Investing in Gold –
Market Timing or Buy-and-Hold?

Dirk G. Baur[1],*, Hubert Dichtl[2], Wolfgang Drobetz[3], and Viktoria-Sophie Wendt[4],‡

This draft: November 2018

Abstract

The literature on gold is dominated by empirical studies on its diversification, hedging, and safe haven properties. In contrast, the question "When to invest in gold?" is generally not analyzed in much detail. We test more than 4,000 seasonal, technical, and fundamental timing strategies for gold and find evidence for some market timing ability and economic gains relative to a passive buy-and-hold benchmark. However, since the results are not robust to data-snooping biases and limited to specific evaluation periods, we conclude that our findings support the efficiency of the gold market.

Keywords: gold; market efficiency; investment strategies; monthly seasonalities; fundamental factors; technical indicators; superior predictive ability test

JEL Classification: G11, G12, G14

[1] University of Western Australia, Business School, Crawley, WA 6009, Australia.

[2] University of Hamburg, Faculty of Business, 20148 Hamburg, Germany.

[3] University of Hamburg, Faculty of Business, 20148 Hamburg, Germany.

[4] University of Hamburg, Faculty of Business, 20148 Hamburg, Germany.

* Corresponding author: E-mail: dirk.baur@uwa.edu.au

‡ *Acknowledgments*: We thank two anonymous referees for helpful comments. We also thank participants at the INFINITI Conference on International Finance (2018) for their feedback.

1. Introduction

Gold is associated with important economic properties such as being a hedge against inflation (Jastram and Leyland, 2009; Blose, 2010; Beckmann and Czudaj, 2013) and currency risk (Sjaastad and Scacciavillani, 1996; Capie, Mills, and Wood, 2005; Reboredo, 2013) as well as providing a safe haven during crisis times (Baur and Lucey, 2010).[1] Based on these economic properties, the literature predominantly focuses on the investment characteristics of gold and its diversification benefits in combination with other asset classes such as stocks and bonds in a multi-asset allocation portfolio (Hillier, Draper, and Faff, 2006; Lucey, Poti, and Tully, 2006; Bruno and Chincarini, 2010; Emmrich and McGroarty, 2013). However, it remains largely unexplored whether the in-sample, i.e., contemporaneous and thus non-predictive, relationship between the gold price and several economic determinants effectively translates into out-of-sample predictability of the gold price. The closely related research question whether the gold market can be actively timed compared to a buy-and-hold strategy has also received relatively little attention. Emmrich and McGroarty (2013) point out that such timing ability would be beneficial given the poor performance of gold over longer time periods.

One of the few studies that addresses these questions is Pierdzioch, Risse, and Rohloff (2014). They study the predictability of monthly gold market excess returns with a real-time forecasting approach based on publicly available financial and macroeconomic predictor variables. Although they provide evidence that the predictor variables under consideration do have market timing ability with respect to the gold market, a simple trading rule that dictates an investment in gold if the forecast of excess gold returns is positive, and otherwise switches to the risk-free asset, does not lead to superior performance relative to a buy-and-hold strategy after accounting for transaction costs. While these findings corroborate the hypothesis of informational efficiency in gold markets with respect to a broad set of fundamental factors, there is empirical evidence for calendar anomalies in gold returns (Baur, 2013),

[1] O'Connor et al. (2015) provide a survey of the broad literature on the financial economics of gold. More recently, the cryptocurrency Bitcoin has been labelled 'new' or 'digital' gold (e.g., Popper, 2016). However, latest studies illustrate that the two assets are very different in terms of return, volatility, and correlation characteristics (Baur et al., 2018; Klein et al., 2018).

and also for the profitability of trend-following trading strategies (Szakmary, Shen, and Sharma, 2010; Moskowitz, Ooi, and Pedersen, 2012).

Starting from this contrasting evidence, we revisit the efficiency of the gold market by simultaneously comparing numerous strategies to time the gold market both in terms of their statistical quality as well as their economic profitability. We build on the extant literature and collect a comprehensive set of market timing signals based on seasonal patterns, trend-following indicators, and fundamental factors. Importantly, we also account for data-snooping. As Sullivan, Timmermann, and White (1999) point out, when many models are evaluated individually, some are bound to show superior performance by chance alone. To formally control for data-snooping when testing for the possible superiority of a market timing strategy, we use Hansen's (2005) test for superior predictive ability (SPA-test) that provides a multiple testing framework without data-snooping bias. In addition, we implement the stepwise extensions of the SPA-test recently proposed by Hsu, Hsu, and Kuan (2010) and Hsu, Kuan, and Yen (2014).

The question whether return predictability, market timing, or any return patterns will persist is fundamental in asset pricing theory.[2] If predictability is "real", i.e., if the outperformance comes from exposure to risk, it is most likely to last. Even if the opportunity is widely publicized, the average investor may not invest because, in equilibrium, the excess risk exactly counteracts the excess return.[3] In this scenario, we should encounter truly superior trading strategies in our gold price data because the average investor understands the risk-return trade-off and refrains from the excess risk exposure. On the other hand, if there is a misperception of risk, it is least likely to persist. In fact, an excess risk premium has the strongest portfolio implications for investors: everyone should try to capture it, portfolio decisions will be adjusted once investors know about the opportunity, and the high return vanishes quickly. In this case, our analysis should not reveal any superior trading strategy. Gold is a highly liquid asset class, and there are many professional market participants (e.g., money managers and hedge fund managers) that

[2] See Cochrane (1999) for a more detailed discussion.

[3] Whether there are economic risk premiums in the gold market is beyond the scope of our analysis. The seminal studies by Ferson and Harvey (1993, 1994) test multifactor models for global stock markets that contain a commodity risk premium. While the estimates for their commodity risk premium is not significant individually, the goodness-of-fit of a multifactor model that contains a commodity risk premium cannot be rejected.

trade the asset class using a variety of instruments (e.g., physical gold or derivative instruments). Therefore, our multiple testing framework, by allowing us to identify superior trading strategies that are free of data snooping biases, can help to better understand the efficiency of the gold market.

Our simulation results reveal that there are several market timing strategies that predict the direction of the gold market better than a simple random walk and outperform a passive buy-and-hold strategy in traditional back-tests. However, using a data-snooping resistant testing framework, we find only weak evidence for superior statistical and economic performance of any market timing strategy over the full evaluation period. For specific sub-samples (i.e., different evaluation periods), we identify stronger evidence for outperformance of the market timing signals, but the analysis implicitly assumes that the optimal sub-sample periods can be identified ex ante. All in all, our findings demonstrate that the gold market is efficient with respect to a broad set of seasonal, technical, and fundamental factors.

The remainder of our study is as follows: Section 2 provides a summary of the existing literature on timing the gold market. Section 3 describes the simulated market timing signals, outlines the multiple testing framework used to account for the data snooping bias, outlines the multiple testing framework used to account for the data snooping bias, and defines the relevant performance measures. Section 4 contains a data description and presents the results of our simulations. Finally, section 5 concludes.

2. Literature review

Pierdzioch, Risse, and Rohloff (2014) use a real-time forecasting approach based on publicly available predictor variables such as the inflation rate, the term spread, and changes in oil prices to study the predictability of monthly gold market excess returns. These fundamental factors serve as predictors of future gold returns by affecting either demand or supply of gold or the expectation of market participants thereof. Using different model selection criteria and model averaging approaches, they conclude that the gold market is informationally efficient with respect to the predictor variables.[4] In a related

[4] In a different strand of literature, Mihaylov, Cheong, and Zurbruegg (2015) document that gold returns can also be predicted using the difference between consensus forecasts for firms primarily operating in the gold market and current earnings.

study, Pierdzioch, Risse, and Rohloff (2015) use boosting algorithms to build a forecasting model for gold from various simple auxiliary models and find that parsimonious forecasting models that tend to perform well in statistical terms do not survive an economic performance evaluation.[5] Baur, Beckmann, and Czudaj (2016) use a dynamic model averaging approach, mitigating both the uncertainty about the variables to include in a prediction model (model uncertainty) and the uncertainty about the time-variation of the parameters (parameter uncertainty), to forecast gold prices. They provide evidence for significant time-variation in the influence of gold price predictors and conclude that the in-sample functions of gold such as the currency hedging property do not translate into out-of-sample predictability.

Nguyen et al. (2017) document that the excess return of gold is predictable using the left jump risk premium and the gold variance risk premium. They interpret their results as evidence for the existence of a gold risk premium that is time-varying and predictable. Charles, Darné, and Kim (2015) also stress that return predictability is time-varying and dependent on the current political and economic environment. While these results endorse the efficiency of the gold market, several other recent studies indicate behavioral based inefficiencies, such as psychological barriers around key reference points in gold prices (Aggarwal and Lucey, 2007; Lucey and O'Connor, 2016) or market optimism (Aggarwal, Lucey, and O'Connor, 2014).

There is also growing empirical evidence for anomalies in gold returns. Baur (2013) finds the so called 'autumn effect' in the gold market, with positive and statistically significant gold returns limited to the months September and November, which is supported by economic intuition, e.g., by the increased wedding season gold jewelry demand in India. Similar seasonal patterns exist in the Chinese gold market (Qi and Wang, 2013) and in U.S. gold ETF returns (Naylor, Wongchoti, and Ith, 2014).

Furthermore, several studies provide evidence on the performance of technical trading strategies in the gold market. Moskowitz, Ooi, and Pedersen (2012) find time series momentum in monthly returns

[5] The weak relationship between statistical performance measures and economic value has also been discussed in the context of stock return prediction (Leitch and Tanner, 1991; Cenesizoglu and Timmermann, 2012).

of several liquid instruments, including gold futures.[6] While Marshall, Cahan, and Cahan (2008) find no evidence for the profitability of technical trading rules using daily data, Szakmary, Shen, and Sharma (2010) examine the profitability of moving average and channel strategies for monthly commodity futures and report significant positive excess returns.[7] Similar results are provided in Batten et al. (2015) for intraday trading. As pointed out by Chevallier, Gatumel, and Ielpo (2013), the gold market exhibit a high degree of return persistence, which potentially benefits trend-following strategies. This conjecture is confirmed by Auer (2016), who shows that trading strategies based on the Hurst coefficient, a simple measure of long-range dependence, significantly outperform a buy-and-hold strategy in the gold market.

3. Empirical procedure

3.1. Market timing signals

To construct a comprehensive collection of market timing signals, we build on the economic properties of gold and incorporate the findings regarding previously surveyed anomalies in gold prices. All our market timing signals are constructed such that they are equal to one if the future gold market excess return is expected to be positive and thus an investment in the gold market is warranted, and zero otherwise. Table 1 provides a short description of all market timing signals including their abbreviations.

[Insert Table 1 here]

Seasonal market timing signals: To exploit monthly seasonality in the gold market, for example the ‘autumn effect’ (Baur, 2013), an investor could follow a simple market timing strategy that signals a positive gold market excess return and thus an investment in gold only in September and November. While this strategy is supported by economic intuition, the strategy itself is merely mechanic. Therefore, to address the concern that this strategy has simply been ‘mined’ from the data, we follow Dichtl and Drobetz (2014) and implement the following approach: Each month, an investor can either be invested in the gold market ($S_{i,t} = 1$) or not ($S_{i,t} = 0$). In this vein, we obtain $2^{12} = 4{,}096$ different monthly

[6] Miffre and Rallis (2007) verify the profitability of momentum strategies also in the cross-section of commodity futures.

[7] In a related study, Narayan, Narayan and Sharma (2013) find similar profitable trading strategies in commodity spot markets.

seasonal market timing signal functions, labeled from *SEA0* to *SEA4095*, where SEA stands for seasonal. While the *SEA0* strategy signals an investment in the gold market for each of the twelve months, the *SEA4086* strategy, for example, invests in the gold market only in September, and the *SEA4095* strategy refrains from investing in the gold market in all twelve months. In contrast to all other monthly allocation strategies, the buy-and-hold strategy (*SEA0*) and the cash-only strategy (*SEA4095*) do not incur any transaction costs. To limit the possible number of combinations, the same monthly allocation is implemented in each year during the sample period for each seasonal market timing strategy.

Technical market timing signals: We include monthly technical signals based on moving average and momentum indicators in our empirical analyses.

A moving average signal is defined as $S_{i,t} = \begin{cases} 1 & \text{if } MA_{s,t} \geq MA_{l,t} \\ 0 & \text{if } MA_{s,t} < MA_{l,t} \end{cases}$ with $s < l$, where $MA_{j,t} = (1/j) \sum_{i=0}^{j-1} P_{t-i}$ for $j = s, l$ based on the gold price index P_t. Following Szakmary, Shen, and Sharma (2010), the short index for the moving average is set to $s = 1, 2, 3$ and the long index to $l = 9, 12$, which results in six moving average signals. We also set $s = 1$ and $l = 10, 24, 36, 48$ to cover some other parameterizations that are popular in the literature and among investors. Overall, we employ ten different moving average signals, labeled as *MAs-l* (e.g., *MA1-10*).

A time series momentum signal equals unity if the gold market exhibits positive time series momentum, i.e., the actual price P_t is equal or greater than the m-month lagged price P_{t-m}, and zero otherwise: $S_{i,t} = \begin{cases} 1 & \text{if } P_t \geq P_{t-m} \\ 0 & \text{if } P_t < P_{t-m} \end{cases}$. Following Moskowitz, Ooi, and Pedersen (2012), we set $m = 1, 3, 6, 9, 12, 24, 36, 48$. The resulting eight momentum signals are labeled as *MOMm* (e.g., *MOM9*).

Fundamental market timing signals: We use a set of eleven fundamental and macroeconomic predictor variables that are economically linked to gold price fluctuations to construct corresponding market timing signals.[8] To derive the market timing signals, we first estimate a simple linear regression model for each predictor variable (see Mui and Chu (1993) for forecasts of the spot price of gold):

$$r_{t+1}^{exc} = \alpha_i + \beta_i x_{i,t} + \varepsilon_{i,t+1} \tag{1}$$

where r_{t+1}^{exc} is the excess gold return from period *t* to *t+1*, $x_{i,t}$ a predictor variable, α_i and β_i are regression parameters that are estimated using OLS, and $\varepsilon_{i,t+1}$ is a zero-mean disturbance term.[9] Once the regression parameters are estimated, we use the fitted model together with the observed value of the predictor variable to forecast next month's excess gold return. Finally, as in Pesaran and Timmermann (1995), we transform the predicted excess gold returns into market timing signals using the following rule: $S_{i,t} = \begin{cases} 1 \text{ if } \hat{r}_{t+1}^{exc} \geq 0 \\ 0 \text{ if } \hat{r}_{t+1}^{exc} < 0 \end{cases}$.

Mui and Chu (1993) show that combining individual forecasts from simple predictive regression models can produce superior out-of-sample forecasts. Therefore, we derive another potential market timing signal based on the simple mean of all individual forecasts ($MEAN$).[10]

However, one shortcoming of this approach is that potential interdependencies between various predictor variables are not considered. To address this drawback, we also consider a market timing signal based on a "kitchen sink" forecast (KSF), which incorporates all available predictor variables simultaneously in a multivariate regression model. Albeit this model often performs poorly in terms of the mean squared forecast error because of over-parameterization, it can still potentially generate profits (Rapach and Zhou, 2013).

[8] Dichtl (2017) provides detailed economic explanation for each of these eleven predictor variables.

[9] We estimate the regression model using a rolling window approach that consists of 120 monthly observations, e.g., the parameter estimates for the first signal in January 1990 depend on the observations from January 1980 through December 1989. Giacomini and White (2006) discuss the advantage of a rolling window approach compared with an expanding window framework. The use of Hansen's (2005) SPA-test depends on the stationary assumption that is violated within an expanding window approach but not within a rolling window approach (see also the detailed discussion in Hansen, 2005).

[10] In the context of stock market prediction, Rapach and Zhou (2013) conclude that, despite considering several more sophisticated combination methods, the simple average of all individual forecasts performs surprisingly well. This simple combination scheme has the advantage that it does not require the estimation of combining weights, thereby reducing forecasting risk.

In order to alleviate the problem of in-sample overfitting when estimating multivariate regression models such as KSF, Tibshirani (1996) proposes the least absolute shrinkage and selection operator (LASSO) to improve prediction accuracy and perform variable selection. The LASSO objective function is given by $\min_{\alpha,\beta}\left(\sum_{t=1}^{T-1}\left(r_{t+1}-\alpha-\sum_{i=1}^{N}\beta_i x_{i,t}\right)^2+\lambda\sum_{i=1}^{N}|\beta_i|\right)$, where λ is a penalty term.[11] We apply LASSO to the multiple regression model that includes all available predictor variables, and transform the resulting forecasts into market timing signals.

Finally, we consider a diffusion index approach, where all available predictor variables are aggregated into a relatively small number of diffusion indices using principal component analysis. Ludvigson and Ng (2007, 2009), for example, use the diffusion index approach to forecasting equity and bond risk premiums. We implement the approach by considering the first principal component ($PC1F$) or the first two principal components (adding $PC2F$) as predictors in our simple predictive regression model and transforming the forecasts in corresponding market timing signals.

3.2. Implementation of multiple testing framework

A drawback of traditional back-test studies that evaluate their market timing signals on a single historical return path is that the result may be purely from chance, and not due to any genuine merit (Sullivan, Timmermann, and White, 1999, 2001). This bias in statistical inference is usually referred to as 'data snooping'. Without properly adjusting for this bias in a multiple testing set-up, we might commit a type I error, i.e., falsely assessing a market timing strategy as being superior when it is not. To control for data snooping when evaluating our market timing strategies, we apply Hansen's (2005) test for superior predictive ability, or SPA-test, that allows us to compare the market timing strategies not only against a benchmark, but also to draw statistical inference from the empirical distribution of a perfor-

[11] If $\lambda = 0$, the LASSO estimates are equivalent to the OLS estimates in equation (1); by increasing λ, the parameters are shrunk towards zero. To select the appropriate value for λ, we use ten-fold cross-validation and chose the value of λ that minimizes the mean cross-validated error.

mance measure by considering the full population of strategies from which the strategy under consideration is selected. Moreover, we implement two stepwise extensions of the SPA-test developed by Hsu, Hsu, and Kuan (2010) and Hsu, Kuan, and Yen (2014) to identify all significant market timing strategies.

SPA-test: We test the null hypothesis that a chosen benchmark model is not inferior to any alternative market timing strategy $H_0\colon \max_{j=1,\dots,J} E(d_{j,t}) \equiv \mu_j \leq 0$, where $d_{j,t}$ is the difference of the performance measure of market timing strategy j and the performance measure of the benchmark at time t. If the null hypothesis can be rejected, there is at least one market timing strategy that outperforms the benchmark.

The SPA-test uses the following studentized test statistic $V_t^{SPA} = max\left(\max_{j=1,\dots,J} \frac{\sqrt{T}\bar{d}_j}{\widehat{\omega}_j}, 0\right)$, where $\overline{d}_j = \sqrt{T}\sum_{t=1}^{T} d_{j,t}$ denotes the average relative performance of strategy j, and $\widehat{\omega}_j^2$ is some consistent estimator for $\omega_j^2 \equiv var(\sqrt{T}\bar{d}_j)$. Hansen (2005) proposes a bootstrap simulation approach to obtain the distribution of the SPA-test statistic under the null hypothesis by implementing the stationary bootstrap simulation approach of Politis and Romano (1994).[12] From the bootstrap population, we generate $B = 10{,}000$ resamples that we re-center according to Hansen (2005). For each resample, the studentized test statistic under the bootstrap $V_{b,t}^{SPA^*}$ is computed. A consistent estimate of the *p*-value is $\widehat{p}_{SPA} = \sum_{b=1}^{B} \frac{\mathbb{1}_{\left\{V_{b,t}^{SPA^*} > V_t^{SPA}\right\}}}{B}$, where the null hypothesis of the SPA-test that the benchmark model is the best market timing strategy is rejected for small *p*-values.

Hansen (2005) shows that both an upper and a lower bound for the *p*-value can be obtained. The upper bound is the *p*-value of a conservative test, which assumes that all competing strategies are exactly as good as the benchmark model. In contrast, the lower bound constitutes a liberal test that assumes that

[12] The chosen bootstrap approach involves combining blocks with random lengths that are chosen to be geometrically distributed with a mean block length of q^{-1}. In our empirical application, we set the smoothing parameter $q = 0.5$.

the strategies with worse performance than the benchmark model are poor models in the limit. A large difference between the upper and lower bound p-value may be an indication of many poor models.

Step-SPA-test: While the SPA-test can answer the question whether there is at least one superior market timing strategy, if any, it is not able to identify all such strategies. Therefore, we implement a stepwise extension of the SPA-test (step-SPA-test) developed by Hsu, Hsu, and Kuan (2010) to identify all significant market timing strategies if the null hypothesis of the SPA-test is rejected. First, we re-arrange the market timing strategies in descending order of their test statistic and reject the top strategy if its test statistic is greater than the critical value, specified as the $1-\alpha$ quantile of the empirical distribution bootstrapped from the entire population of market timing strategies.[13] Second, we remove $\bar{d}_j$ of the rejected strategy and compute a new critical value bootstrapped from the subset of remaining market timing strategies. We again reject the top strategy if its test statistic is greater than the new critical value and repeat this procedure until no further market timing strategy can be rejected. All market timing strategies that have been removed are identified as superior strategies.

Step-SPA(k)-test: The step-SPA-test is able to successfully identify all superior strategies when the null hypothesis of the SPA-test is rejected, but is fairly conservative in doing so, as it controls the family wise error rate, i.e., the probability of at least one false rejection given the pre-specified error rate α. However, one might be willing to tolerate a higher number of false rejections to increase test power and be able to better reject false null hypotheses. Hsu, Kuan, and Yen (2014) develop a refinement of the step-SPA-test, the step-SPA(k)-test, that asymptotically controls the probability of at least k false rejections, with $k \geq 2$, less than or equal to a certain level α. The implementation of the step-SPA(k)-test is similar to the step-SPA-test but relies on the empirical distribution of the k-th largest test statistic and tests all possible combinations of the $k-1$ strategies to determine the maximum critical value among all combinations.[14] In our empirical analysis, we follow Hsu, Kuan, and Yen (2014) and set $k=3$.

[13] In our empirical analysis, we determine the critical values for the pre-specified error rate $\alpha = 5\%$.

[14] Hsu, Kuan, and Yen (2014) provide further details on the implementation of the step-SPA(k)-test.

FDP-SPA-test: A drawback of the step-SPA(k)-test is that the choice of k is arbitrary and does not depend on the underlying data. If, however, the total number of rejections is large, one might be willing to tolerate a higher number of false rejections, or vice versa, and therefore chose to control the proportion of false rejections to the number of total rejections, i.e., the false discovery proportion (FDP), instead. Hsu, Kuan, and Yen (2014) extend their step-SPA(k)-test to asymptotically control the probability of the FDP exceeding a pre-defined proportion γ at the level α (FDP-SPA-test).[15] The implementation of the FDP-SPA-test relies on the sequential application of the step-SPA(k)-test: First, set $k = 1$ and apply the step-SPA(k)-test. If the number of rejected market timing strategies is less than $k/\gamma - 1$, the procedure stops and identifies all rejected strategies as superior. Otherwise, set $k = k + 1$ and repeat this procedure until no further market timing strategy can be rejected. In our empirical analysis, we follow Hsu, Kuan, and Yen (2014) and set $\gamma = 0.1$, i.e., we require less than 10% of rejected strategies to be falsely identified.

3.3. Performance measures and benchmark models

To evaluate the quality of our market timing signals, we employ both statistical as well as economic performance measures and choose applicable benchmark models.

Directional accuracy measures whether the direction of changes in the gold excess returns is correctly predicted and is given by the following indicator function (Pesaran and Timmermann, 1995):

$$DA_{j,t} = S_{i,t} \times I_t + (1 - S_{i,t}) \times (1 - I_t) \tag{2}$$

where $S_{i,t}$ is one of the seasonal, technical, or fundamental market timing signals discussed in section 3.1 and I_t is an indicator function that equals unity if the realized gold excess return at time t is positive, and zero otherwise. As argued in Leitch and Tanner (1991), if a market is considered efficient, predicting the changes in excess returns should be equivalent to tossing a coin. Therefore, we choose

[15] γ denotes the desired false discovery proportion (FDP). $\gamma = 0.1$ implies that we do not want more than 10% of all identified (rejected) strategies to be falsely rejected. The probability that the FDP is greater than 10% is controlled at the significance level α=5%. See Chordia et al. (2017) for more detailed discussion.

the random walk model without drift as the benchmark for statistical evaluation (Baur, Beckmann, and Czudaj, 2016).[16]

However, even if we were able to establish that the considered market timing signals can correctly predict the sign of future gold market excess returns, such statistical performance may not indicate opportunities for profit-making, especially when taking transaction costs into account. Therefore, in order to assess the economic value of the market timing signals, we devise a simple trading strategy that takes an investment in gold if the market timing signal is unity, or otherwise switches to an alternative risk-free one-period investment, while assuming 20 basis points as turnover-dependent costs.

As performance measures, we consider absolute returns, $r_{j,t}^{abs}$, and risk-adjusted excess returns (or Sharpe ratio; Sharpe, 1994), $r_{j,t}^{exc}/\sigma_j$, where $r_{j,t}^{exc}$ denotes the excess return above the risk-free rate (defined as a simple return) and σ_j the volatility of the excess return. As demonstrated by Zakamouline (2011), Ornelas, Silva, and Fernandes (2012), and Adcock et al. (2014), the choice of the performance measure can strongly influence the evaluation of risky portfolios. While higher moments of the return distribution play a significant role in performance evaluation, the effects of skewness or kurtosis on performance evaluation depend on the choice of the performance measure. Given that gold excess returns deviate from normality (see Table 2 above), it may be important to use performance measures that are not simply monotonically increasing functions of the Sharpe ratio.[17] Therefore, we also apply downside risk-adjusted excess returns (or Sortino ratio; Sortino and Price, 1994), $r_{j,t}^{exc}/\sigma_{d,j}$, where $\sigma_{d,j}$ denotes the volatility of negative excess returns. We compare the performance of the resulting market timing strategies against a simple buy-and-hold benchmark (Pierdzioch, Risse, and Rohloff, 2014).

[16] Our results are unchanged when we assume a random walk with drift as our benchmark model and estimate the drift term as the current mean gold return using an expanding window starting from the beginning of the sample period.

[17] Adcock et al. (2014) show that, subject to regularity conditions, all performance measures, which are increasing functions of reward and decreasing functions of risk, are monotonically increasing functions of the Sharpe ratio. By contrast, performance measures which employ upper partial moments or conditional expected excess returns as measures of reward are not robust to differences in distributions. However, analyzing commodity investment returns, Auer (2015) finds a high degree of ranking similarity across performance measures although commodity returns are non-normal.

4. Empirical results

Our dataset comprises monthly data from December 1979 through December 2017. We use end-of-month spot gold fixing prices from the London Bullion Market (3:00 PM, London time) in USD and compute continuously compounded monthly returns in excess of the risk-free rate.[18] After an initial estimation period for the fundamental regression models, we evaluate the performance of all market timing signals over the period from January 1990 to December 2017.

4.1. Descriptive statistics

Panel A of Figure 1 shows the price of gold as well as monthly gold excess returns over the full sample period. The price of gold varies substantially over time: The relatively high gold price at the beginning of the sample period was substantially corrected over the course of the following years (from around 670 USD per ounce in September 1980 to around 290 USD in February 1985). Subsequent to a period of relatively minor price movements, the gold price experienced a major jump from around 280 USD at the beginning of 2002 up to 1,800 USD in August 2011. Monthly excess returns show a similar pattern; highly volatile returns until the mid-1980s are followed by a period of relative calamity up to the beginning of the price correction at the end of 2011. Panel B of Figure 1, taking different starting years, shows forcefully that an investment in gold rather than in the risk-free asset was detrimental to an investor's wealth as cumulative gold excess returns have been negative for prolonged periods of time.

[Insert Figure 1 here]

Table 2 shows descriptive statistics for monthly gold excess returns and results of simple weak-form market efficiency tests (autocorrelation tests and runs tests). The corresponding values are reported for both the full sample period (1980:01-2017:12) and the evaluation period (1990:01-2017:12). Panel A of Table 2 reveals that the mean monthly excess return is higher in the evaluation period compared to

[18] The gold price series is from the Federal Reserve Bank of St. Louis (https://research.stlouisfed.org/fred2). The London Bullion Market (LBMA) gold price index is also used as the benchmark for exchange traded commodities (ETCs) that are easily investible even for private investors (e.g., the iShares Physical Gold ETC). In addition, we use the Treasury bill rate provided in Goyal and Welch (2008) data set as the risk-free rate.

the full sample period, with 0.12% and -0.16%, respectively. As already shown in Figure 1, this is attributable to the steady increase in gold prices between 2002 and 2012. Conversely, the standard deviation is higher during the full evaluation period, which resembles the impact of the volatility spikes in the early 1980s. The Jarque-Bera test statistics suggest that the null hypothesis of a normal distribution of gold excess returns must be rejected for both time periods.

[Insert Table 2 here]

The autocorrelation coefficients in panel B of Table 2 indicate that, at least for the evaluation period, there is higher-order serial correlation in monthly gold excess returns. However, none of the autocorrelation coefficients up to lag 36 is greater than 0.2 (not reported). The test statistics of run tests in panel C of Table 2 indicate that the distribution of monthly gold excess returns is random in both periods, regardless of the cutoff point used to define runs. However, when we apply rank-based (R_1) and sign-based (S_1) variance-ratio tests (Wright, 2000) for three different holding periods (3, 12, and 36 months) to our gold return series, both the random walk and the martingale difference hypothesis is rejected for the full sample period. For the evaluation period, we still identify deviations from weak-form efficiency at longer holding periods. Overall, we find contrasting evidence regarding the weak-form efficiency of monthly gold market returns that could potentially be exploited by appropriate market timing strategies.

4.2. Performance of the simulated market timing strategies

We assess the performance of all market timing signals over the evaluation period from January 1990 to December 2017 and measure their performance in statistical terms, i.e., mean directional accuracy, as well as economic profitability, i.e., mean absolute returns, mean risk-adjusted excess returns (Sharpe ratios), and mean downside risk-adjusted excess returns (Sortino ratios). Table 3 lists the five best market timing strategies for each of the three strategy groups: seasonal, technical, and fundamental market timing signals.

[Insert Table 3 here]

The best performing seasonal market timing strategies exhibit a mean directional accuracy of no less than 57% (panel A), which is much higher than the accuracy of the random walk model. Moreover, the mean return of the best performing seasonal market timing strategies exceeds 0.55% per month (panel B), outperforming the buy-and-hold strategy by more than 0.19 percentage points per month (see Appendix 1 for descriptions of monthly allocations). Turning to the Sharpe ratios (panel C), our results are similar, i.e., buy-and-hold underperforms the five best monthly seasonality strategies by at least eight percentage points in terms of average risk-adjusted excess returns. Outperformance is even more pronounced when measured in terms of Sortino ratios (panel D).

In comparison with the set of seasonal market timing strategies, the best technical market timing strategies exhibit slightly lower mean directional accuracy as well as lower monthly returns, Sharpe ratios, and Sortino ratios. Similar results are obtained regarding the performance of fundamental market timing strategies. Nevertheless, the best fundamental market timing strategies still outperform the buy-and-hold strategy by at least eleven percentage points in terms of average monthly return and four (five) percentage points in terms of Sharpe (Sortino) ratios.

Overall, while the results in Table 3 suggest that market timing would have been beneficial over our evaluation period and already provide a first indication which strategies perform well, these analyses neither test for statistical significance nor account for the data-snooping problem (Lo and MacKinlay, 1990; Harvey, 2017). To evaluate the statistical significance of performance relative to the benchmark, while controlling for data snooping, we apply Hansen's (2005) SPA-test to all three types of strategies. In each test, the strategies are compared with the performance of the benchmark model. The first test considers all 4,095 different monthly seasonal market timing strategies,[19] the second test all 18 technical market timing strategies, and the third test all 16 fundamental market timing strategies. Table 4 summarizes the SPA-test results.

[Insert Table 4 here]

[19] In total, we have 4,096 seasonal market timing strategies (see section 3.1 for details). However, the SEA0 strategy is identical to the buy-and-hold benchmark model and excluded from the test set.

Panel A of Table 4 shows the results based on mean directional accuracy, while the following panels display the outcomes based on mean monthly absolute returns as well as the monthly Sharpe and Sortino ratios. Column (1) describes the set of strategies, column (2) indicates the chosen benchmark model, and column (3) denotes the best strategy within the test set, i.e., the market timing strategy with the highest performance measure. The nominal *p*-value in column (4) results from a pairwise comparison of the best strategy with the benchmark model. In contrast to the *p*-value of the SPA-test, this *p*-value does not account for the entire set of market timing strategies. Columns (5) and (6) report the consistent *p*-value together with the lower and upper bound *p*-values of the SPA-test as well as the number of significant strategies identified by the step-SPA-test, the step-SPA(*3*)-test, and the FDP-SPA-test, respectively.[20]

The results in panel A confirm that the best market timing strategies, whether based on seasonal, technical, or fundamental signals, can outperform the random walk model based on mean directional accuracy when considered in isolation, as indicated by the nominal *p*-values below 0.05. However, when correcting for data snooping biases, we are unable to reject the null hypothesis of the SPA-test for any set of market timing strategies; all consistent *p*-values exceed 5%, i.e., there is no market timing strategy that significantly outperforms the random walk. While we are able to identify several superior fundamental market timing strategies that outperform the random walk model at the 5% level of significance when allowing for a higher number of false rejections (step-SPA(*3*)-test), the FDP-SPA-tests do not identify any strategy as significantly superior. The discrepancy between the results of the step-SPA(*3*)-tests and the FDP-SPA-tests suggest that the arbitrarily chosen value of $k = 3$ for the step-SPA(*k*)-tests is very liberal within this test set. This indicates that, when applying the step-SPA(*3*)-tests in this case, the realized false discovery proportion (FDP) exceeds 10% (the pre-defined proportion γ of the FDP-SPA-tests), and these ‘superior’ fundamental market timing strategies might be in fact falsely identified.

[20] As suggested by an anonymous referee, we also apply the more commonly known Bonferroni and Holm (1979) methods to our set of strategies in Table 4. The Bonferroni adjusted *p*-values are higher than the consistent SPA *p*-values, effectively serving as a conservative bound. Using the Holm (1979) method leads to the same conclusion as the step-SPA-tests. Given that the Bonferroni and Holm (1979) methods are less powerful than the (step-) SPA-test because they disregard the dependence structure of the individual statistics, we do not report the results of these robustness checks in the paper, but provide them upon request.

Turning to economic profitability, the best seasonal market timing strategy outperforms the buy-and-hold benchmark based on mean monthly returns (panel B of Table 4) only at a statistical significance level of 10% in a pairwise comparison, as indicated by the nominal *p*-value of 0.0868. Consequently, the consistent *p*-value of 0.7458 indicates that there is no seasonal market timing strategy that significantly outperforms the buy-and-hold benchmark once accounting for data snooping biases. This result holds even for the more liberal test (with lower bound *p*-value of 0.5833). However, the spread between lower and upper *p*-values is substantial, indicating the presence of many poor seasonal market timing strategies. Allowing for a higher number of false rejections does not qualitatively change this conclusion, as indicated by the result of the step-SPA(*3*)-test.

Turning to technical market timing strategies, the best strategy, based on the cross-over of the one-month and four-year moving averages (*MA1-48*), outperforms the buy-and-hold benchmark only at a statistical significance level of 10% in a pairwise comparison, as indicated by the nominal *p*-value of 0.0733. Once accounting for the full set of technical market timing strategies, we again find no evidence of outperformance relative to the buy-and-hold strategy, even if relaxing the number of false rejections.

Finally, the results for fundamental market timing strategies indicate that even the best fundamental market timing strategy based on a "kitchen sink" forecast does not statistically outperform the buy-and-hold benchmark in a pairwise comparison, and we are not able to identify any fundamental market timing strategy that exhibits superior performance in a multiple-testing setting, confirming the results of Pierdzioch, Risse, and Rohloff (2014).

Panel C of Table 4 shows the results based on monthly Sharpe ratios. The best strategies among all three types of strategies show a better performance relative to buy-and-hold at least in pairwise comparisons and significantly outperform the buy-and-hold benchmark in a single-testing setting, with nominal *p*-values below 0.05. However, we cannot reject the null hypothesis of the SPA-test, with all consistent *p*-values exceeding 10%. While the step-SPA(*3*)-test identifies at least one significant technical market timing strategy, namely the *MA1-48* strategy, the pursuant FDP-SPA-test does not reject the null hypothesis when controlling the false discovery proportion (FDP) at 10%. The somewhat better performance of market timing strategies when measuring performance in terms of monthly Sharpe ratios rather than absolute returns is attributable to the reduced return variance due to the prolonged investment in the risk-free rate.[21] Finally, the results based on monthly Sortino ratios (panel D of Table 4) are qualitatively similar to the results based on monthly Sharpe ratios.

Our results so far suggest that, while several strategies show superior market timing ability relative to the random walk model, this ability does not necessarily translate into robust economic gains. Although we can identify several market timing strategies that generate a strong performance relative to a buy-and-hold strategy over the full evaluation period, especially in terms of risk-adjusted excess returns, and are also statistically significant at least at the 10% level in isolation, this outperformance is not robust to data snooping concerns and vanishes when we apply the SPA-test. Therefore, our results cast doubt on the future success of the identified 'best' gold timing strategies under realistic trading conditions.

4.3. Sub-sample analyses

Although the bootstrap procedure implemented for the SPA-tests provides us with random subsamples, it still relies on the original data series and is therefore influenced by the choice of the evaluation period. Moreover, the performance of the buy-and-hold benchmark strategy is highly influenced by the explosive price behavior between 2002 and 2012 (see Figure 1). To verify whether our results are robust to the choice of the evaluation period, we repeat our analyses for three sub-sample periods: a period characterized by a relatively stable gold price development from 1990 to 2001, a period exhibiting explosive bubble-like price dynamics from 2002 to 2012 (Baur and Glover, 2015), and the subsequent price correction from 2013 onwards. Table 5 summarizes the number of significant strategies identified by the stepwise extensions of the SPA-test and the FDP-SPA-test in each sub-sample period.

[21] The excess returns of technical market timing strategies, for example, exhibit an average monthly standard deviation of 3.56% compared to 4.53% for the buy-and-hold strategy during the evaluation period. Given that the technical market timing strategies are invested in the gold market only for 56% of the evaluation period, on average, this difference is driven by the investment in the risk-free asset.

[Insert Table 5 here]

With respect to directional accuracy, the results in panel A of Table 5 indicate that the random walk benchmark would have been significantly outperformed by all three groups of market timing signals during the gold market boom from 2002 to 2012. Given that the random walk benchmark implicitly assumes an expected return of zero, this result is not surprising. In the remaining sub-sample periods, however, the random walk constitutes a hard benchmark to beat, confirming the results of Baur, Beckmann, and Czudaj (2016). Furthermore, when comparing the results of panel A with panels B to D, it becomes apparent that the periods of superior statistical performance do not coincide with periods of superior economic profitability, confirming previous results by Leitch and Tanner (1991).

With respect to economic profitability, the results in panels B to D of Table 5 show that the buy-and-hold benchmark would have been significantly outperformed by seasonal as well as fundamental market timing strategies both in terms of mean absolute returns and Sharpe ratios during the sub-sample period from 1990 to 2001. While the buy-and-hold strategy was hurt by negative gold excess return over this period (-0.25% per month), most fundamental factors delivered the correct signal and preserved investment value by switching into the risk-free asset. In fact, the fundamental market timing strategies signaled, on average, a positive gold excess return only in 22 of the 144 months in this sub-sample period. In contrast, in the second sub-sample period, no market timing strategy was able to outperform the buy-and-hold benchmark, which is expected given the extreme positive price trend over this period. For the third sub-sample, the buy-and-hold strategy once again suffered from the negative average gold excess returns (-0.42% per month), but none of the market timing strategies was able to outperform the buy-and-hold strategy over this period.

These patterns further support our conclusion that the gold market is informationally efficient. Several recent studies show that knowledge about an anomaly, in particular, its first publication in the academic literature, is important for future profitability (Schwert, 2003; Jacobsen and Visaltanachoti, 2009; McLean and Pontiff, 2016; Dichtl and Drobetz, 2015). If publication generates the attention of sophisticated investors, who learn about mispricing and start trading against the mispricing, we expect an anomaly to fade away.

5. Conclusion

The question when to invest in gold has received considerably less attention than the question whether gold is a portfolio diversifier, a hedge, or a safe haven. This study aims to reduce this imbalance by analyzing if the direction of the gold market can be accurately predicted, and if active market timing strategies can outperform a passive buy-and-hold strategy. For our evaluation period from January 1990 to December 2017, our back-test results for more than 4,000 different seasonal, technical, and fundamental market timing signals reveal that the directional accuracy of many signals exceed those of a simple random walk, and that the corresponding market timing strategies deliver significant excess returns. While the best fundamental and technical market timing strategy outperformed a buy-and-hold strategy by about 2.3 percentage points per year, the best seasonal trading strategy outperformed a passive strategy by at least 2.7 percentage points per year.

However, because all market timing strategies are data-mined and thus exist by pure chance, these strategies do not necessarily work in the future. To account for this data snooping bias, we apply a multiple testing framework and find that, although some market timing strategies predict the direction of the gold market better than a simple random walk model, no strategy shows a statistically significant economic outperformance relative to the buy-and-hold strategy over the full evaluation period. The weak relationship between economic and statistical performance measures confirms previous findings in the literature.

Additional sub-sample analyses reveal that the statistical outperformance of the market timing signals is confined to the period from 2002 to 2012, which is characterized by explosive gold price dynamics. Any economic gains are mostly generated in the period prior to 2002, when gold excess returns were negative, and many market timing strategies revert to holding the risk-free asset. Overall, our findings lend support to the efficiency of the gold market with respect to a broad set of seasonal, technical and fundamental factors.

Figures

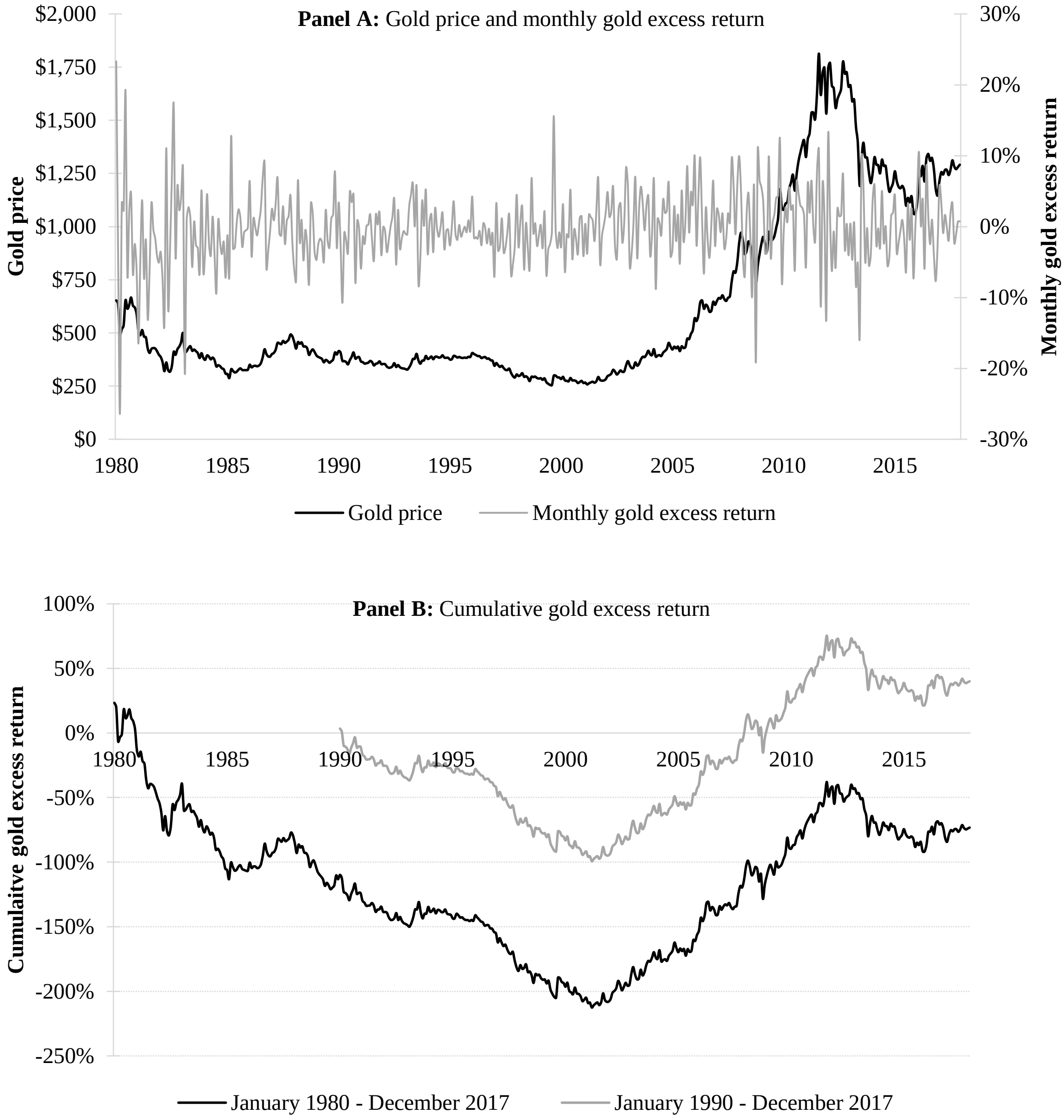

Figure 1. Panel A shows the gold price in USD (black line, left axis) and monthly gold excess returns (grey line, right axis) over the sample period from January 1980 to December 2017. Panel B depicts the cumulative gold excess return over the sample period from January 1980 to December 2017 (black line) and over the evaluation period from January 1990 to December 2017 (grey line).

Tables

Table 1
Overview of market timing signals

Model	Description
Seasonal market timing signals	
SEA*XXXX*	Seasonal allocation models are consecutively numbered. The monthly allocations of all documented seasonal market timing strategies are listed in Appendix 1.
Technical market timing signals	
MA*s*-*l*	Moving-average indicator, calculated as the difference between short-term (*s*) and long-term (*l*) moving averages of the gold price index P_t: $MAs\text{-}l_t = \begin{cases} 1 \; if \; MA_{s,t} \geq MA_{l,t} \\ 0 \; if \; MA_{s,t} < MA_{l,t} \end{cases}$ for $s = \{1,2,3\}$ and $l = \{9,12\}$ where $MA_{j,t} = \frac{1}{j} \sum_{i=0}^{j-1} P_{t-i}$ for $j = s, l$.
MOM*m*	Momentum-indicator, calculated as the difference between the current gold price index and the gold price index m months ago: $MOMm = \begin{cases} 1 \; if \; P_t \geq P_{t-m} \\ 0 \; if \; P_t < P_{t-m} \end{cases}$ for $m = \{9,12\}$.
Fundamental market timing signals	
DFR	Predictive regression (eq. (1)) based on default return spread (long-term corporate bond return minus the long-term government bond return).
DFY	Eq. (1) based on default yield spread (difference between Moody's BAA- and AAA-rated corporate bond yields).
DY	Eq. (1) based on dividend yield (log of a twelve-month moving sum of dividends paid on the S&P 500 index, minus the log of lagged stock prices (S&P 500 index)).
ERP	Eq. (1) based on equity risk premium (difference between the log return on the S&P 500 index (including all dividends) and the log return on a risk-free bill).
INFL	Eq. (1) based on inflation rate (calculated from the Consumer Price Index (CPI) for all urban consumers, lagged by one month).
LTR	Eq. (1) based on return on long-term government bonds.
LTY	Eq. (1) based on long-term government bond yield.
RVOL	Eq. (1) based on volatility of the equity risk premium (based on a twelve-month moving standard deviation estimator).
TBL	Eq. (1) based on interest rate on a three-month Treasury bill.
TMS	Eq. (1) based on term spread (long-term yield minus Treasury bill rate).
USD	Eq. (1) based on USD exchange rate (continuously compounded year-on-year change in trade-weighted effective nominal U.S. exchange rate, lagged by one month).
MEAN	Mean combination forecast.
KSF	Kitchen sink forecast.
PC1F	Forecast based on first principal component.
PC2F	Forecast based on first and second principal component.

Table 2

Descriptive statistics and simple tests for weak-form market efficiency

This table reports descriptive statistics (panel A) as well as the results of simple tests for weak-form market efficiency (panels B, C, and D) for monthly gold excess returns over the evaluation period (January 1990 to December 2017) and the full sample period (January 1980 to December 2017). All calculations are based on monthly log returns. The mean return and standard deviation in panel A are not annualized. Mean returns are multiplied by 100, i.e. in percent per month. The kurtosis is adjusted so that the normal distribution exhibits a kurtosis of zero (excess kurtosis). The Jarque-Bera test statistic tests the null hypothesis that the monthly gold excess returns are normally distributed. AC(1) in panel B denotes autocorrelation in the excess returns at lag 1. The statistical significance of autocorrelation is tested using the Ljung-Box test. Q(·) denotes the Q-test statistic that tests for an autocorrelation up to the lag shown in brackets. Panel C gives the test statistics of the runs test, where the definition of a run is based on three different cutoff values: mean, median, and zero. Panel D gives the test statistics of the rank-based (R1) and sign-based (S1) variance-ratio tests developed by Wright (2000) for three holding periods as shown in brackets (3, 12, and 36 months). *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

	1990:01-2017:12		**1980:01-2017:12**	
Panel A: Descriptive statistics				
Mean	0.12		-0.16	
Std.	4.53		5.17	
Min	-19.19		-26.39	
Max	15.60		23.33	
Skewness	-0.07		-0.10	
Kurtosis	1.37		3.40	
Jarque-Bera	26.50	***	220.00	***
Panel B: Autocorrelations				
AR(1)	-0.11		-0.08	
Q(1)	3.89	**	2.84	*
Q(3)	5.58		5.92	
Q(12)	21.06	**	18.69	*
Q(36)	57.36	**	43.60	
Panel C: Run tests				
Cut-off: Mean	0.24		-0.19	
Cut-off: Median	0.22		-0.19	
Cut-off: Zero	0.00		-0.53	
Panel D: Wright (2000) tests				
R1 (3)	-1.56		-1.27	***
R1 (12)	0.27		1.36	*
R1 (36)	2.72	***	2.55	**
S1 (3)	-0.51		0.06	
S1 (12)	1.61	*	2.27	**
S1 (36)	4.01	***	3.37	***

Table 3

Traditional back-test results

This table lists the performance of the five best market timing strategies for each set of signals based on monthly seasonalities, technical indicators, and fundamental factors (refer to Table 1 for a description), as well as the benchmark strategy over the evaluation period from January 1990 to December 2017. Four performance measures are used: mean directional accuracy (panel A), mean absolute returns (panel B), Sharpe ratios (panel C) and Sortino ratios (panel D). The mean absolute returns are multiplied by 100, i.e., in percent per month.

Set of signals	***Panel A***		***Panel B***			***Panel C***		***Panel D***	
	Directional accuracy		*Absolute returns*			*Risk-adjusted excess returns*		*Downside risk-adjusted excess returns*	
	Strategy	**Mean**	**Strategy**	**Mean**	**St.dev.**	**Strategy**	**Mean**	**Strategy**	**Mean**
Seasonal signals	SEA0656	57.74%	SEA2313	0.574	3.264	SEA4010	0.114	SEA4010	0.216
	SEA1378	57.74%	SEA3190	0.571	3.066	SEA3761	0.113	SEA3756	0.203
	SEA1433	57.74%	SEA1337	0.569	3.509	SEA3190	0.111	SEA3761	0.200
	SEA2343	57.74%	SEA2302	0.567	3.326	SEA3756	0.110	SEA3190	0.194
	SEA0213	57.14%	SEA3201	0.547	2.989	SEA3201	0.106	SEA3201	0.184
Technical signals	MA1-36	56.55%	MA1-48	0.553	3.569	MA1-48	0.090	MA1-48	0.138
	MOM24	56.55%	MOM24	0.541	3.608	MOM24	0.086	MOM24	0.130
	MA1-48	56.25%	MA1-36	0.530	3.630	MA1-36	0.082	MA1-36	0.124
	MA2-12	55.95%	MOM12	0.499	3.681	MOM12	0.073	MA2-12	0.108
	MOM12	55.95%	MA1-24	0.479	3.705	MA2-12	0.070	MOM12	0.108
Fundamental signals	TBL	58.63%	KSF	0.544	3.450	KSF	0.090	KSF	0.141
	DFY	56.85%	DFY	0.498	3.671	DFR	0.076	DFR	0.121
	TMS	56.85%	DFR	0.494	3.861	DFY	0.073	DFY	0.108
	USD	56.55%	TBL	0.472	3.632	TBL	0.068	TBL	0.099
	ERP	55.95%	LTY	0.462	3.065	TMS	0.067	TMS	0.096
Benchmark strategy	Random walk	49.40%	Buy-and-hold	0.350	4.515	Buy-and-hold	0.026	Buy-and-hold	0.038

Table 4
Tests for superior predictive ability

This table reports the results of the SPA-test, the step-SPA-test, the step-SPA(*3*)-test, and the FDP-SPA-test (with the false discovery proportion (FDP) not exceeding $\boldsymbol{\gamma = 10\%}$) for the pre-specified error rate $\boldsymbol{\alpha = 5\%}$ for the performance of market timing strategies based on monthly seasonalities, technical indicators, and fundamental factors (refer to Table 1 for a description) compared to the benchmark model over the evaluation period from January 1990 to December 2017. Four performance measures are used: mean directional accuracy (panel A), mean absolute returns (panel B), Sharpe ratios (panel C) and Sortino ratios (panel D). Column (1) indicates the set of signals the data snooping tests are applied to. The "best" strategy is the market timing strategy with the highest performance measure. The nominal *p*-value ignores the search over all strategies that preceded the selection of the strategy being compared to the benchmark, i.e., it does not account for the entire set of market timing strategies. In addition, we report the consistent *p*-value of the SPA-test, as well as its lower and upper bounds. Finally, we state the number of outperforming strategies identified by the step-SPA-, the step-SPA(*3*)- and the FDP-SPA-tests. "-" indicates that there is no outperforming strategy at the 5% level of significance.

(1)	(2)	(3)	(4)	(5)	(6)
Set of signals	**Benchmark**	**Best strategy**	**Nominal *p*-value**	**SPA *p*-value** **[lower bound; upper bound]**	**# of strategies identified by** **[step-SPA; step-SPA(*3*); FDP-SPA]**
***Panel A**: Mean directional accuracy*					
Seasonal signals	Random walk	SEA0656	0.0262	0.4301 [0.3902; 0.4322]	[-; -; -]
Technical signals	Random walk	MA1-36	0.0376	0.1289 [0.1163; 0.1289]	[-; -; -]
Fundamental signals	Random walk	TBL	0.0126	0.0612 [0.0612; 0.0612]	[-; 5; -]
***Panel B**: Mean absolute returns*					
Seasonal signals	Buy-and-hold	SEA2313	0.0868	0.7458 [0.5833; 0.7565]	[-; -; -]
Technical signals	Buy-and-hold	MA1-48	0.0733	0.2633 [0.2170; 0.2633]	[-; -; -]
Fundamental signals	Buy-and-hold	KSF	0.1018	0.3467 [0.3467; 0.3467]	[-; -; -]

(continued)

Table 4 *(continued)*

Set of signals	**Benchmark**	**Best strategy**	**Nominal *p*-value**	**SPA *p*-value** **[lower bound; upper bound]**	**# of strategies identified by** **[step-SPA; step-SPA(*3*); FDP-SPA]**
***Panel C**: Sharpe ratios*					
Seasonal signals	Buy-and-hold	SEA4010	0.0431	0.7634 [0.5918; 0.7809]	[-; -; -]
Technical signals	Buy-and-hold	MA1-48	0.0212	0.1178 [0.1022; 0.1320]	[-; 1; -]
Fundamental signals	Buy-and-hold	KSF	0.0372	0.2070 [0.2070; 0.2070]	[-; -; -]
***Panel D**: Sortino ratios*					
Seasonal signals	Buy-and-hold	SEA4010	0.0227	0.6304 [0.4801; 0.6432]	[-; -; -]
Technical signals	Buy-and-hold	MA1-48	0.0178	0.1178 [0.0868; 0.1116]	[-; 1; -]
Fundamental signals	Buy-and-hold	KSF	0.0303	0.1690 [0.1690; 0.1690]	[-; -; -]

Table 5
Sub-sample analyses

This table reports the results of step-SPA-test, the step-SPA(*3*)-test and the FDP-SPA-test (with the false discovery proportion (FDP) not exceeding $\boldsymbol{\gamma = 10}$%) for the pre-specified error rate $\boldsymbol{\alpha = 5\%}$ for the performance of market timing strategies based on monthly seasonalities, technical indicators, and fundamental factors (refer to Table 1 for a description) compared to the benchmark model over three sub-sample periods: January 1990 to December 2011, January 2001 to December 2012, and January 2013 to December 2017. Four performance measures are used: mean directional accuracy (panel A), mean absolute returns (panel B), Sharpe ratios (panel C) and Sortino ratios (panel D). We state the number of outperforming strategies identified by the step-SPA-, the step-SPA(*3*)- and the FDP-SPA-tests. "-" indicates that there is no outperforming strategy at the 5% level of significance.

Set of signals	**Benchmark**	**# of strategies identified by [step-SPA; step-SPA(*3*); FPD-SPA]**		
		1990-2001	**2002-2012**	**2013-2017**
***Panel A**: Mean directional accuracy*				
Seasonal signals	Random walk	[-; -; -]	[834; 1097;]	[-; -; -]
Technical signals	Random walk	[-; -; -]	[16; 16; 16]	[-; -; -]
Fundamental signals	Random walk	[-; -; -]	[18; 18; 18]	[-; -; -]
***Panel B**: Mean absolute returns*				
Seasonal signals	Buy-and-hold	[27; 55; 118]	[-; -; -]	[-; 2; -]
Technical signals	Buy-and-hold	[-; -; -]	[-; -; -]	[-; -; -]
Fundamental signals	Buy-and-hold	[15; 15; 15]	[-; -; -]	[-; -; -]
***Panel C**: Sharpe ratios*				
Seasonal signals	Buy-and-hold	[1; 7; 1]	[-; -; -]	[-; 3; -]
Technical signals	Buy-and-hold	[-; -; -]	[-; -; -]	[-; -; -]
Fundamental signals	Buy-and-hold	[2; 2; 2]	[-; -; -]	[-; -; -]
***Panel D**: Sortino ratios*				
Seasonal signals	Buy-and-hold	[-; -; -]	[-; -; -]	[1; 2; 1]
Technical signals	Buy-and-hold	[-; -; -]	[-; -; -]	[-; -; -]
Fundamental signals	Buy-and-hold	[-; 2; -]	[-; 2; -]	[-; -; -]

References

Adcock, C., N. Areal, M. R. Armada, M. C. Cortez, B. Oliveira, and F. Silva, 2014, New Tests of Correlation and the Choice of Measures of Portfolio Performance, Working Paper, Sheffield University Management School.

Aggarwal, R., and B.M. Lucey, 2007, Psychological Barriers in Gold Prices, *Review of Financial Economics* 16, 217-230.

Aggarwal, R., B.M. Lucey and F.A. O'Connor, 2014, Rationality in Precious Metals Forward Markets: Evidence of Behavioural Deviations in the Gold Markets, *Journal of Multinational Financial Management* 25-26, 110-130.

Auer, B.R., 2016, On the Performance of Simple Trading Rules Derived from the Fractal Dynamics of Gold and Silver Price Fluctuations, *Finance Research Letters* 16, 255-267.

Batten, J.A., B.M. Lucey, F.J. McGroarty, M. Peate, and A. Urquhart, 2015, Does Technical Analysis Beat the Market? Evidence from High Frequency Trading in Gold and Silver.

Baur, D.G., 2013, The Autumn Effect of Gold, *Research in International Business and Finance* 27, 1–11.

Baur, D.G., J. Beckmann, and R. Czudaj, 2016, A Melting Pot – Gold Price Forecasts under Model and Parameter Uncertainty, *International Review of Financial Analysis* 48, 282–291.

Baur, D.G., T. Dimpfl, and K. Kuck, 2018, Bitcoin, Gold and the US Dollar – A Replication and Extension, *Finance Research Letters* 25, 103-110.

Baur, D.G., and K.J. Glover, 2015, Speculative Trading in the Gold Market, *International Review of Financial Analysis* 39, 63–71.

Baur, D.G., and B.M. Lucey, 2010, Is Gold a Hedge or a Safe Haven? An Analysis of Stocks, Bonds, and Gold, *Financial Review* 45, 217–229.

Beckmann, J., and R. Czudaj, 2013, Gold as an Inflation Hedge in a Time-varying Coefficient Framework, *North American Journal of Economics and Finance* 24, 208–222.

Białkowski, J., M.T. Bohl, P.M. Stephan, and T.P. Wisniewski, 2015, The Gold Price in Times of Crisis, *International Review of Financial Analysis* 41, 329–339.

Blose, L.E., 2010, Gold Prices, Cost of Carry, and Expected Inflation, *Journal of Economics and Business* 62, 35–47.

Bruno, S., and L. Chincarini, 2010, A Historical Examination of Optimal Real Return Portfolios for Non-US Investors, *Review of Financial Economics* 19, 161–178.

Capie, F., T.C. Mills, and G. Wood, 2005, Gold as a Hedge against the Dollar, *Journal of International Financial Markets, Institutions and Money* 15, 343–352.

Cenesizoglu, T., and A. Timmermann, 2012, Do Return Prediction Models Add Economic Value?, *Journal of Banking & Finance* 36, 2974–2987.

Charles, A., O. Darné, and J.H. Kim, 2015, Will Precious Metals Shine? A Market Efficiency Perspective, *International Review of Financial Analysis* 41, 284–291.

Chevallier, J., M. Gatumel, and F. Ielpo, 2013, Understanding Momentum in Commodity Markets, *Applied Economics Letters* 20, 1383–1402.

Chordia, T., A. Goyal, and A. Saretto, 2017, p-Hacking: Evidence from two Million Trading Strategie, Working paper, Swiss Finance Research Paper Series 17–37.

Cochrane, J., 1999, Portfolio Advice in a Multifactor World, *Economic Perspective*, Federal Reserve Bank of Chicago 23, 59-78.

Dichtl, H., 2017, Forecasting Excess Returns of the Gold Market: Can we learn from the Stock Market Predictions?, Working Paper.

Dichtl, H., and W. Drobetz, 2014, Are Stock Markets Really so Inefficient? The Case of the "Halloween Indicator", *Finance Research Letters* 11, 112–121.

Dichtl, H., and W. Drobetz, 2015, Sell in May and Go Away: Still Good Advice for Investors, *International Review of Financial Analysis* 38, 29–43.

Emmrich, O., and F.J. McGroarty, 2013, Should Gold Be Included in Institutional Investment Portfolios?, *Applied Financial Economics* 23, 1553–1565.

Ferson, W., and C. Harvey, 1993, The Risk and Predictability of International Equity Returns, Review of Financial Studies 6, 527–66

Ferson, W., and C. Harvey, 1994, Sources of Risk and Expected Returns in Global Equity Markets, *Journal of Banking and Finance* 18, pages 775–803.

Giacomini, R., and H. White, 2006, Tests of Conditional Predictive Ability, *Econometrica* 74, 1545–1578.

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

Hansen, P.R., 2005, A Test for Superior Predictive Ability, *Journal of Business & Economic Statistics* 23, 365–380.

Harvey, C.R., 2017, Presidential Address: The Scientific Outlook in Financial Economics, *Journal of Finance* 72, 1399–1440.

Hillier, D., P. Draper, and R. Faff, 2006, Do Precious Metals Shine? An Investment Perspective, *Financial Analysts Journal* 62, 98–106.

Holm, S., 1979, A Simple Sequentially Rejective Multiple Test Procedure, *Scandinavian Journal of Statistics* 6, 65–70.

Hsu, P.-H., Y.-C. Hsu, and C.-M. Kuan, 2010, Testing the Predictive Ability of Technical Analysis Using a New Stepwise Test without Data Snooping Bias, *Journal of Empirical Finance* 17, 471–484.

Hsu, Y.-C., C.-M. Kuan, and M.-F. Yen, 2014, A Generalized Stepwise Procedure with Improved Power for Multiple Inequalities Testing, *Journal of Financial Econometrics* 12, 730–755.

Jastram, R.W., and J. Leyland, 2009, *The Golden Constant*, Edward Elgar Publishing Ltd., Cheltenham, United Kingdom.

Jacobsen, B., and N. Visaltanachoti, 2009, The Halloween Effect in U.S. Sectors, *Financial Review* 44, 437–459.

Klein, T., H.P. Thu, and T. Walther, 2018, Bitcoin is not the New Gold – A Comparison of Volatility, Correlation, and Portfolio Performance, *International Review of Financial Analysis* 59, 105-116.

Leitch, G., and J.E. Tanner, 1991, Economic Forecast Evaluation: Profits Versus the Conventional Error Measures, *American Economic Review* 81, 580–590.

Lo, A.W., and A.C. MacKinlay, 1990, Data-Snooping Biases in Tests of Financial Asset Pricing Models, *Review of Financial Studies* 3, 431–467.

Lucey, B.M., V. Poti, and E. Tully, 2006, International Portfolio Formation, Skewness and the Role of Gold, Working Paper.

Lucey, M.E., and F.A. O'Connor, 2016, Mind the Gap: Psychological Barriers in Gold and Silver Prices, *Finance Research Letters* 17, 135-140.

Ludvigson, S.C., and S. Ng, 2007, The Empirical Risk-Return Relation: A Factor Analysis Approach, *Journal of Financial Economics* 83, 171–222.

Ludvigson, S.C., and S. Ng, 2009, Macro Factors in Bond Risk Premia, *Review of Financial Studies* 22, 5027–5067.

Marshall, B.R., R.H. Cahan, and J.M. Cahan, 2008, Can Commodity Futures Be Profitably Traded with Quantitative Market Timing Strategies, *Journal of Banking & Finance* 32, 1810–1819.

McLean, R.D., and J. Pontiff, 2016, Does Academic Research Destroy Stock Return Predictability?, *Journal of Finance* 71, 5–31.

Miffre, J., and G. Rallis, 2007, Momentum Strategies in Commodity Futures Markets, *Journal of Banking & Finance* 31, 1863–1886.

Mihaylov, G., C.S. Cheong, and R. Zurbruegg, 2015, Can Security Analyst Forecasts Predict Gold Returns?, *International Review of Financial Analysis* 41, 237–246.

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

Mui, H.W., and C.W. Chu, 1993, Forecasting the Spot Price of Gold: Combined Forecast Approaches versus a Composite Forecast Approach, *Journal of Applied Statistics* 20, 13–23.

Narayan, P.K., S. Narayan, and S.S. Sharma, 2013, An Analysis of Commodity Markets: What Gain for Investors?, *Journal of Banking & Finance* 37, 3878-3889.

Naylor, M.J., U. Wongchoti, and H. Ith, 2014, Market Microstructure of Precious Metal ETFs, *Journal of Index Investing* 5, 48–56.

Nguyen, B, M. Prokopczuk, and C.W. Simen, 2017, The Risk Premium of Gold, Working paper, University of Hannover and University of Reading.

O'Connor, F.A., B.M. Lucey, J.A. Batten, and D.G. Baur, 2015, The Financial Economics of Gold - A Survey, *International Review of Financial Analysis* 41, 186–205.

Ornelas, H., F. Silva, and B. Fernandes, 2012, Yes, the Choice of Performance Measure Does Matter for Ranking of US Mutual Funds, *International Journal of Finance and Economics* 17, 61–72.

Pesaran, M.H., and A. Timmermann, 1995, Predictability of Stock Returns: Robustness and Economic Significance, *Journal of Finance* 50, 1201–1228.

Pierdzioch, C., M. Risse, and S. Rohloff, 2014, On the Efficiency of the Gold Market: Results of a Real-Time Forecasting Approach, *International Review of Financial Analysis* 32, 95–108.

Pierdzioch, C., M. Risse, and S. Rohloff, 2015, Forecasting Gold-Price Fluctuations: A Real-Time Boosting Approach, *Applied Economics Letters* 22, 46–50.

Politis, D.N., and J.P. Romano, 1994, Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions, *Annals of Statistics* 22, 2031–2050.

Popper, N., 2016, Digital Gold: Bitcoin and the Inside Story of the Misfits and Millionaires Trying to Reinvent Money, *HarperCollins Publisher*, New York, NY.

Qi, M., and W. Wang, 2013, The Monthly Effects in Chinese Gold Market, *International Journal of Economics and Finance* 5.

Rapach, D., and G. Zhou, 2013, Forecasting Stock Returns, in: Elliott, G., and A. Timmermann, *Handbook of Economic Forecasting* Volume 2A, Elsevier B.V., Amsterdam.

Reboredo, J.C., 2013, Is Gold a Safe Haven or a Hedge for the US Dollar? Implications for Risk Management, *Journal of Banking & Finance* 37, 2665–2676.

Schwert, G.W., 2003, Anomalies and market efficiency, in: Constantinides, G.M., M. Harris, and R.M. Stulz (eds.), Handbook of the Economics of Finance, North-Holland, Amsterdam: Elsevier Science, 939–974.

Sharpe, W.F., 1994, The Sharpe Ratio, *Journal of Portfolio Management* 21, 49–58.

Sjaastad, L.A., and F. Scacciavillani, 1996, The Price of Gold and the Exchange Rate, *Journal of International Money and Finance* 15, 879–897.

Sortino, F.A., and L.N. Price, 1994, Performance Measurement in a Downside Risk Framework, *Journal of Investing* 3, 59-64.

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

Sullivan, R., A. Timmermann, and H. White, 2001, Dangers of Data Mining: The Case of Calendar Effects in Stock Returns, *Journal of Econometrics* 105, 249–286.

Szakmary, A.C., Q. Shen, and S.C. Sharma, 2010, Trend-following Trading Strategies in Commodity Futures: A Re-examination, *Journal of Banking & Finance* 34, 409–426.

Tibshirani, R., 1996, Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological) 58, 267–288.

Wright, J.H., 2000, Alternative Variance-Ratio Tests Using Ranks and Signs, *Journal of Business & Economic Statistics* 18, 1–9.

Zakamouline, V., 2011, The Performance Measure You Choose Influences the Evaluation of Hedge Funds, *The Journal of Performance Measurement* 15, 48–64.

Appendix

Appendix 1
Monthly signals of all documented seasonal market timing strategies

Strategy	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
SEA0213	1	1	0	1	1	1	1	1	1	0	1	0
SEA0656	1	1	0	1	1	1	0	1	1	0	1	0
SEA1337	1	1	0	0	0	0	1	1	1	0	1	1
SEA1378	1	1	0	0	1	1	0	1	1	0	1	0
SEA1433	1	1	0	1	1	0	0	1	1	0	1	0
SEA2302	1	1	0	0	0	0	0	1	1	0	1	1
SEA2313	1	1	0	0	0	0	1	1	1	0	1	0
SEA2343	1	1	0	0	1	0	0	1	1	0	1	0
SEA3190	1	1	0	0	0	0	0	1	1	0	1	0
SEA3201	1	1	0	0	0	0	1	1	1	0	0	0
SEA3756	1	1	0	0	0	0	0	0	1	0	1	0
SEA3761	1	1	0	0	0	0	0	1	1	0	0	0
SEA4010	1	1	0	0	0	0	0	0	1	0	0	0