts to 0.48% for stocks on the NYSE and 0.74% for stocks on the NASDAQ. Siegel (2002) also advocates for using 0.50% one-way transaction costs in order to evaluate the impact of realistic transaction costs on the performance of a market timing strategy. Therefore in our study we assume that the one-way transaction costs in the stock market amount to 0.50%, that is, $\lambda = 0.005$.

The government bonds are more liquid securities as compared to stocks and, therefore, the average bid-ask spread in bond trading is smaller than that in stock trading. Chakravarty and Sarkar (2003) and Edwards, Harris, and Piwowar (2007) estimate the average bid-ask half-spread in bond trading to be about 0.10%. Therefore in our study we assume that the one-way transaction costs in the bond market amount to 0.10%, that is, $\lambda = 0.001$.

2.4 Robust method of performance measurement in out-of-sample tests

The traditional out-of-sample performance measurement method is based on simulating the real-life trading where a trader has to make a choice of which trading rule to use (in our case, how many months k to use in the computation of moving averages at each given time) given the information about the past performances of different trading rules. Supposedly, such an out-of-sample simulation should remove the data-mining bias in performance measurement and allow one to find out whether a trader could really beat the market by following some trading rules. The traditional out-of-sample performance measurement procedure, described below, resembles the procedure employed by Lukac, Brorsen, and Irwin (1988) and Lukac and Brorsen (1990) for testing the profitability of some technical trading rules in commodity futures markets. To shorten the exposition, we outline the procedure for the $SMA(k)$ trading rule only. The out-of-sample simulation procedure and performance measurement for the $MOM(k)$

trading rule is implemented identically to that for the $SMA(k)$ trading rule.

The set of trading rules for the simple moving average market timing strategy is denoted by $SMA(k)$, $k \in [2, 24]$. The $SMA(k)$ rule prescribes investing in the stocks/bonds when the monthly price is above a k-month simple moving average, and moving to the risk-free asset when the monthly price is below a k-month simple moving average. Which k to use at each given month is chosen by finding the best trading rule in the past. The choice of k can be done using either an expanding-window or rolling-window estimation scheme. Specifically, out-of-sample simulation of a market timing strategy using an expanding-window estimation of k is performed as follows. First of all, the historical data are segmented into in-sample and out-of-sample subsets. The split point between the initial in-sample and out-of-sample subsets is denoted by τ, $1 < \tau < T$, where T is the number of months in the total sample. The initial in-sample period of $[1, \tau]$ is used to complete the procedure of selecting the best trading rule given some optimization criterion $O(r_1, r_2, \ldots, r_\tau)$ defined over the returns of the market timing strategy up to month τ. That is, the choice of the optimal k_τ^* is given by

$$\max_{k \in [2,24]} O(r_1, r_2, \ldots, r_\tau).$$

Subsequently, the trading signal for month $\tau + 1$ is determined using the $SMA(k_\tau^*)$ rule. One then expands the in-sample period by one month, performs the best trading rule selection procedure once again using the new in-sample period of $[1, \tau + 1]$, and determines the trading signal for month $\tau + 2$ using the $SMA(k_{\tau+1}^*)$ rule. One repeats this procedure, pushing the endpoint of the in-sample period ahead by one month with each iteration of this process, until the trading signal for the last month T is determined. In the end, the performance of the market timing strategy is measured using the returns over the out-of-sample period, $(r_{\tau+1}, r_{\tau+2}, \ldots, r_T)$.

In the rolling-window estimation scheme the choice of k is done using the most recent n observations. In this case the choice of the optimal k_s^* (for $\tau \leq s < T - 1$) is given by

$$\max_{k \in [2,24]} O(r_{s-n+1}, r_{s-n+2}, \ldots, r_s).$$

The rolling-window estimation scheme is used when parameter instability is suspected. For our

purpose it is natural to employ the expanding window estimation scheme. This is because in the literature on market timing with the SMA and MOM rules one supposes, either explicitly or implicitly, that some specific k is optimal regardless of the time period or asset class.

Now we turn to focusing on two serious deficiencies of the traditional out-of-sample performance measurement procedure described above. Firstly, the estimate for the out-of-sample performance is based on only one historical realization of the underlying data series. Thus, the estimate is not especially precise. Secondly and most importantly, the decision about how to divide the data between the in-sample and out-of-sample subsets is arbitrary. This raises the question that the out-of-sample performance may depend crucially on the choice of a split point. If this is the case, there are two types of concerns. The first concern is that different choices of the split point may lead to different empirical results. For example, depending on the choice of a split point the out-of-sample performance of a market timing rule might by either superior or inferior as compared to that of the passive counterpart. The second concern is a "data-mining" issue when multiple split points might have been considered and the reported out-of-sample performance could be that which favors most a market timing rule.

In order to overcome the deficiencies of the traditional out-of-sample performance measurement procedure, we propose and implement a robust method of performance measurement in out-of-sample tests. Our method is based on simulating many alternative historical realizations of the underlying data series. After each simulation, the traditional out-of-sample performance measurement procedure is repeated. In the end, the average out-of-sample performance is computed. Our experiments show that the the average out-of-sample performance is insensitive to how we apportion data between the in-sample and out-of-sample subsets. It should be emphasized, however, that in order to employ this method to produce a reliable estimate of performance, one needs to use a historical sample of data which represents a wide range of market conditions and a simulation method which preserves all relevant statistical features of the historical data series.

2.5 Optimization criteria

There is a big uncertainty about what optimization criterion to use in the determination of the best trading rule using the past data. To limit the choice of optimization criteria, we consider an investor who decides whether to follow the passive buy-and-hold strategy or to follow the

active market timing strategy. Since the two strategies are supposed to be mutually exclusive, it is natural to employ a reward-to-risk performance measure as the optimization criterion. That is, our investor chooses the value of k which maximizes some portfolio performance measure in a back test, that is, using the past (in-sample) data.

The most widely recognized reward-to-risk measure is the Sharpe ratio. Thus, the Sharpe ratio represents the natural optimization criterion to find the best trading rule. The Sharpe ratio uses the mean excess returns as a measure of reward, and the standard deviation of excess returns as a measure of risk. However, the Sharpe ratio has often been criticized on the grounds that the standard deviation seems to be an inappropriate measure of risk. In particular, the standard deviation penalizes similarly both the downside risk and upside return potential. There have been proposed many alternatives to the Sharpe ratio. In most of these alternative reward-to-risk ratios the standard deviation is replaced by another risk measure that takes into account only the downside risk. We decided, in addition to the Sharpe ratio, to employ a set of the most popular alternatives to the Sharpe ratio. The alternative performance measures used in our study include the Omega ratio, Sortino ratio, Kappa 3 ratio (where the lower partial moments of order 1, 2, and 3 respectively are used as risk measures), the Upside potential ratio (where both the reward and risk measures are based on partial moments of distribution), Reward-to-VaR ratio, Reward-to-CVaR ratio (where the risk is measured by either Value-at-Risk or Conditional Value-at-Risk), and the Calmar ratio (where the risk is measured by the maximum drawdown).[7]

Yet it turns out that *regardless of the performance measure used, the comparative performance of the passive buy-and-hold strategy and the active market timing strategy remains virtually the same.* This result seems to be rather puzzling, because alternative performance measures employ quite different approaches to risk and performance measurement. However this agrees very well with the findings reported by Eling and Schuhmacher (2007) and Eling (2008). In particular, these authors computed the rank correlations between the rankings of risky portfolios according to some alternative performance measures, including the Sharpe ratio, and found that the rankings are extremely positively correlated. The authors concluded that the choice of performance measure is irrelevant, since, judging by the values of rank correlations, all measures give virtually identical rankings. Since the comparative performance

[7]See, for example, Eling and Schuhmacher (2007) or Eling (2008) a brief overview of all these measures.

of the passive and active strategies virtually does not depend on the choice of performance measure, in order to save the space we report only the results based on the use of the Sharpe ratio as the optimization criterion.

Note, however, that the ultimate investor's goal is to maximize the utility of wealth at the end of the investment horizon. In order to do this, the investor has to solve not one, but two different optimization problems: the optimal portfolio choice problem and the optimal capital allocation problem. The first one consists in determining the optimal risky portfolio to invest in. This problem is, in principle, easily resolved by choosing the portfolio with the highest performance measure. The other optimization problem is how to allocate money optimally between the risky portfolio and the Treasury bills. The latter problem has no unique resolution, because the optimal capital allocation depends on the risk preferences of a particular investor. As a result, it is difficult in practice to determine the optimal capital allocation for each specific investor. This task is further complicated due to existence of market imperfections, for example, borrowing and short sale restrictions. Therefore in practice it might be the case that a combination of an inferior risky portfolio with the optimal capital allocation can result in a much higher utility of final wealth than a combination of the optimal risky portfolio with an inferior capital allocation. In order to shed light on this issue, we decided to report not only the Sharpe ratios of the buy-and-hold and market timing strategies, but also the final wealth from investing initially \$100 in either of the two distinct strategies.

It is worth noting that the final wealth criterion (from the initial investing of the same amount of money into different strategies) is consistent with the *capital growth theory* (see, for example, Hakansson and Ziemba (1995), for a good review of the theory) and the Kelly criterion (see Kelly (1956)). In brief, the growth-optimal portfolio is the portfolio which maximizes the long-run geometric return. It can be shown that the growth-optimal strategy is consistent with a logarithmic utility of final wealth. Since the geometric return is approximately given by $g \approx \mu - \frac{1}{2}\sigma^2$, where μ is the arithmetic return and σ is the standard deviation of return, the maximization of the geometric return can naturally be interpreted as the maximization of a specific utility function.

2.6 The choice of a proper simulation method

In order to simulate many alternative historical realizations of the underlying data series that contain all relevant statistical features of the original data series, we rely on bootstrap methods. A bootstrap is a computer-intensive method of estimation of the sampling distribution of a test statistics by resampling the original (i.e., historical) data. The problem is that there are many different bootstrap methods (for a brief review see Mammen and Nandi (2012)). How to choose the proper one? We do not take for granted that some particular bootstrap method is suitable for our purpose. Suppose there are N different bootstrap methods to choose among, each bootstrap method is labelled as BM_i, $i \in [1, 2, \ldots, N]$. We follow the usual line of statistical reasoning and formulate the following **null hypothesis**:

> *BM_i is able to simulate alternative historical realizations of the underlying data series that preserve all relevant statistical properties of the original data series.*

Our test statistics is the difference in the Sharpe ratios, $\Delta = SR_A - SR_P$, where SR_A and SR_P are the Sharpe ratios of the active and passive strategies respectively. The estimation of the significance level of our test statistics is performed as follows. First, using the original data series, we simulate out-of-sample the returns to the market timing strategy. Then we compute the Sharpe ratios of the active and passive strategies and obtain the actual historical estimate for Δ. After that, using bootstrap method i we randomize the original series to get random resamples that could have been realized in the past. This is repeated 1,000 times, each time we simulate out-of-sample the returns to the market timing strategy using the same procedure as for the original series and obtain an estimate for Δ^*.[8] Finally, to estimate the significance level, we count how many times the computed value for Δ^* after bootstrapping happens to be above the value of the actual historical estimate for Δ. In other words, under the null hypothesis we compute the probability of obtaining an even greater difference in performances (between the market timing strategy and the passive benchmark) than the actual historical estimate. We reject the null hypothesis when the probability of observing a higher value of Δ^* than the actual historical estimate Δ falls below a conventional significance level (1% or 5%). This is because it is highly unlikely that the application of the trading rules to the simulated data series can produce the same performance difference as the actually observed one. Apparently,

[8] The superscripted asterisk is the typical symbol for denoting a bootstrapped observation.

the bootstrapped data series do not exhibit the relevant statistical properties existed in the original data series.

We entertain two different semi-parametric bootstrap methods. The first semi-parametric bootstrap method follows closely Nelson and Kim (1993), Rapach and Wohar (2005), and Goyal and Welch (2008). In this method we assume that the capital appreciation return series (denoted by CAR) are independent over time (i.e., not predictable), whereas both the dividend yield and the risk-free rate of return (denoted by DIV and RF respectively) follow the first-order autoregressive (AR(1)) process. Therefore the data generating process is assumed to be

$$\begin{aligned} CAR_t &= \mu + e_t, \\ DIV_t &= \alpha + \beta\, DIV_{t-1} + w_t, \\ RF_t &= \gamma + \theta\, RF_{t-1} + u_t. \end{aligned} \tag{1}$$

In this case all the data series are generated using the semi-parametric standard bootstrap method. Specifically, first of all, the parameter μ (the mean capital appreciation return) is estimated using the full sample of observations, and the random disturbances e_t are stored for resampling. Secondly, the parameters α, β, γ, and θ are estimated by OLS using the full sample of observations, and the residuals w_t and u_t stored for resampling. Afterwards, to generate, for example, a random resample for $(DIV_1^*, DIV_2^*, \ldots, DIV_T^*)$, we pick up an initial observation DIV_1^* from the actual data at random. Then a series is generated using the AR(1) model and by drawing w_t^* with replacement from the residuals. The simulated total return is computed as the sum of the simulated capital appreciation return and the dividend yield, that is, $R_t^* = CAR_t^* + DIV_t^*$. All the disturbance terms e_t, w_t, and u_t are assumed to be correlated and identically distributed over time random variable. In order to retain the historical correlations among the disturbance terms,[9] we resample the entire vector of disturbance terms instead of resampling each term separately. As a result, the bootstrapped observation of the vector $\{e_s^*, w_s^*, u_s^*\}$ for a time s represents the actual historical observation of the vector $\{e_t, w_t, u_t\}$ for a time t.

It is worth emphasizing that using the first bootstrap method we test whether the observed superior performance of the market timing strategy can be explained by pure luck when the capital appreciation return, used to compute the trading signals, is an i.i.d. random variable

[9]For example, the capital appreciation return and the dividend yield are usually negatively correlated.

(in this case the security prices, not adjusted for cash dividends, follow a random walk). Yet, as a matter of fact, the real-life capital appreciation return is often a persistent random variable. The descriptive statistics presented in Table 1 advocate that for 3 out of 4 market indices used in our study the capital appreciation return exhibits a positive and statistically significant autocorrelation. In addition, there might be some other types of serial dependence in real-life data series. In order to retain the serial dependence in the capital appreciation return and the other data series, we use the second bootstrap method. Our second bootstrap method employs the stationary block-bootstrap method of Politis and Romano (1994). This stationary block-bootstrap method seems to be the preferred method of incorporating real-life time-series dependencies in data while performing the bootstrap (see, for example, Sullivan et al. (1999) and White (2000)). Note that using the second bootstrap method we test whether the empirical success of the market timing strategy can be explained by serial dependence in data.

Our second bootstrap method represents a refinement of the first bootstrap method where the data generating process is assumed to be

$$\begin{aligned} CAR_t &= \mu + \phi\, CAR_{t-1} + e_t, \\ DIV_t &= \alpha + \beta\, DIV_{t-1} + w_t, \\ RF_t &= \gamma + \theta\, RF_{t-1} + u_t. \end{aligned} \tag{2}$$

In essence, it is a semi-parametric stationary block-bootstrap method. It has two major differences as compared to the first method. First of all, here it is assumed that the capital appreciation return also follows the AR(1) process. Secondly, whereas in the first bootstrap method each of the disturbance terms is supposed to be serially independent, in the second bootstrap method we explicitly assume that there might be some serial dependence. In order to incorporate the historical serial dependence in data series and retain the cross-sectional correlation between the disturbance terms, following the method of Politis and Romano (1994) the disturbance terms are resampled using blocks of data instead of individual observations. To simulate a block, first a random index i is generated from the discrete uniform distribution on $\{1, \ldots, T\}$. Then the block length l is generated from a geometric distribution with probability p. Thus, the average block length equals $1/p$. We use $1/p = 16$ which means that the average

block length equals to 16 months.[10] Consequently, the blocks of simulated disturbances for e^*, w^*, and u^* are given by $\{e_i, e_{i+1} \ldots, e_{i+l-1}\}$, $\{w_i, w_{i+1} \ldots, w_{i+l-1}\}$, and $\{u_i, u_{i+1} \ldots, u_{i+l-1}\}$. Since a generated block length is not limited from above, $l \in [1, \infty)$, and, for example, the block for e^* can begin with observation e_T, the stationary bootstrap method "wraps" the data around in a "circle", so that e_1 follows e_T and so on.

3 Empirical Results

3.1 Best trading rules in a back test

First of all, for each stock/bond market index we want to determine the best trading rules in a back test over our total sample period. In order to do this, for each trading rule we simulate the market timing strategy over 1926 to 2012 using different (fixed) lengths $k \in [2, 24]$, and find the value of k which produces the best performance. As in Faber (2007) and many other papers, when we examine the performance of the best trading rule in a back test, we do not take into account the transaction costs in order to show the best possible performance. The results are reported in Table 2. Observe that the $SMA(10)$ rule, advocated by Faber and many others, is the best trading rule when the passive benchmark is the Standard and Poor's Composite index. For the Dow Jones Industrial Average index the optimal trading rule in a back test is $SMA(13)$ (close to the 52-week moving average rule). For the bond market indices the optimal k in the $SMA(k)$ rule is lower than 10. For the $MOM(k)$ rule, the optimal k is usually lower than 12 (as advocated by Moskowitz et al. (2012) and some others), the only exception is the index of long-term bonds.

Even though we find that neither the $SMA(10)$ rule nor the $MOM(12)$ rule is the optimal rule for all the indices, this result is determined by our specific choices of the sample period and the type of return used to compute the trading signals. The choice of a sample period varies from paper to paper; the most popular choices are either 1926-present, 1970-present, or 1995-present (the reasons for these popular choices are discussed in the next subsection). Similarly, the type of return used to compute the trading signals varies from paper to paper;

[10] Sullivan et al. (1999) use $1/p = 10$ and report that the simulation results are quite insensitive to the choice of the average block size. Our choice is motivated by the fact that in our experiments the maximum k in the best trading rules in a back test amounts to 15. This suggests that there might be time-series dependencies in underlying data series over time horizons up to 16 months.

it is either the capital appreciation return, the total return, or the excess return. If we use, as in the paper by Moskowitz et al. (2012), the period 1970-present and the excess return in the computation of the trading signal of the time-series momentum rule, then the optimal k would be either 11 or 12 for practically all market indices.

Market index	**SMA(k) rule**	**MOM(k) rule**
Standard and Poor Composite	$SMA(10)$	$MOM(5)$
Dow Jones Industrial Average	$SMA(13)$	$MOM(7)$
Long-term bonds	$SMA(6)$	$MOM(12)$
Intermediate-term bonds	$SMA(5)$	$MOM(4)$

Table 2: The best trading rules in a back test for the simple moving average and momentum market timing strategies over the period 1926 to 2012.

3.2 Out-of-sample performance dependence on the choice of a split point

The main goal of this subsection is to demonstrate that the out-of-sample performance of a trading rule is sensitive to the choice of a split point between the initial in-sample and out-of-sample subsets. What is more crucial is that depending on the choice of a split point the out-of-sample performance of a trading rule might by either superior or inferior as compared to that of the passive counterpart. The passive benchmark in our demonstration is the Standard and Poor's Composite stock price index. For different historical periods, Table 3 reports the performance of the buy-and-hold strategy, the performance of the best trading rules in a back test and the absence of transaction costs, as well as the performance of the trading rules in the out-of-sample tests and the presence of realistic transaction costs. Before turning to the discussion of performance dependence on the split point, we would like to draw the reader's attention to the fact that in every historical period the best trading rule in a back test outperforms the passive benchmark with a solid margin. This demonstrates that, when many trading rules are back tested, one can virtually always find a rule that performs better than the passive benchmark.

Consider in details the performance of the market timing strategies versus the performance of the passive strategy when the split point between the initial in-sample and out-of-sample subsets is chosen to be December 1929. In this case the initial in-sample period is January 1926 to December 1929, whereas the initial out-of-sample period is January 1930 to December 2012.

Period	Buy&Hold	SMA(k) rule		MOM(k) rule	
		BBT	OOS	BBT	OOS
1930 to 2012	0.11	0.16	0.12	0.16	0.12
1932 to 2012	0.13	0.16	0.12	0.17	0.13
1973 to 2012	0.10	0.15	0.11	0.14	0.10
1975 to 2012	0.13	0.16	0.13	0.16	0.12
1995 to 2008	0.08	0.25	0.21	0.23	0.19
2009 to 2012	0.25	0.38	0.09	0.35	0.12

Table 3: The table reports the Sharpe ratios of the buy-and-hold strategy (Buy&Hold), the k-month simple moving average rule (SMA(k) rule), and the k-month time-series momentum rule (MOM(k) rule) for different historical periods. For each trading rule reported are the Sharpe ratio of the best trading rule in a back test and the absence of transaction costs (BBT), as well as the Sharpe ratio of the trading rule in the out-of-sample tests and the presence of realistic transaction costs (OOS). The Sharpe ratios are computed using monthly data.

With this choice, the Sharpe ratio of either of the best trading rules in a back test amounts to 0.16 which is about 50% higher than the Sharpe ratio of the buy-and-hold strategy. In the out-of-sample tests, the Sharpe ratio of both the trading rules amounts to 0.12 which is about 10% higher than the Sharpe ratio of the buy-and-hold strategy. Even though the real-life performance[11] of the trading rules turns out to be much poorer than the performance of the best trading rules in a back test, still, judging by the Sharpe ratio criterion, we have to conclude that either of the market timing strategies delivers a better performance than that of the buy-and-hold strategy. Note, however, that if the split point is chosen to be December 1931 (which seems to be a relatively negligible relocation of the original split point) then we arrive at the opposite conclusion: the real-life performance of the market timing is either worse than or equal to the performance of the buy-and-hold strategy. Similarly, our conclusion on whether the market timing rules outperform a buy-and-hold strategy changes depending on whether the split point is chosen to be December 1972 or December 1974. In the first case we conclude that the real-life performance of the trading rules is either better than or equal to the performance of the buy-and-hold strategy, whereas in the second case we conclude that the real-life performance of trading rules is either worse than or equal to the performance of the buy-and-hold strategy.

So why the conclusion on whether the market timing rules outperform a buy-and-hold strategy is so sensitive to the choice of a split point between the initial in-sample and out-of-

[11] Throughout the paper by a "real-life performance" we mean the out-of-sample performance in the presence of realistic transaction costs.

sample subsets? Our analysis reveals that the superior performance of market timing strategies is confined to relatively short historical episodes. In particular, over our total sample 1926 to 2012 there have been only four relatively short historical periods that contributed most to the superior performance of the market timing strategies. Specifically, they are the periods of severe bear markets of 1930-31, 1973-74, 2001-02, and 2007-08. As a result, the performance of market timing strategies, relative to the performance of the passive benchmark, depends largely on whether the split point is located before or after a major stock market downturn. Hence, the proponents of market timing might be tempted to manipulate the out-of-sample performance. In particular, the out-of-sample performance of market timing rules can be improved by allocating the split point right before a major stock market downturn. By contrast, the performance of market timing rules can be deteriorated by allocating the split point right after a major stock market downturn. Given this knowledge, it is not surprising that the benefits of market timing are most obvious when one focuses exclusively on the end of the sample where the two major stock market downturns are located. For example, if the split point is chosen to be December 1994, then the real-life performance of both the market timing strategies over 1995-2008 turns out to be substantially better than that of the buy-and-hold strategy. Yet, focusing on this specific historical period, the advantages from market timing prove to be greatly overstated. As a counter-example, which demonstrates that the superior performance of market timing rules is mainly confined to a few particular historical episodes, over the period 2009 to 2012 the real-life performance of both the market timing strategies was much worse than the performance of the buy-and-hold strategy.

3.3 The real-life historical performance of trading rules over 1930 to 2012

In the previous subsection we demonstrated that the out-of-sample performance of a trading rule, with an ad-hoc choice of a split point, does not provide a trustworthy estimate of it's real-life performance. However, in this subsection we examine the out-of-sample performance of trading rules over a specific historical period 1930 to 2012. The goal of this exercise is threefold. Firstly, we want to systematically compare the performance of the best trading rules in a back test (that is, the performance that is usually reported) and the real-life performance. Secondly, this exercise provides some useful insights into the properties of market timing strategies. Thirdly, since the start of this particular historical period coincides with the beginning of the

most severe stock market downturn, the out-of-sample performance of the stock market timing strategies over this period contains a definite upward bias as compared with the "true" long-term performance. This insures that the choice of a proper bootstrap method is made under strict statistical scrutiny when we use this specific historical period.

Table 4 reports the performance of the best trading rules in a back test (exclusive of transaction costs) and the out-of-sample performance of the trading rules (inclusive of transaction costs) for each benchmark index. For the purpose of comparison, this table also reports the performance of the passive buy-and-hold strategies over the same period. Consider first the performance of the best trading rules in a back test. As compared to that of the buy-and-hold strategy, it is clearly impressive. Regardless of the passive benchmark and trading rule, all the market timing strategies optimized in a back test show better risk-adjusted performance as compared with the performance of the corresponding buy-and-hold strategy. In particular, the market timing strategy, optimized in a back test, has only a bit lower mean returns and a substantially lower risk than the buy-and-hold strategy. For example, if the S&P Composite index is used as the passive benchmark, then the Sharpe ratio of the best trading rules in a back test is about 50% higher than the Sharpe ratio of the buy-and-hold strategy. At the same time, the capital growth provided by the market timing strategy is also about 50% higher than that of the buy-and-hold strategy.

By contrast, the real-life performance of the market timing strategies is substantially worse than that of the best trading rules in a back test. There are only three instances out of eight when the real-life market timing strategy shows a better risk-adjusted performance as compared with the passive benchmark. In particular, both the market timing rules have a better risk-return tradeoff than the buy-and-hold strategy when the S&P Composite index is used as the passive benchmark. In addition, the active timing strategy based on the $SMA(k)$ rule has a better risk-return tradeoff than the buy-and-hold strategy when the intermediate-term government bond index is used as the passive benchmark. In all instances the capital growth provided by the real-life market timing strategy is substantially lower than that of the buy-and-hold strategy.

Specifically, if the S&P Composite index is used as the passive benchmark, the real-life market timing strategy has a Sharpe ratio which is only about 10% higher than that of the buy-and-hold strategy. Even though judging by the Sharpe ratio the real-life market timing

Panel A: Standard and Poor's Composite

	Buy&Hold	SMA(k) rule		MOM(k) rule	
		BBT	OOS	BBT	OOS
Mean returns	0.90%	0.83%	0.71%	0.85%	0.75%
Standard deviation	5.51%	3.32%	3.44%	3.37%	3.71%
Skewness	0.45	-0.50	-0.65	-0.30	0.81
Best month return	42.91%	16.35%	15.85%	16.87%	42.41%
Worst month return	-29.42%	-21.54%	-23.52%	-21.54%	-23.52%
Maximum drawdown	-79.18%	-33.69%	-61.68%	-32.67%	-48.57%
Sharpe ratio	0.109	0.160	0.120	0.164	0.122
Growth of $100	166,155	221,436	64,345	269,275	88,871

Panel B: Dow Jones Industrial Average

	Buy&Hold	SMA(k) rule		MOM(k) rule	
		BBT	OOS	BBT	OOS
Mean returns	0.88%	0.76%	0.65%	0.77%	0.60%
Standard deviation	5.27%	3.28%	3.33%	3.28%	3.34%
Skewness	0.12	-0.54	-0.57	-0.53	-0.68
Best month return	40.46%	14.76%	14.76%	14.76%	14.13%
Worst month return	-29.80%	-22.91%	-22.91%	-22.91%	-22.91%
Maximum drawdown	-81.49%	-41.07%	-51.11%	-24.19%	-47.10%
Sharpe ratio	0.110	0.141	0.106	0.145	0.089
Growth of $100	152,931	113,246	36,960	127,636	21,605

Panel C: Long-Term Government Bonds

	Buy&Hold	SMA(k) rule		MOM(k) rule	
		BBT	OOS	BBT	OOS
Mean returns	0.50%	0.46%	0.40%	0.48%	0.40%
Standard deviation	2.45%	1.80%	1.76%	1.81%	1.76%
Skewness	0.61	0.36	0.26	0.74	0.31
Best month return	15.23%	11.45%	11.45%	14.43%	11.45%
Worst month return	-11.24%	-11.24%	-11.24%	-11.24%	-11.24%
Maximum drawdown	-20.76%	-21.98%	-22.64%	-14.35%	-15.21%
Sharpe ratio	0.083	0.093	0.063	0.105	0.063
Growth of $100	9,787	7,684	4,492	9,522	4,470

Panel D: Intermediate-Term Government Bonds

	Buy&Hold	SMA(k) rule		MOM(k) rule	
		BBT	OOS	BBT	OOS
Mean returns	0.45%	0.45%	0.41%	0.43%	0.38%
Standard deviation	1.30%	0.93%	0.92%	0.91%	0.92%
Skewness	0.88	0.90	0.76	0.71	-0.21
Best month return	11.98%	6.24%	6.14%	5.31%	5.31%
Worst month return	-6.41%	-3.87%	-3.87%	-3.87%	-6.41%
Maximum drawdown	-8.89%	-5.62%	-6.59%	-5.62%	-9.02%
Sharpe ratio	0.125	0.181	0.138	0.155	0.092
Growth of $100	7,823	8,270	5,642	6,420	3,849

Table 4: The performance of the passive buy-and-hold strategy versus the performance of the best trading rules in a back test (BBT) and out-of-sample performance (OOS) of the trading rules over the period January 1930 to December 2012. Panel A reports the performances when the passive benchmark is the S&P Composite index. Panel B reports the performances when the passive benchmark is the DJIA index. Panel C reports the performances when the passive benchmark is the long-term government bond index. Panel D reports the performances when the passive benchmark is the intermediate-term government bond index. The descriptive statistics and performance measures are computed using monthly data.

strategy still outperforms the buy-and-hold strategy, the capital growth provided by the real-life market timing strategy happens to be from two to three times as less as that of the buy-and-hold strategy. Put it differently, the real-life capital growth provided by the market timing strategy is about four times as less as the capital growth provided by the best trading rule in a back test. On the other hand, if the DJIA index is used as the passive benchmark, then the real-life performance of the market timing strategies is worse than the performance of the buy-and-hold strategy. In this case, even the back-test optimized market timing strategy provides a capital growth which is less than that of the buy-and-hold strategy. In real-life, the capital growth provided by the market timing strategy happens to be from five to seven times as less as that of the buy-and-hold strategy. When the passive benchmark is a bond market index, the capital growth provided by the real-life market timing strategy is about two times as less as that of the buy-and-hold strategy.

Also observe that, when the passive benchmark is a stock market index, whereas the returns of the buy-and-hold strategy are positively skewed, the returns of both the best trading rules in a back test and the real-life market timing strategies are usually negatively skewed. This means that the return distribution of the buy-and-hold strategy has a fat right tail (higher variation on gains), while the return distribution of the market timing strategies has a fat left tail (higher variation on losses). Somewhat similar conclusion can be reached by comparing the best and worth month returns of the buy-and-hold strategy and market timing strategies: whereas the worst month returns are more or less the same, the best month return of the passive buy-and-hold strategy is usually substantially higher than that of the market timing strategies. That is, as a rule the market timing strategy lets the "big downward mover" months pass through, but plainly misses the "big upward mover" months.

3.4 Tests of the bootstrap simulation methods

All in all, in our study we use 4 passive benchmark indices and test 2 trading rules on each benchmark; the total number of tested active strategies amounts to 8. We found that 3 of them outperform the passive benchmark on the risk-adjusted basis in out-of-sample tests over the period 1930 to 2012. Given the actual historical difference in the performances between the market timing and passive strategies, we implement the tests of the bootstrap simulation methods described in Section 2.6. Such a test consists in generating many possible historical

realizations of the underlying data series over the 87-year period 1926 to 2012 and then estimating the out-of-sample performance of market timing strategies over the period 1930 to 2012. The goal is to find a bootstrap method which is able to simulate alternative historical realizations of the underlying data series that preserve all relevant statistical features of the original data series. The results of the tests are reported in Table 5.

We remind the reader that we test two particular bootstrap methods. In the first bootstrap method the security prices, not adjusted for cash dividends, are assumed to follow a random walk. In this case there is no serial dependence in data series besides the normal persistence in the dividend yield and the risk-free rate of return. In essence, when we use the first bootstrap method, we test whether the superior performance of the market timing strategies can be explained by pure luck when the stock/bond prices are not predictable. The p-value of this test in Table 5 is denoted by $\mathbf{pval}_1$. As our results suggest, all the observed superior performances *cannot* be explained by pure luck. For instance, when the prices of the S&P Composite index are not predictable, the probability of observing a greater discrepancy in performances (than the actual one) amounts to 0.03 for either of the trading rules. That is, on average in only 3 out of 100 simulations the difference in the Sharpe ratios of the market timing and buy-and-hold strategies happens to be greater than the actual historical difference. Since this probability is rather small, we reject the hypothesis that the first bootstrap method is able to simulate alternative historical realizations of the underlying data series.

In the second bootstrap method we retain the historical serial dependencies and cross-correlations in all data series, including the security prices not adjusted for cash dividends. Therefore, when we use the second bootstrap method, we test whether the superior performance can be explained by the statistical properties of the underlying data. The p-value of this test in Table 5 is denoted by $\mathbf{pval}_2$. Our results suggest that all the observed superior performances *can* be explained by serial dependence in data. For instance, if we retain the historical serial dependencies and cross-correlations in the data series when the S&P Composite index is used as the passive benchmark, the probability of observing an even greater difference in performances amounts to 0.38 for the $SMA(k)$ trading rule. That is, in this case on average in 38 out of 100 simulations the difference in the Sharpe ratios of the market timing and buy-and-hold strategies happens to be greater than the actual historical difference. Since this probability is rather high, it is logically feasible to believe that the second bootstrap method is able to

simulate alternative historical realizations of the underlying data series. In other words, the observed superior performance is not surprising given the underlying data generating process.

	SMA(k) rule		MOM(k) rule	
Market index	**pval$_1$**	**pval$_2$**	**pval$_1$**	**pval$_2$**
Standard and Poor Composite	**0.03**	0.38	**0.03**	0.27
Dow Jones Industrial Average	0.16	0.37	0.28	0.67
Long-term bonds	0.29	0.55	0.32	0.54
Intermediate-term bonds	**0.01**	0.41	0.27	0.86

Table 5: The p-values of the test of the hypothesis that a particular bootstrap method (1 or 2) is able to simulate alternative historical realizations of the underlying data series that preserve all relevant statistical properties of the original data series. The values pval$_1$ and pval$_2$ are computed using the bootstrap method 1 and 2 respectively. The first bootstrap method imposes no predictability in the stock/bond index prices not adjusted for dividends (these prices are used to compute the trading signals). The second bootstrap method retains the historical serial dependencies and cross-correlations in the underlying data series. Bold text indicate values that are statistically significant at the 5% level (when we can reject the null hypothesis).

3.5 Real-life performance of market timing strategies

The results reported in the preceding subsection suggest that the empirical success of the considered market timing strategies can be fully explained by serial dependencies in the underlying data series. Therefore the second bootstrap simulation method, which incorporates serial dependence in data, allows us to obtain a robust estimate of the real-life performance of a market timing strategy. In particular, using the second bootstrap method we randomize the original series to get random resamples that could have been realized over the course of the period from January 1926 to December 2012. Then we simulate out-of-sample the returns to the market timing strategy using the traditional procedure described in Section 2.4 and evaluate its performance. This is repeated 1,000 times and in the end the average out-of-sample performance is computed. We varied the length of the initial in-sample subset from 4 to 50 years, and these experiments showed that the average out-of-sample performance is quite insensitive to how we apportion data between the in-sample and out-of-sample subsets. This suggests, among other things, that even using a rather long price history to find the best trading rule in the past, one cannot improve the expected performance of the rule in the future.

In order to provide more details on the performance of a market timing strategy, after

each simulation we compute three different measures of performance: relative difference in the mean returns, relative difference in the standard deviations of returns, and relative difference in the Sharpe ratios. All the relative differences are computed with respect to the parameters of the buy-and-hold strategy. In particular, after each resampling of historical data series we simulate technical trading and obtain one alternative historical realization of the returns to the buy-and-hold strategy, R_t^*, the market timing strategy, r_t^*, and the risk-free rate of return, RF_t^*. This allows us to compute

$$\delta_\mu^* = \frac{E[r_t^*] - E[R_t^*]}{E[R_t^*]}, \quad \delta_\sigma^* = \frac{Std[r_t^*] - Std[R_t^*]}{Std[R_t^*]}, \text{ and } \delta_{SR}^* = \frac{SR_A^* - SR_P^*}{SR_P^*}.$$

After all the simulations are completed, we compute the averages of the relative performance measures. As a result, $\widehat{\delta}_\mu^*$ allows us to estimate the relative difference between the mean returns of the market timing strategy and the buy-and-hold strategy; $\widehat{\delta}_\sigma^*$ allows us to estimate the relative difference between the standard deviations of returns of the market timing strategy and the buy-and-hold strategy; $\widehat{\delta}_{SR}^*$ allows us to estimate the relative difference between the Sharpe ratios of the market timing strategy and the buy-and-hold strategy.

The results of our robust method of performance measurement in out-of-sample tests, for each market index and trading rule, are reported in Table 6. For instance, consider the case when the S&P Composite index is used as the passive benchmark and the market timing strategy is based on the $SMA(k)$ trading rule. In this case the market timing strategy has mean returns and standard deviation of returns that are on average respectively 20% and 32% lower than the corresponding parameters of the passive benchmark. In addition, on average the Sharpe ratio of the $SMA(k)$ rule is 7% higher than the Sharpe ratio of the S&P Composite index. That is, we find that the $SMA(k)$ rule indeed outperforms the buy-and-hold strategy when the S&P Composite index is used as a passive benchmark. Yet the real-life degree of outperformance is much lower than it is usually reported (7% versus 50% to 100% increase in the Sharpe ratio). The results of our method of performance measurement suggest that only 2 out of 8 tested active strategies are able to produce a slightly better performance than that of the passive counterparts. In both cases it is the $SMA(k)$ rule, and the passive benchmark is either the S&P Composite index or the intermediate-term government bond index. It is somewhat surprising that according to our estimates the performance of the

popular $MOM(k)$ rule is never higher than that of the passive counterpart: in the best case its average risk-adjusted performance is similar to that of the passive benchmark. We believe that the explanation for this result lies in the fact that the $MOM(k)$ trading rule generates more false signals as compared to the $SMA(k)$ rule (since the former operates with lower values of k than the latter). As a result, the investors who use the $MOM(k)$ rule are more whipsawed than the investors who use the $SMA(k)$ rule.

	SMA(k) rule			**MOM(k) rule**		
Market index	**Mean**	**Std**	**Sharpe**	**Mean**	**Std**	**Sharpe**
Standard and Poor Composite	-20	-32	+7	-21	-30	+0
Dow Jones Industrial Average	-26	-30	-11	-24	-29	-8
Long-term bonds	-19	-28	-17	-18	-28	-21
Intermediate-term bonds	-9	-28	+8	-12	-29	-6

Table 6: The relative differences in the expected performances of the market timing strategy and the corresponding buy-and-hold strategy. These relative differences are computed using the robust method of performance measurement. All the relative differences are calculated with respect to the parameters of the buy-and-hold strategy. **Mean** denotes the relative difference in the mean returns, **Std** and **Sharpe** denote the relative differences in the standard deviations of returns and the Sharpe ratios respectively. The differences are computed using monthly data and reported in percents.

4 Discussions

4.1 Revisiting the myths about the superior performance of market timing

The proponents of the market timing with moving averages and momentum rules often advocate that such a strategy allows investors both to reduce risk and enhance returns. We have to acknowledge that there is indeed a small chance of this happening. For instance, if the S&P Composite index is used as the passive benchmark, there have been 4 relatively short periods of severe market downturns when the real-life market timing strategy provided higher returns with lower risk than the buy-and-hold strategy. It happened that 2 out of 4 such periods occurred in the decade of 2000s. As a result, this decade was an incredibly successful decade for the market timing strategies. Yet this is not a typical performance of the real-life market timing strategy. Our findings reveal that over a long run the market timing strategy is indeed less risky, but the reduction of risk *always* comes at the expense of reduction of returns.

In particular, our results suggest that the standard deviation of returns of the market

timing strategy is on average about 30% less than that of the buy-and-hold strategy regardless of the choice of the passive benchmark. That is, the market timing strategy is expected to be about 30% less risky; the reduction in risk comes from the fact that about 30% of time the money is allocated risk-free. By contrast, the reduction in mean returns depends on the choice of the passive benchmark. Specifically, depending on the passive benchmark, the market timing strategy delivers from 9% to 26% lower mean returns as compared to that of the buy-and-hold strategy. All in all, our estimates of the real-life performance of the market timing strategies suggest that over a long run one can expect that a market timing strategy delivers only marginally better risk-adjusted returns than the buy-and-hold strategy, yet with a substantially lower long-term capital growth.

In addition to the inferior capital growth, when the passive benchmark is a stock index, the return distribution of the real-life market timing strategy usually has a fatter left tail and a thinner right tail than the return distribution of the buy-and-hold strategy. Put it differently, while the buy-and-hold strategy has a higher variation on gains, the market timing strategy typically has a higher variation on losses. Thus, the market timing strategy can produce precisely the opposite result: instead of reducing the downside risk it may inflate the downside risk.

4.2 Why the market timing strategy sometimes works

Our results suggest that the success of the market timing strategies can be fully explained by the statistical properties of the underlying data series. In particular, the stronger the serial dependence in the underlying data series, the better the expected performance of the market timing strategy. It is no coincidence that the higher the autocorrelations in return series, the better the market timing performance. The intermediate-term government bond index and the S&P Composite stock price index exhibit the highest degree of autocorrelation in the (capital appreciation) return series, see Table 1. Similarly, the real-life market timing strategy that uses the $SMA(k)$ rule delivers the best outperformance when either the intermediate-term government bond index or the S&P Composite stock price index is used as a passive benchmark. The lowest degree of autocorrelation is observed for the long-term government bond index prices. Similarly, when the long-term bond index is used as a passive benchmark, the market timing strategy shows the worst performance regardless of the choice of trading

rule.

An interesting insight into the real-life performance of a stock market timing strategy can be provided by examining its "failure rate". The term "failure rate" is motivated by the following. Recall that every market timing strategy prescribes moving to cash when a Sell signal is generated. The investor, who employs the market timing strategy, hopes that during a Sell period the Treasury bills provide better return than the stocks. We define by "failure" an event in which the Treasury bills fail to deliver a higher return than the stocks during a Sell period. So how often a real-life stock market timing strategy generates false signals? Our analysis reveals that the average failure rate amounts to about 80%. In words, this means that when the market timing strategy generates a Sell signal, in approximately 80% of cases at the end of a Sell period the market timing strategy will provide a lower return than the buy-and-hold strategy.[12] Again, this result demonstrates that, when the passive benchmark is a stock price index, the superior performance of the market timing strategy is confined to some relatively short particular episodes. That means that in order the market timing strategy delivers a superior performance as compared to that of the passive counterpart, investors sometimes have to wait a very long time and experience painful emotions because their active portfolios consistently lag the benchmark. For example, the real-life performance of market timing was inferior during a 25-year period from 1975 to 1999 regardless of the choice of a passive stock market index. It is quite probable that in order the market timing delivers once again a superior performance as that during the decade of 2000s, investors have to wait for another 20+ years.

So why does the market timing strategy sometimes work? There is a couple of simple explanations. Firstly, since the market timing strategy generates a Sell signal after the market prices have been falling over some period, it could therefore be argued that when a Sell signal is generated and the investor switches to cash, the investor is basically betting that the market downturn will continue. Sometimes, because of the small yet statistically significant persistence in price series, the investor guesses right. Other times, this guess comes true by a pure chance. It is true that the market timing as a rule protects from big losses when the stock market downturn is long-lasting and severe. However, in such a case a trailing stop-loss order would

[12] Yet in this case a lower return is always accompanied by a lower risk. As a result, the failure rate for the risk-adjusted performance is lower than 80%.

probably provide better protection than market timing.

4.3 The optimality of market timing for long-term risk-tolerant investors

Risk-tolerant long-term investors must consider very carefully the consequences of the fact that the market timing strategy always produces inferior capital growth in the long run. To elaborate more on this, suppose that the S&P Composite index is used as the passive benchmark. In this case we find that the out-of-sample performance of the market timing strategy over the period 1930 to 2012, judging by the Sharpe ratio criterion, was better than the performance of the buy-and-hold strategy. However, if in 1930 an investor had invested all his money into the market timing strategy instead of the buy-and-hold strategy, this investor would be very disappointed at the end of the period (supposing that he is still alive), because over this period the buy-and-hold strategy provided the capital growth that was almost three times as much as that of the market timing strategy. This example highlights the importance of the optimal capital allocation decision.

Paraphrasing our results, we find that over a long run the buy-and-hold strategy is more risky, but at the same time more rewarding than the market timing strategy. The riskiness of the real-life market timing strategy, when the S&P Composite index is used as the passive benchmark, over the period 1930 to 2012 roughly corresponds to the riskiness of the portfolio which consists of allocating 65% into stocks and, consequently, 35% into Treasury bills. Suppose that the investor can allocate only between stocks and Treasury bills and this investor is rather risk averse so that he allocates no more than 65% of his initial fortune to stocks. Then this investor would be better off if he could allocate his money between the market timing strategy and the Treasury bills. But what if the investor is rather risk tolerant? Then, if the investor chooses to allocate between the market timing strategy and Treasury bills, the optimal capital allocation requires implementing a *leveraged* market timing strategy, that is, borrowing money. Yet in real markets the borrowing costs are prohibitive large. Nowadays there are leveraged Exchange Traded Funds (ETF) that could serve this purpose. However, the tracking error relative to the benchmark index is usually very large (see, for example, Shin and Soydemir (2010)). In addition, Asness, Frazzini, and Pedersen (2012) postulate that investors have aversion to leverage, that is, they are not willing to borrow. Thus, if the leveraged market timing strategy is not feasible, a long-term risk tolerant investor will be better off by investing

in the buy-and-hold strategy even if the Sharpe ratio of the buy-and-hold strategy is somewhat lower than the Sharpe ratio of the market timing strategy.

5 Conclusions

In this paper we presented a detailed study of the real-life performance of the two popular market timing strategies. We demonstrated that the real-life performance of the market timing strategies is substantially worse than the reported performance. Under the stipulation that the real-life performance of the market timing strategies provides us with unbiased estimates of their expected future performance, what performance can investors anticipate from these timing strategies in the future? Our estimates suggest that over a long run investors can expect at best only marginally better risk-adjusted returns as compared to the passive investing. Over a medium run, on the other hand, a stock market timing strategy is more likely to underperform than to outperform the passive strategy. This is because the superior performance of a stock market timing strategy is usually confined to some relatively short particular historical episodes.

Our results suggest that the observed superior performance of technical trading rules can be explained by short-term serial dependence in data. Yet we find that not every passive benchmark exhibits serial dependence that can be potentially exploited in market timing. We confirm that the market timing strategy is indeed less risky, but a lower risk always comes with a lower return and lower capital growth in the long run. Thus, risk-tolerant long-term investors must consider very carefully whether it is wise to follow the market timing strategy even if they believe that the active strategy delivers better risk-adjusted returns than the passive strategy. Last but not least, the fact that the market timing strategy showed a better risk-adjusted tradeoff in the past does not automatically mean that the same timing strategy will show a better risk-adjusted tradeoff in the future as well.

References

Aronson, D. (2006). *Evidence-Based Technical Analysis: Applying the Scientific Method and Statistical Inference to Trading Signals*. John Wiley & Sons, Ltd.

Asness, C. S., Frazzini, A., and Pedersen, L. H. (2012). "Leverage Aversion and Risk Parity", *Financial Analysts Journal*, *68*(1), 47–59.

Bauer, Richard J., J. and Dahlquist, J. R. (2001). "Market Timing and Roulette Wheels", *Financial Analysts Journal*, *57*(1), 28–40.

Berkowitz, S. A., Logue, D. E., and Noser, E. A. (1988). "The Total Costs of Transactions on the NYSE", *Journal of Finance*, *43*(1), 97–112.

Bessembinder, H. (2003). "Issues in Assessing Trade Execution Costs", *Journal of Financial Markets*, *6*(3), 233–257.

Brock, W., Lakonishok, J., and LeBaron, B. (1992). "Simple Technical Trading Rules and the Stochastic Properties of Stock Returns", *Journal of Finance*, *47*(5), 1731–1764.

Chakravarty, S. and Sarkar, A. (2003). "Trading Costs in Three U.S. Bond Markets", *Journal of Fixed Income*, *13*(1), 39–48.

Chan, L. K. C. and Lakonishok, J. (1993). "Institutional Trades and Intraday Stock Price Behavior", *Journal of Financial Economics*, *33*(2), 173–199.

Dermody, J. C. and Prisman, E. Z. (1993). "No Arbitrage and Valuation in Markets with Realistic Transaction Costs", *Journal of Financial and Quantitative Analysis*, *28*(1), 65–80.

Edwards, A. K., Harris, L. E., and Piwowar, M. S. (2007). "Corporate Bond Market Transaction Costs and Transparency", *Journal of Finance*, *62*(3), 1421–1451.

Eling, M. (2008). "Does the Measure Matter in the Mutual Fund Industry?", *Financial Analysts Journal*, *64*(3), 54–66.

Eling, M. and Schuhmacher, F. (2007). "Does the Choice of Performance Measure Influence the Evaluation of Hedge Funds?", *Journal of Banking and Finance*, *31*(9), 2632 – 2247.

Faber, M. T. (2007). "A Quantitative Approach to Tactical Asset Allocation", *Journal of Wealth Management*, *9*(4), 69–79.

Freyre-Sanders, A., Guobuzaite, R., and Byrne, K. (2004). "A Review of Trading Cost Model: Reducing Transaction Costs", *Journal of Investing*, *13*(3), 93–116.

Gartley, H. M. (1935). *Profits in the Stock Market*. Lambert Gann Pub.

Goyal, A. and Welch, I. (2008). "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction", *Review of Financial Studies*, *21*(4), 1455–1508.

Gwilym, O., Clare, A., Seaton, J., and Thomas, S. (2010). "Price and Momentum as Robust Tactical Approaches to Global Equity Investing", *Journal of Investing*, *19*(3), 80–91.

Hakansson, N. H. and Ziemba, W. T. (1995). "Chapter 3 Capital Growth Theory", In R.A. Jarrow, V. M. and Ziemba, W. (Eds.), *Finance*, Vol. 9 of *Handbooks in Operations Research and Management Science*, pp. 65 – 86. Elsevier.

Hansen, P. R. and Timmermann, A. (2013). "Choice of Sample Split in Out-of-Sample Forecast Evaluation", Working paper, European University Institute, Stanford University and CREATES.

Hudson, R., Dempsey, M., and Keasey, K. (1996). "A Note on the Weak form Efficiency of Capital Markets: The Application of Simple Technical Trading Rules to UK Stock Prices–1935 to 1994", *Journal of Banking and Finance*, *20*(6), 1121–1132.

Jegadeesh, N. and Titman, S. (1993). "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency", *Journal of Finance*, *48*(1), 65–91.

Kelly, J. L. (1956). "A New Interpretation of Information Rate", *Bell System Technical Journal*, *2*(3), 185 –189.

Kilgallen, T. (2012). "Testing the Simple Moving Average across Commodities, Global Stock Indices, and Currencies", *Journal of Wealth Management*, *15*(1), 82–100.

Knez, P. J. and Ready, M. J. (1996). "Estimating the Profits from Trading Strategies", *Review of Financial Studies*, *9*(4), 1121–1163.

Lukac, L. P. and Brorsen, B. W. (1990). "A Comprehensive Test of Futures Market Disequilibrium", *Financial Review*, *25*(4), 593–622.

Lukac, L. P., Brorsen, B. W., and Irwin, S. H. (1988). "A Test of Futures Market Disequilibrium Using Twelve Different Technical Trading Systems", *Applied Economics*, *20*(5), 623–639.

Mammen, E. and Nandi, S. (2012). "Bootstrap and Resampling", In Gentle, J. E., Härdle, W. K., and Mori, Y. (Eds.), *Handbook of Computational Statistics*, Springer Handbooks of Computational Statistics, pp. 499–527. Springer Berlin Heidelberg.

Moskowitz, T. J., Ooi, Y. H., and Pedersen, L. H. (2012). "Time Series Momentum", *Journal of Financial Economics*, *104*(2), 228–250.

Nelson, C. R. and Kim, M. J. (1993). "Predictable Stock Returns: The Role of Small Sample Bias", *Journal of Finance*, *48*(2), 641–661.

Park, C.-H. and Irwin, S. H. (2007). "What Do We Know About the Profitability of Technical Analysis?", *Journal of Economic Surveys*, *21*(4), 786–826.

Politis, D. and Romano, J. (1994). "The Stationary Bootstrap", *Journal of the American Statistical Association*, *89*(428), 1303–1313.

Rapach, D. E. and Wohar, M. E. (2005). "Valuation Ratios and Long-Horizon Stock Price Predictability", *Journal of Applied Econometrics*, *20*(3), 327–344.

Rossi, B. and Inoue, A. (2012). "Out-of-Sample Forecast Tests Robust to the Choice of Window Size", *Journal of Business and Economic Statistics*, *30*(3), 432–453.

Shin, S. and Soydemir, G. (2010). "Exchange-traded funds, persistence in tracking errors and information dissemination", *Journal of Multinational Financial Management*, *20*(4-5), 214–234.

Siegel, J. (2002). *Stocks for the Long Run*. McGraw-Hill Companies.

Sullivan, R., Timmermann, A., and White, H. (1999). "Data-Snooping, Technical Trading Rule Performance, and the Bootstrap", *Journal of Finance*, *54*(5), 1647–1691.

White, H. (2000). "A Reality Check for Data Snooping", *Econometrica*, *68*(5), 1097–1126.