Direction-of-change forecasting in commodity futures markets

Liu, J., Papailias, F., & Quinn, B. (2021). Direction-of-change forecasting in commodity futures markets. *International Review of Financial Analysis*, *74*, Article 101677. https://doi.org/10.1016/j.irfa.2021.101677

Published in:
International Review of Financial Analysis

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Direction-of-Change Forecasting in Commodity Futures Markets*

Jiadong Liu[a,*], Fotis Papailias[b], Barry Quinn[a]

[a] *Queen's Management School, Queen's University Belfast, UK*
[b] *King's Business School, King's College London, UK*

Abstract

This paper examines direction-of-change predictability in commodity futures markets using a variety of binary probabilistic techniques. As well as traditional techniques, we apply Variable Length Markov Chain (VLMC) analysis, an innovative technique popularised in computational biology when predicting DNA sequences (Bühlmann et al., 1999). To the best of our knowledge, this is the first application of VLMC in finance. Our results show that both VLMC and technical analysis methods provide strong predictability of the direction-of-change of commodity returns, with annualised mean returns of approximately 8%, substantially higher than the passive long strategy. Our results suggest that a short-term learning effect is present in commodities market which can be exploited using innovative direction-of-change forecasting techniques.

Keywords: Forecasting commodity futures, Direction-of-change, Dynamic probit model, Variable Length Markov Chain, Return signal momentum
JEL: G11, G12, G13

*We are grateful to the associate editor and two anonymous referees for their insightful comments and suggestions.

*Corresponding author. Queen's Management School, Queen's University Belfast, Riddel Hall, 185 Stranmillis Road, BT9 5EE, UK. email: liu.jiadong@qub.ac.uk.

1. Introduction

There are a large number of studies in the asset pricing literature that seek to predict market movements. However, most focus on forecasting an exact future return (level forecasting), see, e.g., Ferreira and Santa-Clara (2011), Pettenuzzo et al. (2014) and Joseph et al. (2011). The best models are usually selected based on minimising forecast errors such as Mean Absolute Forecast Error (MAFE) and Root Mean Squared Forecast Error (RMSFE). In contrast, financial decisions are primarily made based on maximising profitability or risk-adjusted profitability, a goal that does not correspond directly to maximising forecast accuracy, see, e.g., (Leitch and Tanner, 1991; Brooks et al., 2001). The immediate focus of investors is whether they should buy or sell the asset in advance of the next trading period, i.e., the trading signals. In this sense, direction-of-change forecasting[1], which estimates the probability of the next period being a particular return sign, can be more informative than level forecasting.

This paper examines direction-of-change predictability in commodity futures markets using a variety of approaches[2]. These methods have different economic or data-driven rationales to model direction-of-change dependence. First, we use a Return Signal Momentum (RSM) strategy introduced by Papailias et al. (2017), which simply assumes that the current return sign is dependent on past signs. We extend RSM to an exponential setting and call it ERSM, where the more recent observations are given a higher weighting. Second, we employ the dynamic probit model of Nyberg (2011) which has been proven to be successful in forecasting stock market direction-of-change [3]. Finally, we apply the Variable Length Markov Chain (VLMC), which has

[1]Evidence of direction-of-change predictability in financial data is seen in Pesaran and Timmermann (1995), Gencay (1998), Leung et al. (2000), Hong and Chung (2003), Christoffersen and Diebold (2006), Liu and Kemp (2019), among others.

[2]Investing in commodity futures markets has gained increasing popularity in recent years amongst institutional investors. According to the World Federation of Exchanges, global investment in commodity futures has increased more than 10 times between 2005 and 2018.

[3]The probability models are one of the standard methods to forecast binary outcomes such as the stock return signs, see, e.g., Anatolyev and Gospodinov (2010), Rydberg and Shephard (2003), and Kauppi and Saikkonen (2008).

been shown to successfully forecast categorical data series such as DNA sequences, see Bühlmann et al. (1999). To the best of our knowledge, this is the first application of this innovative binary forecasting approach in finance. The VLMC model is ideal for forecasting sequences of behavioural events in financial markets, otherwise hidden from traditional techniques. When the forecasting problem is re-framed as the direction-of-change in financial asset returns, it offers an convenient approach to capture repeated patterns in discrete and categorical financial time series.

In this study, we ask a fundamental question: What drives the direction-of-change in commodity futures markets? If the RSM class models successfully capture this direction-of-change, it suggests that there exists a momentum effect causing the persistence in return signs in the short-term. If the binary probability models uncover predictability, this means that the market movement is in line with the market states and macroeconomic conditions. Finally, if VLMC models are successful, then we conclude that investors tend to use past market movements as a reference. In other words, there is a learning effect in futures markets driven by the past data with the pattern repeating as a result. Our results help develop the understanding of the fundamental forces underpinning commodity markets.

Based on a dataset that includes weekly returns of 24 commodity futures indices from January, 1971 to August, 2018, we compare the forecast accuracy and strategy performance all models. As a further robustness check, we conduct both rolling and recursive window approaches based on different in-sample periods ranging from 26 weeks to 780 weeks (15 years). We first specify each model according to their in-sample goodness of fit. Then, we examine the forecast error and strategy performance using out-of-sample data.

Finally, we apply the Model Confidence Set (MCS) procedure to establish robust out of sample performance (Hansen et al., 2011). This framework allows us to compare multiple candidate models simultaneously and eliminating those that are significantly inferior to the rest by rejecting the null hypothesis of *Equal Predictive Ability* (EPA).

Our empirical results highlight that both the RSM filter and the VLMC model exhibit strong direction-of-change predictability. The two best out-of-sample per-

formers are RSM and the bootstrapped VLMC model specified with an in-sample window of 52 weeks. In general, models with short-term in-sample windows (26-104 weeks) outperform models with long-term in-sample windows (3-15 years). By contrast, the dynamic probit models do not show superior return sign predictability with success rates only slightly higher than 50% across a variety of in-sample periods. The MCS procedure results confirm our findings by ranking the 52 weeks RSM and 52 weeks in sample window bootstrapped VLMC model as the top two.

To highlight the economic benefits of adopting the approaches, we perform trading strategies based on the signals generated by the models. Our empirical trading results highlight that the RSM and VLMC models result in superior absolute return and Sharpe ratios. In particular, the RSM 52 weeks strategy has an average annualised portfolio return of 7.8% ($t = 4.45$), while the bootstrapped VLMC with 52 weeks model has an average return of 7.7% ($t = 4.50$)[4]. Although the dynamic probit models generate statistically significant positive returns in some settings, these are much lower in magnitude than the profits arising from the use of the RSM and VLMC approaches.

An interesting insight of this paper is the innovative application of the VLMC model in direction-of-change forecasting of commodity markets. In financial time series analysis, it is common to model market states instead of level forecasting. For example, Brooks et al. (2015) used a switching regression to model the booms and busts in commodity markets; Pettenuzzo et al. (2014) predicts the size and duration of breaks (state changes) via a hierarchical hidden Markov Chain model. In our study, we simply employ the direction-of-change in prices as a measure of market movements. Our findings emphasise the benefits of the data driven machine learning approach in predicting commodity return signs. In comparison with the extant models used as benchmarks in this paper, our research constitutes the first time the VLMC models show its value in this context.

A possible explanation for the VLMC uncovering predictability in commodity

[4]The t-statistics reported are calculated based on a Newey–West standard error (Newey and West, 1986)

markets is that commodity prices are partly driven by supply side shocks[5]. For example, the crude oil price is largely determined by oil supply from OPEC countries; agricultural commodity price is partly determined by supply and storage in different products. Any cyclicalities caused by supply side shocks can lead to learning effects in commodity price movements. In light of the recent advance in machine learning techniques and their implication to empirical asset pricing and market micro-structure (Gu et al., 2020; Easley et al., 2020), we propose that the VLMC should be adopted as an effective approach and uncover latent sequential patterns in commodity markets.

The remainder of this paper is organised as follows. Section 2 describes our commodity index dataset and the external variables used to construct the probability models. In Section 3, we discuss the three main approaches we employ in this study. We present results of the out-of-sample forecasting exercise in Section 4. Finally, Section 5 provides concluding remarks.

2. Data

We obtain daily closing prices for 24 constituents of the Standard and Poor's Goldman Sachs Commodity Index (S&P GSCI) from Datastream. Our dataset is similar to the sample used in Bianchi et al. (2015) and Koijen et al. (2018). The constituent commodities are classified into five categories: energy, precious metals, live stock, industrial metals and agricultural; and are listed on the ICE, NYMEX, COMEX, CME, LME, CBOT and KCBT exchanges. The benefit of using the S&P GSCI indices is that they reflect the real returns of investing in the most liquid futures that is traded on different exchanges. Moreover, these indices are continuous price series and not subject to the issue of rolling contracts, where one has to shift a future contract over to another on expiry. Our sample period spans from January 1971 to August 2018.

For each price series, we calculate the daily percentage returns and aggregate

[5]Supply side shocks are often found to be an important factor in determining commodity prices, see, e.g., Baumeister and Peersman (2013) in oil market, Shafiee and Topal (2010) in gold market, and Roberts and Schlenker (2013) in agriculture commodities.

these daily returns into weekly returns, similar to the approach in Moskowitz et al. (2012). Weekly data is used in comparison to daily data, as we expect to observe stronger sign dependence, based on Christoffersen and Diebold (2006) who find that sign dependence is weak in high data frequency.

Table 1 summarises the descriptive statistics for each commodity index. The sector, listed exchange, annualised arithmetic mean, standard deviation, skewness and kurtosis of the weekly returns of each index are presented. The long-term return for different commodities varies across sectors with most of energy futures being relatively high and agriculture futures returns being low. Given the speculative nature of commodity futures markets, the volatility associated with them is high relative to other assets.

Our second approach, the dynamic probability model, requires a set of explanatory variables to model the probability of positive returns. Following Leung et al. (2000) and Nyberg (2011), we calculate weekly returns of the MSCI world index as a proxy of the general market state, the U.S. Dollar Index, the NBER-based recession indicator controlling for crisis periods, the U.S. short-term (3-month) and long-term (10-year) treasury yield. As our data contains commodity futures, we include the S&P GSCI commodity index as another important control variable. They data was sourced from Bloomberg for the period January 1971 to August 2018.

3. Binary Forecasting Approaches and Model Specification

Unlike traditional returns forecasting models which estimate the level of returns, direction-of-change dependence models usually focus on estimating the probability of an event (i.e. the occurrence of a positive or negative return). In this section, we discuss different approaches used to forecast such binary time series. Although these methods are substantially different in terms of their mechanics, they aim to estimate the same output, the probability of a positive return in the next period. Consider a binary time series $\{X_t\}$, where the value of each variable x_t is either 1 for a positive return or 0 for a negative return. Then, the probability of having a positive return at time $t+1$ is given by: $P_{t+1} = Pr(x_{t+1} = 1|\Omega_t)$, where Ω_t is the information set including the lagged value of x_{t+1} and any explanatory variables we need until time

t. Based on forecasting the probability of the direction-of-change, these models are able to predict market movements, and hence, investors can make trading decisions using these models. Furthermore, we highlight the best specifications for each type of these approaches.

3.1. Return Signal Momentum

RSM is a non-parametric simple moving average smoother for estimating the probability of return signs, which has been proven to be able to predict return signs by Papailias et al. (2017). We calculate the probability of positive return signs, P, over the past k periods from time $t-k+1$ to t for a single instrument using a simple moving average method. Mathematically, this probability measures the average frequency of positive signs over the look-back period k. The RSM probability equation is given by:

$$P_{t+1} = \frac{1}{k} \sum_{i=t-k+1}^{t} x_i, \tag{1}$$

where x_i denotes the binary variable which takes a value of 1 when the return is positive and 0 otherwise.

To extend the studies of RSM, we estimate an exponential return signal moving average (ERSM). By applying an exponential moving average, ERSM assumes that the recent return signs have a greater impact on the present than earlier ones. Hence, the weight for each past return sign, w_i is a function of the smoothing parameter α, which is calculated as follows:

$$w_i = \alpha(1-\alpha)^{t-i}, \tag{2}$$

where $i \in \{t-k+1, t-k+2, \ldots, t-1, t\}$. α is specified as $\alpha = 2/(k+1)$, which accounts for approximately 86% of the weight[6]. Hence, the estimated probability of

[6]Adjusting α leads to qualitatively similar in-sample results. These are available upon request.

ERSM is the sum of product of weight w_i and binary variable x_i:

$$P_{t+1}^{E} = \sum_{i=t-k+1}^{t} w_i x_i. \tag{3}$$

3.2. Dynamic Probit Model

Next, we consider another type of binary choice model, namely a dynamic probability model. Leung et al. (2000) and Nyberg (2011) show how it can be used to successful predict stock markets. However, such a method has yet to be applied to commodity markets. Given the recent financialisation of commodity markets (Cheng and Xiong, 2014; Tang and Xiong, 2012), commodity assets are more correlated with financial assets and macroeconomic conditions. Therefore, we believe this approach is appropriate in forecasting the direction-of-change for commodity assets. More specifically, we employ a recent improved approach called the dynamic probit model that uses lagged dependent binary variables to predict its future value under the Kauppi and Saikkonen (2008) framework. This model captures market movements better as it incorporates an autoregressive component, accounting for the serial dependency in return signs (Nyberg, 2011; Pönkä, 2017).

The dynamic probit model assumes that the binary time series $\{X_t\}$ follows a Bernoulli distribution as $X_{t+1}|\Omega_t \sim B(P_{t+1})$. The probability P_{t+1} can be expressed as a standard normal cumulative distribution function as follows:

$$P_{t+1} = \Phi(\pi_{t+1}), \tag{4}$$

where π_{t+1} is the latent variable which is linearly related to the information set Ω_t.

In this study, the dynamic framework proposed by Nyberg (2011) is employed, where the lagged binary time series x_t itself is included in the information set Ω_t. Hence, the dynamic probit model is specified by estimating the linear function of π_{t+1}:

$$\pi_{t+1} = w + \delta_1 x_t + D_t'\beta, \tag{5}$$

where w is the constant, x_t denotes the lagged binary dependent variable, and D_t'

is the vector of the explanatory variables. The parameters in the dynamic probit models are estimated using maximum likelihood.

Next, we fit dynamic probit models to test the in-sample predictability of the probability P_t. The control variables we use are the returns of the MSCI world index, the S&P GSCI commodity index and the U.S Dollar Index. These variables are considered together with the first difference of short-term (3-month), long-term (10-year) US Treasury yield, and the NBER recession indicator. We also include the lagged binary time series of the return signs x_{t-1} to form the dynamic models as in Nyberg (2011).

Several model selection criteria are employed to evaluate the model performance, including the Schwarz information criterion, BIC (Schwarz et al., 1978) and the pseudo-R^2 (Estrella, 1998). Both criteria are calculated based on the maximum log-likelihood value. The BIC is a typical model selection criterion, while the pseudo-R^2 is a measure of goodness-of-fit by estimating the relationship between the maximum log-likelihood values of the candidate model itself and its constrained version.

Optimising our model specification on the above criteria, our results suggest that the dynamic model containing all the above mentioned variables has the best in-sample performance. The equation is expressed as follows:

$$\pi_{t+1} = w + \delta_1 x_t + \beta_1 MSCI + \beta_2 GSCI + \beta_3 USD + \beta_4 \Delta SI + \beta_5 \Delta LI + \beta_6 REC, \quad (6)$$

where ΔSI and ΔLI denote the short-term and long-term treasury yield, respectively. The probit model specified above has the highest pseudo-R^2 and the lowest BIC. The results are consistent when using a 26-week, 2-year, 5-year and 10-year sample period. These model specification results are available on request.

Table 2 summarises the results of the performance of the model specified in Equation 6 based on in-sample periods of 26 weeks, 52 weeks, two years and three years. The evaluation criteria including BIC, pseudo-R^2 and success rate SR are presented

for the dynamic probit regressions run on each of the 24 commodity indices[7]. We observe that the dynamic probit models show reduced predictability as the in-sample period increases. Across the 24 indices, the models using a 26 and 52-week estimation period exhibit higher pseudo-R^2 and lower BIC compared to the ones using longer in-sample period (2 and 3 years). The SR of the listed models also decreases as the sample period increases suggesting reduced in-sample predictability. Furthermore, predictability varies across different commodities. For example, in the results using 26 weeks in-sample estimation, energy (e.g., Gas oil, Heating oil and RBOB gas) and a few metal commodities (e.g., Platinum and Zinc) show strong in-sample predictability as reflected by high pseudo-R^2 and success rate SR and low BIC. By contrast, the model predictability of some agricultural commodities (e.g., Cocoa and Cotton) is weak.

3.3. Variable Length Markov Chain

The final approach we adopt is the VLMC model for stationary categorical time series (Bühlmann et al., 1999), which translates into the binary time series $\{X_t\}$ in our context. The VLMC model was originally designed to predict DNA sequences and, to the best of our knowledge, this is the first time that it has been used to forecast financial time series data. The similarity between the DNA sequence and the direction-of-change in asset returns is that they are both discrete and categorical[8]. It is a variation of the full Markov Chain model of finite order commonly used in economic decision making[9], however, it is more flexible and efficient (Bühlmann et al., 1999; Bejerano, 2004; Mächler and Bühlmann, 2004).

More specifically, it has two main advantages compared to a full Markov Chain model: (i) it is more suitable for fitting to data with high dimensional order, as the

[7]The success rate SR is simply the proportion of correct prediction in the in-sample period which is illustrated in Equation 9 of Section 4.

[8]Any financial time series data can have different market states which, indirectly, have an underlying relationship, e.g., stable, bullish and bearish. Asset returns can also be classified into multiple categories, e.g., $R_t < -2\%$ (I), $-2\% \leqslant R_t < 0$ (II), $0 \leqslant R_t < 2\%$ (III) and $R_t \geqslant 2\%$ (IV). However, we use the simplest and most intuitive way to classify returns into positive and negative.

[9]Application of Markov Chain models in economic decision making are seen in Hassan and Nath (2005) and Grégoir and Lenglart (2000), among others

dimension of VLMC does not increase exponentially as the order increases, and (ii) it results in a structurally richer model with memories of variable length through optimally pruning a tree-based algorithm. Therefore, VLMC is particularly useful for financial datasets as is computationally more efficient while reduces estimation variance.

The core feature of VLMC is its use of a context model. This allows VLMC to decipher the relevant information over the assumed infinite past of the process $\{X_t\}$. The model starts from a function called the preliminary context function $c_{pre}(\cdot)$. This function aims to find out the shortest possible past period λ which carries the relevant information while omitting the rest:

$$C_{pre} : x^0_{-\infty} \mapsto x^0_{-\lambda+1}, \tag{7}$$

where λ is determined by minimising the number of periods m which contains the relevant past information:

$$\lambda = \lambda(x^0_{-\infty}) = min\{m; Pr(X_1 = x_1|X^0_{-\infty} = x^0_{-\infty}) = Pr(X_1 = x_1|X^0_{-m+1} = x^0_{-m+1})\}. \tag{8}$$

If the largest λ in $c_{pre}(\cdot)$ is equal to p which is a finite number, then we have a VLMC model of order p. The context function $c(\cdot)$ is formed as a further simplification of $c_{pre}(\cdot)$ by lumping terminal nodes[10] with the same values into their second last symbols.

The context function $c(\cdot)$ can also be expressed as a context tree, τ, with different terminal nodes of variable lengths. Figure 1 gives a simple example of how such a context tree looks. The context tree presented has five terminal nodes with a maximum order of three. If none of the terminal nodes are pruned, then such a VLMC model is equivalent to a full Markov Chain of order three.

To fit the VLMC model we tune one parameter, K, to establish the optimal pruning of the context tree τ. The cut-off value of K is an asymptotic $\frac{1}{2}\chi^2_v$ distribution

[10] A terminal node refers to a state of possible outcome in the Markov Chain model.

with degrees of freedom $v = card(x) - 1$. In our case, as x_t is a binary time series, the cardinality is two, and $v = 1$. Hence, $K = 1.92$ and $K = 3.32$ for significance level of 5% and 1% respectively. The larger the value K, the larger the number of nodes pruned, and the smaller the context tree. The optimal cutoff K is chosen based on the information criteria. For more details about implementing the VLMC model, see Mächler and Bühlmann (2004).

Next, we fit the VLMC models with the different in-sample periods ranging from 26 weeks to 10 years. For each index, the best model is chosen based on a smallest Akaike information criterion (AIC) rule introduced by Mächler and Bühlmann (2004)[11].

In an effort to obtain more robust results for the cutoff K, we employ the bootstrap simulation to search for the optimal cutoff as in Bühlmann (2000) and Mächler and Bühlmann (2004). First, we select a relatively small initial K_0 so that not many terminal nodes are pruned. VLMC bootstrap is then performed to simulate the fitted VLMC expression $X_1^*, ..., X_n^*$. Second, we try different K values greater than K_0 in the simulation to see if they add any accuracy for one-step ahead forecasting. The loss function utilised in VLMC bootstrap is the zero-one loss $1_{[\hat{X}_{n+1}^* \neq X_{n+1}^*]}$, which is commonly used in classification modelling. Finally, the optimal K is determined by minimising the bootstrapped mean of the zero-one loss. A full example of a VLMC model specification with bootstrap is detailed in Appendix A.

4. Forecast Accuracy

4.1. Out-of-sample Forecasting

It has been well documented that good in-sample fit does not necessarily correspond to accurate out-of-sample forecasts, see, for example, (Meese and Rogoff, 1983) and Welch and Goyal (2007). In the next step of our study, we test the out-of-sample predictability of the three approaches specified in Section 3 using different in-sample window lengths. For robustness we apply both rolling and recursive window approaches for our out-of-sample forecasting.

[11] An example illustrating the selection process is presented in Appendix A and Figure A.1.

To compare the out-of-sample accuracy of different models, we employ the success rate SR. SR is a simple measure of probabilistic estimation accuracy, which has been adopted in prior studies see, for example, Leung et al. (2000) and Nyberg (2011). All observations are divided into four groups: (1) I^{uu} is the number of instances where a model's "up" prediction corresponds with a future observed "up" market movement (a correct prediction); (2) I^{ud} is the number of instances where a model's "up" prediction corresponds with a future observed "down" market movement (a false prediction); (3) I^{du} is the number of instances where a model's "down" prediction corresponds with a future observed "up" market movement (a false prediction); and (4) I^{dd} is the number of instances where a model's "down" prediction corresponds with a future observed "down" market movement (a correct prediction). The success rate SR is the ratio of number of correct predictions over the total number of predictions:

$$SR = \frac{I^{uu} + I^{dd}}{I^{uu} + I^{ud} + I^{du} + I^{dd}}. \tag{9}$$

Table 3 performs an out-of-sample comparison of success rates across RSM, ERSM, dynamic probit, VLMC, and the first order Markov Chain models. Both rolling and recursive window approaches are employed based on in-sample windows of 26, 52, 104 (2 years), 156 (3 years), 208 (4 years), 260 (5 years), 520 (10 years) and 780 weeks (15 years). The success rates are calculated by using the model produced signals that correctly predict the actual return signs across all 24 commodity futures indices. The models exhibiting the highest success rates for each of the approaches are shown in bold. We also employ a proportion test developed by Newcombe (1998a,b) to examine whether the success rates of these models are statistically different from 0.5.

According to the results of SR shown in Table 3, the predictive power of the various models is stronger using short-term in-sample windows (e.g., 52 and 104 weeks) and weaker using long-term windows (e..g, three years or longer). We argue that this is because of the reversal effect in the long-term (3-5 years) that offsets the short-term trend in commodity returns which reduces the predictability of the

models, see, e.g., Bianchi et al. (2015) and Moskowitz et al. (2012). The RSM models perform best when using the 52 weeks in-sample window, which is consistent with Papailias et al. (2017). The ERSM models show above 52% SR based on 52, 104, 156 week in-sample windows. The success rates of both RSM and ERSM are statistically different from 0.5, regardless of the choice of in-sample window. The dynamic probit models generate the highest SR (above 51%) when 104, 156 and 208 week in-sample windows are employed. This indicates that the macroeconomic conditions have intermediate-term impact on the commodity asset return signs which improves the predictability. Similarly, the VLMC and first-order Markov Chain models yield their highest SR when selecting 52 and 104 week in-sample windows. For the Markov Chain class models, the rolling window approaches perform much better than the recursive window approaches.

Next, we analyse the out-of-sample success rates of different models for each of the 24 commodity indices and report them in Table 4. We select the 52 weeks in-sample window approach as it generates the highest overall performance for the majority of the models[12]. The success rates of the RSM, ERSM, dynamic probit model, VLMC (rolling window) and the first-order Markov Chain (rolling window) models are summarised[13]. The results are quite mixed as different models show strong predictability across different commodities. Overall, the three best performing approaches are RSM, ERSM and the VLMC model with bootstrap. The probit class of models lead to the worst performance with their average success rate being close to 51%, however, they still beat their rivals in predicting Brent oil, gas and cotton.

In summary, both the return signal momentum and VLMC class models show superior out-of-sample predictability. The RSM works best based on a look-back period of 52 weeks which is consistent with the literature in momentum studies, and especially the time series momentum of Moskowitz et al. (2012). Using longer in-sample windows lead to worse predictability. The VLMC models also work in

[12]Results of the success rates using other in-sample windows are available in Appendix B.

[13]We do not include the recursive window approach for the Markov Chain class models as they do not generate superior predictability based on the results in Table 3.

the short-term in-sample window but a bootstrap practise is essential and results in a higher success rate compared to a traditional first order Markov Chain model. This implies that the market patterns repeat over the short-term but disappears over the long-term. Finally, the dynamic probit models perform the best with the intermediate-term in-sample window (104-208 weeks) settings. However, they only show marginal outperformance compared to the one with 52 weeks window.

4.2. Trading Strategies

A key consideration for practitioners is whether these models can be utilised to earn excess returns in the real world. In this section, we outline out-of-sample trading strategies based on the same in-sample windows as in Section 4.1. The position signals of commodity s at time $t+1$ is based on the model outcome $P_{t+1} = Pr(x_{t+1} = 1|\Omega_t)$ using the information available up to time t. We use a simple threshold strategy as suggested by Leung et al. (2000) where a buying signal is generated when P_{t+1} is no less than 0.5, otherwise a selling signal is formed. We use various measures to evaluate the strategy performance including annualised mean returns, standard deviation, Sharpe Ratio, cumulative returns and maximum drawdown. Details of these performance evaluation methods are provided in Appendix C.

For each individual commodity index s, its strategy return R^s_{t+1} is calculated by multiplying its actual return by the above-mentioned signal based on P_{t+1}. Next, we construct a portfolio using the equally-weighted method. For a universe of S instruments, the equally-weighted portfolio return at time $t+1$ is the average of the strategy return for each individual commodity R^s_{t+1}:

$$R^p_{t+1} = \frac{1}{S}\sum_{s=1}^{S} R^s_{t+1}. \tag{10}$$

We compute the weekly cumulative returns for each period and for different models.

Table 5 summarises the performance of the trading strategies based the seven candidate models: RSM, ERSM, dynamic probit models using rolling and recursive approaches, VLMC, VLMC with bootstrap and the first order MC model. We consider the annualised mean return and calculate its t-statistics using Newey–West

standard error (Newey and West, 1986) to check whether the returns are statistically significant. The Sharpe ratio and cumulative return are also adopted as a robustness check. We compare the key statistics of these strategies versus a buy-and-hold benchmark strategy. The first panel of Table 5 shows that none of these buy-and-hold benchmark strategies yield statistically significant mean returns in the period assessed.

In contrast with the statistically insignificant passive long returns, most of our candidate strategies show statistically significant average returns using 26, 52 and 104 week in-sample window approaches. This indicate that the commodity futures markets are indeed predictable. Both RSM and ERSM strategy returns are statistically significant at the 1% level for these three in-sample periods. The best performing RSM and ERSM strategies are those with 52 weeks in-sample windows, which have annualised mean returns of 0.078 ($t = 4.446$) and 0.067 ($t = 3.987$), respectively. This is followed by the VLMC class models which generate slightly lower mean returns but are still significant. Among the VLMC class models, the bootstrapped VLMC model using 52 weeks in-sample windows generates the highest average return of 0.077 ($t = 4.490$). Finally, although the strategies based on dynamic probit models yield returns that are mostly statistically significant at 10% level, they are not as profitable as the other types of approaches considered.

Figure 2 shows the cumulative performance of the best trading strategies based on four types of models, namely, RSM, dynamic probit, VLMC and MC models. We use the 52 weeks in-sample window approach as it generates the highest average returns across all candidate strategies. Both RSM and bootstrapped VLMC strategies exhibit steady increases over the 46 year investment horizon. A dollar invested in both strategies in 1972 yield similar cumulative returns of almost \$30 by 2018. In contrast, the first-order Markov Chain model has a cumulative return of 7.75 times and the probit rolling strategy results in 5.48 times. All of these models outperform the equally-weighted buy-and-hold strategy, indicating the presence of return predictability in commodity markets.

4.3. Robustness Check with MCS Procedure

To test which models statistically outperform the others, we apply the *Model Confidence Set* (MCS) approach of Hansen et al. (2011). Compared to traditional tests for forecast accuracy such as Diebold and Mariano (1995) test, the MCS procedure allows us to simultaneously test the performance of multiple models and select a set of superior models. It has been used in forecasting financial time series models and trading strategies as seen in e.g., Neumann and Skiadopoulos (2013) and Wang et al. (2015), among others. To the best of our knowledge, this is the first application of the MCS in binary/categorical time series setting.

The superior set of models are chosen when the null hypothesis of *Equal Predictive Ability* (EPA) is not rejected at a selected level of significance. The EPA hypothesis test is constructed based on a given loss function, which in our case is a function of the forecast error. More specifically, let Y_t be the realised value at time t, and $\hat{Y}_{i,t}$ the estimated value of model i at time t. The loss function $\ell_{i,t}$ is then calculated as:

$$\ell_{i,t} = \ell(Y_t, \hat{Y}_{i,t}). \tag{11}$$

In a binary case, Y_t is the realised return sign that takes the value of 1 if positive and 0 otherwise. $\hat{Y}_{i,t}$ is the model estimated binary indicator. We use the absolute error to calculate the loss function.

From an initial set of models M_0 which contains m candidate models, the superior set of models $M^*_{1-\alpha}$, based on a level of significance $1-\alpha$, is selected with the number of selected models $m^* \leq m$. According to Bernardi and Catania (2018), we set α as 0.2. Then, the difference between loss functions across the various models $d_{ij,t}$ is given by:

$$d_{ij,t} = \ell_{i,t} - \ell_{j,t}, \;\; i,j = 1,...,m, \;\; t = 1,...,n. \tag{12}$$

Next, to determine the best set M^*, we start an iteration where a series of EPA hypothesis tests are run to eliminate an inferior model one at a time. The EPA

hypothesis can be expressed as:

$$\begin{aligned} H_{0,M} &: E(d_{ij,t}) = 0, \ \text{ for all } \ i,j = 1,2,...,m \\ H_{A,M} &: E(d_{ij,t}) \neq 0, \ \text{ for some } \ i,j = 1,2,...,m \end{aligned} \tag{13}$$

According to Hansen et al. (2011), in order to test $H_{0,M}$, an equivalence test δ_M is constructed as:

$$T_M = \max_{i,j \in M} |t_{ij}|, \tag{14}$$

where the t-statistic

$$t_{ij} = \frac{\bar{d}_{ij}}{\sqrt{\widehat{\text{var}}(\bar{d}_{ij})}}. \tag{15}$$

In Equation 15, $\bar{d}_{ij}$ denotes the average of the loss differential and $\widehat{\text{var}}$ the bootstrapped variance. This t-statistic is also used in the other forecast error comparison methods such as Diebold and Mariano (1995) and West (1996). The iterative procedure continues until no more statistically significant inferior models are detected. The remaining m^* models are the *Superior Set Models* (SSM)[14].

To implement the MCS procedure, we first concatenate the estimated binary series for all 24 commodity indices to form a long time series data. Then, the loss differential function $d_{ij,t}$ is calculated from outputs of the the estimated binary series and the realised binary series. We repeat the above procedure for the seven candidate models, namely, RSM, ERSM, Probit rolling/recursive window approaches, VLMC, bootstrapped VLMC and first order MC models using eight in-sample window settings (26-780 weeks). This leads to 56 (7 models $\times$ 8 in-sample settings) candidate models in total. As the dynamic probit models using the recursive method are very similar across different in-sample settings, we only keep the best performing dynamic probit recursive model with an in-sample window of 104 weeks[15]. Therefore, our final reduced number of considered MCS models is 50.

[14] See Bernardi and Catania (2018) for more technical details.

[15] The MCS procedure reports errors when the candidate loss functions are too similar to one another.

Table 6 summarises the 5000 bootstrapped samples of the MCS procedure results for the 50 candidate models. 40 models are selected as the Superior Set Models (SSM) based on the EPA when $\alpha = 0.2$ and a confidence level of $1 - \alpha = 0.8$. The models contained in the SSMs are ranked based the p-value of the EPA hypothesis test. The average losses, the strategy mean returns and their Newey–West t-statistics are reported[16]. The results indicate that the two best models are the bootstrapped VLMC and 52 weeks in-sample period and the RSM with 52 weeks in-sample period, models which also had the highest mean return and t-statistics. However, some models, such as the ERSM and RSM, both with 26 week in-sample periods and ranks of 27 and 28, respectively, are ranked lower despite still generating considerable profits. This is due to the MCS procedure measuring the predictability error and not considering the return magnitude.

For comparison purpose, we have also employed the Diebold-Mariano test (Diebold and Mariano, 1995) to examine the robustness of the MCS procedure. As the Diebold-Mariano test is single EPA test which can only compare the predictive accuracy of two strategies simultaneously. For each strategy, we perform one-sided Diebold-Mariano test against all the rest strategies to find out the best strategy that outperforms the most. In the last two columns in Table 6, we report the number of outperformed models for each strategy based on significance levels of 0.1 and 0.05. The VLMC-BS and RSM, both with 52 weeks in-sample window, outperformed the majority the other candidates (45 for VLMC-BS and 44 for RSM when p-value is 0.1). In general, the number of outperformed models based on the Diebold-Mariano test decreases from rank 1 to rank 40, suggesting that the results of the MCS procedure is robust.

[16]The mean returns and t-statistics are slightly different from the investment results in Table 5. This is because we calculate each strategy's performance starting Week 781 to align the investment horizon for models with different in-sample periods.

5. Concluding Remarks

This study explores direction-of-change predictability in commodity futures markets. We employ a number of contrasting approaches including the Return Signs Momentum (RSM), the dynamic probit models and the Variable Length Markov Chain (VLMC) models. Our results suggest that the RSM and VLMC models with short-term in-sample windows are valuable in forecasting commodity price direction-of-change. To the best of our knowledge, this is the first time VLMC models have been applied in finance, with their success indicating that the strong learning effect present in commodity markets. By contrast, the binary probability models which work well in stock markets, show lower predictability than the former two approaches. Furthermore, the MCS procedure is implemented for the first time to examine the accuracy of a binary time series model. The results show consistency with our uncovered out-of-sample outcomes with the framework identifying that the RSM and VLMC 52 weeks models are superior.

Given that VLMC model is suitable in predicting stationary categorical time series data, it could be further employed in other prediction challenges in finance such as volatility direction-of-change forecasting or other discrete time series processes which have a cardinality greater than two. The advantage of this model is that it captures the market movement using a simple model-free mechanism and does not require much information other than the price itself. However, this is also its limitation where it does not take into account other relevant variables.

References

Anatolyev, S. and Gospodinov, N. (2010). Modeling financial return dynamics via decomposition. *Journal of Business & Economic Statistics*, 28(2):232–245.

Baumeister, C. and Peersman, G. (2013). Time-varying effects of oil supply shocks on the US economy. *American Economic Journal: Macroeconomics*, 5(4):1–28.

Bejerano, G. (2004). Algorithms for variable length markov chain modeling. *Bioinformatics*, 20(5):788–789.

Bernardi, M. and Catania, L. (2018). The model confidence set package for R. *International Journal of Computational Economics and Econometrics*, 8(2):144–158.

Bianchi, R. J., Drew, M. E., and Fan, J. H. (2015). Combining momentum with reversal in commodity futures. *Journal of Banking & Finance*, 59:423–444.

Brooks, C., Prokopczuk, M., and Wu, Y. (2015). Booms and busts in commodity markets: bubbles or fundamentals? *Journal of Futures Markets*, 35(10):916–938.

Brooks, C., Rew, A. G., and Ritson, S. (2001). A trading strategy based on the lead–lag relationship between the spot index and futures contract for the ftse 100. *International Journal of Forecasting*, 17(1):31–44.

Bühlmann, P. (2000). Model selection for variable length markov chains and tuning the context algorithm. *Annals of the Institute of Statistical Mathematics*, 52(2):287–315.

Bühlmann, P., Wyner, A. J., et al. (1999). Variable length markov chains. *The Annals of Statistics*, 27(2):480–513.

Cheng, I.-H. and Xiong, W. (2014). Financialization of commodity markets. *Annual Review of Financial Economics*, 6(1):419–441.

Christoffersen, P. F. and Diebold, F. X. (2006). Financial asset returns, direction-of-change forecasting, and volatility dynamics. *Management Science*, 52(8):1273–1287.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13:253–265.

Easley, D., López de Prado, M., O'Hara, M., and Zhang, Z. (2020). Microstructure in the machine age. *The Review of Financial Studies, forthcoming*.

Estrella, A. (1998). A new measure of fit for equations with dichotomous dependent variables. *Journal of Business & Economic Statistics*, 16(2):198–205.

Ferreira, M. A. and Santa-Clara, P. (2011). Forecasting stock market returns: The sum of the parts is more than the whole. *Journal of Financial Economics*, 100(3):514–537.

Gencay, R. (1998). The predictability of security returns with simple technical trading rules. *Journal of Empirical Finance*, 5(4):347–359.

Grégoir, S. and Lenglart, F. (2000). Measuring the probability of a business cycle turning point by using a multivariate qualitative hidden markov model. *Journal of Forecasting*, 19(2):81–102.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Hassan, M. R. and Nath, B. (2005). Stock market forecasting using hidden markov model: a new approach. In *Intelligent Systems Design and Applications, 2005. ISDA'05. Proceedings. 5th International Conference on*, pages 192–196. IEEE.

Hong, Y. and Chung, J. (2003). Are the directions of stock price changes predictable? statistical theory and evidence. *Manuscript, Cornell University*.

Joseph, K., Wintoki, M. B., and Zhang, Z. (2011). Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. *International Journal of Forecasting*, 27(4):1116–1127.

Kauppi, H. and Saikkonen, P. (2008). Predicting US recessions with dynamic binary response models. *The Review of Economics and Statistics*, 90(4):777–791.

Koijen, R. S., Moskowitz, T. J., Pedersen, L. H., and Vrugt, E. B. (2018). Carry. *Journal of Financial Economics*, 127(2):197–225.

Leitch, G. and Tanner, J. E. (1991). Economic forecast evaluation: profits versus the conventional error measures. *The American Economic Review*, pages 580–590.

Leung, M. T., Daouk, H., and Chen, A.-S. (2000). Forecasting stock indices: a comparison of classification and level estimation models. *International Journal of Forecasting*, 16(2):173–190.

Liu, J. and Kemp, A. (2019). Forecasting the sign of US oil and gas industry stock index excess returns employing macroeconomic variables. *Energy Economics*, 81:672–686.

Mächler, M. and Bühlmann, P. (2004). Variable length markov chains: methodology, computing, and software. *Journal of Computational and Graphical Statistics*, 13(2):435–455.

Meese, R. A. and Rogoff, K. (1983). Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics*, 14(1-2):3–24.

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

Neumann, M. and Skiadopoulos, G. (2013). Predictable dynamics in higher-order risk-neutral moments: Evidence from the s&p 500 options. *Journal of Financial and Quantitative Analysis*, 48(3):947–977.

Newcombe, R. G. (1998a). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, 17(8):873–890.

Newcombe, R. G. (1998b). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8):857–872.

Newey, W. K. and West, K. D. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix. *Econometrica*, 55:703–708.

Nyberg, H. (2011). Forecasting the direction of the US stock market with dynamic binary probit models. *International Journal of Forecasting*, 27(2):561–578.

Papailias, F., Liu, J., and Thomakos, D. D. (2017). Returns signal momentum. *Available at SSRN: https://ssrn.com/abstract=2971444*.

Pesaran, M. H. and Timmermann, A. (1995). Predictability of stock returns: Robustness and economic significance. *The Journal of Finance*, 50(4):1201–1228.

Pettenuzzo, D., Timmermann, A., and Valkanov, R. (2014). Forecasting stock returns under economic constraints. *Journal of Financial Economics*, 114(3):517–553.

Pönkä, H. (2017). Predicting the direction of US stock markets using industry returns. *Empirical Economics*, 52(4):1451–1480.

Roberts, M. J. and Schlenker, W. (2013). Identifying supply and demand elasticities of agricultural commodities: Implications for the us ethanol mandate. *American Economic Review*, 103(6):2265–95.

Rydberg, T. H. and Shephard, N. (2003). Dynamics of trade-by-trade price movements: decomposition and models. *Journal of Financial Econometrics*, 1(1):2–25.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shafiee, S. and Topal, E. (2010). An overview of global gold market and gold price forecasting. *Resources policy*, 35(3):178–189.

Tang, K. and Xiong, W. (2012). Index investment and the financialization of commodities. *Financial Analysts Journal*, 68(6):54–74.

Wang, Y., Wu, C., and Yang, L. (2015). Hedging with futures: Does anything beat the naïve hedging strategy? *Management Science*, 61(12):2870–2889.

Welch, I. and Goyal, A. (2007). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica: Journal of the Econometric Society*, pages 1067–1084.

Table 1: Summary statistics.

Sector	Commodity	Exchange	Start date	Mean	Std. dev.	Skewness	Kurtosis
Energy	Brent	ICE	Jan-1999	0.126	0.315	-0.441	5.032
	WTI	NYMEX	Jan-1987	0.077	0.325	-0.345	5.508
	Gas oil	ICE	Jan-1999	0.120	0.307	-0.386	4.112
	Heating oil	NYMEX	Jan-1983	0.073	0.312	-0.086	5.061
	Natural gas	NYMEX	Jan-1994	-0.169	0.463	0.253	4.004
	RBOB gas	NYMEX	Jan-1988	0.133	0.329	-0.196	5.114
Precious metals	Gold	COMEX	Jan-1978	0.017	0.190	0.718	17.573
	Platinum	NYMEX	Jan-1984	0.038	0.216	-0.175	5.136
	Silver	COMEX	Jan-1973	0.036	0.313	-0.036	7.286
Live stock	Lean hogs	CME	Jan-1976	-0.014	0.237	-0.086	5.110
	Live cattle	CME	Jan-1971	0.040	0.174	-0.168	5.093
Industrial metals	Aluminum	LME	Jan-1991	-0.024	0.199	-0.042	4.976
	Copper	LME	Jan-1977	0.070	0.249	-0.010	6.284
	Lead	LME	Jan-1995	0.073	0.304	0.033	5.738
	Nickel	LME	Jan-1993	0.076	0.335	0.188	4.602
	Zinc	LME	Jan-1991	0.014	0.260	0.080	5.168
Agriculture	Cocoa	ICE	Jan-1984	-0.031	0.284	0.256	4.816
	Coffee	ICE	Jan-1981	-0.011	0.340	1.210	13.438
	Corn	CBOT	Jan-1971	-0.033	0.240	0.292	6.276
	Cotton	ICE	Jan-1977	0.009	0.229	0.375	5.466
	Soybean	CBOT	Jan-1971	0.052	0.246	0.289	6.002
	Sugar	ICE	Jan-1973	0.011	0.367	0.345	6.071
	Wheat Chicago	CBOT	Jan-1971	-0.024	0.259	0.483	5.249
	Wheat Kansas	KCBT	Jan-1999	-0.049	0.273	0.266	4.327

This table reports the mean return, standard deviation, skewness and kurtosis for the 24 commodity futures indices which are available from January 1971 to August 2018. The arithmetic monthly mean returns and standard deviation are both annualised. The detailed data sources are described in Appendix C.

Table 2: In-sample performance of the dynamic probit model.

	26 Weeks			52 Weeks			2 Years			3 Years		
	PS.R2	BIC	SR	PS.R2	BIC	SR	PS.R2	BIC	SR	PS.R2	BIC	SR
Brent	0.35	31.84	0.80	0.12	71.12	0.73	0.06	145.83	0.62	0.05	215.85	0.59
WTI	0.18	35.44	0.64	0.03	76.84	0.61	0.03	148.43	0.56	0.01	223.28	0.53
Gas oil	0.46	28.70	0.76	0.15	70.71	0.67	0.06	146.11	0.59	0.07	213.15	0.65
Heating oil	0.48	27.92	0.76	0.17	69.24	0.65	0.02	150.32	0.55	0.01	223.43	0.52
Natural gas	0.36	30.61	0.72	0.08	74.02	0.55	0.06	145.11	0.55	0.05	217.69	0.60
RBOB gas	0.50	27.04	0.84	0.26	62.20	0.71	0.05	144.52	0.61	0.01	220.57	0.58
Gold	0.29	31.63	0.72	0.15	61.78	0.76	0.09	141.16	0.65	0.04	218.11	0.59
Platinum	0.40	30.16	0.76	0.06	73.92	0.63	0.03	146.66	0.53	0.05	213.82	0.59
Silver	0.23	35.13	0.76	0.17	69.57	0.65	0.12	138.67	0.59	0.06	214.73	0.58
Lean hogs	0.37	30.94	0.76	0.04	74.83	0.57	0.05	142.15	0.61	0.04	216.44	0.59
Live cattle	0.31	31.99	0.68	0.10	70.18	0.63	0.02	144.10	0.60	0.06	206.01	0.68
Aluminum	0.27	33.80	0.64	0.07	75.14	0.63	0.07	142.77	0.62	0.06	215.53	0.61
Copper	0.21	35.32	0.64	0.10	71.76	0.67	0.08	138.96	0.64	0.06	214.19	0.63
Lead	0.07	38.93	0.64	0.08	74.29	0.57	0.05	145.25	0.58	0.04	214.01	0.56
Nickel	0.16	36.88	0.64	0.06	74.71	0.63	0.04	147.48	0.60	0.00	224.06	0.52
Zinc	0.52	26.38	0.76	0.11	72.65	0.59	0.04	145.36	0.55	0.02	219.78	0.55
Cocoa	0.13	37.82	0.60	0.16	70.29	0.71	0.03	148.84	0.59	0.05	216.94	0.61
Coffee	0.22	34.36	0.76	0.14	70.51	0.59	0.03	145.72	0.58	0.02	211.48	0.63
Corn	0.35	27.13	0.80	0.07	71.72	0.69	0.08	140.47	0.60	0.09	210.88	0.53
Cotton	0.13	37.47	0.68	0.11	72.52	0.57	0.08	142.10	0.59	0.04	217.96	0.60
Soybean	0.18	36.47	0.60	0.08	74.31	0.63	0.11	139.89	0.60	0.06	213.98	0.59
Sugar	0.20	32.68	0.76	0.17	63.80	0.65	0.08	142.33	0.67	0.09	210.55	0.64
Wheat Chicago	0.14	34.18	0.68	0.07	74.60	0.57	0.03	148.30	0.53	0.02	217.10	0.55
Wheat Kansas	0.48	26.11	0.80	0.15	69.35	0.63	0.06	141.03	0.65	0.03	217.83	0.56

This table summarises the performance of the dynamic probit models specified in Equation 6 based on in-sample periods of 26 weeks, 52 weeks, two years and three years. The evaluation criteria BIC, pseudo-R^2 (PS.R2) and success rate SR are presented for the dynamic probit regressions run on each of the 24 commodity indices. The independent variables consists of the lagged binary time series of return signs x_t, the returns of the MSCI world index, the S&P GSCI index, the USD index, the first difference of US short-term interest rate (3-month treasury yield) Δ_{SI}, US long-term interest rate (10-year treasury yield) Δ_{LI} and the NBER recession indicator.

Table 3: Out-of-sample comparison of success rates across different approaches and in-sample windows.

	26	52	104	156	208	260	520	780
RSM	**0.517***	**0.523***	**0.518***	0.513*	0.513*	0.512*	0.510*	0.513*
ERSM	**0.516***	**0.521***	**0.522***	**0.521***	0.518*	0.515*	0.512*	0.513*
Probit-rolling	0.501	0.509*	**0.511***	0.508*	**0.510***	0.509*	0.506	0.502
Probit-recursive	0.509*	0.510*	**0.510***	**0.511***	**0.511***	0.510*	0.509*	0.508*
VLMC-rolling	0.509*	**0.513***	**0.511***	0.505	0.508*	0.508*	0.506	0.510*
VLMC-BS-rolling	**0.514***	**0.522***	**0.515***	0.508*	0.506	0.509*	0.502	0.506
VLMC-recursive	0.500	0.497	0.503	0.502	0.505	0.506	0.509*	0.509*
VLMC-BS-recursive	0.506*	0.505	0.504	0.504	0.505	0.505	0.504	0.502
MC1-rolling	0.509*	**0.516***	**0.517***	0.513*	0.510*	0.509*	0.505	0.507
MC1-recursive	0.509*	0.509*	0.508*	0.508*	0.508*	0.508*	0.507	0.508

This table compares the out-of-sample success rates across the RSM, ERSM, dynamic probit, VLMC and first order Markov Chain models. Both rolling and recursive window approaches are employed based on in-sample windows of 26, 52, 104 (2 years), 156 (3 years), 208 (4 years), 260 (5 years), 520 (10 years) and 780 weeks (15 years). The models exhibiting the highest success rates for each of the approaches are shown in bold. The success rates are calculated by averaging model signals that correctly predict actual return signs across all of the 24 commodity futures indices. A proportion test is run to examine whether the success rates are statistically significant different from 0.5 ($*$: $p < 0.01$.

Table 4: Out-of-sample comparison of success rates for individual indices (52 weeks in-sample window).

	RSM	ERSM	Probit-roll	Probit-rec	VLMC	VLMC-BS	MC1
Brent	0.521	0.526	0.504	**0.540**	0.504	0.517	0.501
WTI	**0.521**	0.514	0.503	0.505	0.514	0.509	0.506
Gas oil	0.543	0.528	0.524	**0.548**	0.496	0.510	0.516
Heating oil	**0.526**	0.522	0.509	0.503	0.522	0.521	0.519
Natural gas	0.519	0.513	0.488	0.512	**0.524**	0.520	0.485
RBOB gas	0.522	0.517	0.499	0.507	**0.530**	0.527	0.514
Gold	**0.536**	0.522	0.510	0.500	0.535	**0.536**	0.532
Platinum	0.541	0.536	0.515	0.525	0.548	**0.552**	0.536
Silver	0.526	0.519	0.522	0.491	0.521	0.533	**0.534**
Lean hogs	0.517	0.510	0.511	0.503	0.516	0.508	**0.518**
Live cattle	**0.519**	0.499	0.509	0.514	0.512	0.519	0.511
Aluminum	0.545	**0.547**	0.524	0.517	0.503	0.540	0.531
Copper	**0.533**	0.532	0.520	0.518	0.520	**0.533**	0.521
Lead	**0.544**	0.533	0.505	0.516	0.509	0.504	0.522
Nickel	0.529	**0.537**	0.508	0.503	0.488	0.525	0.529
Zinc	0.515	**0.535**	0.499	0.498	0.513	0.522	0.522
Cocoa	0.520	0.515	0.487	0.501	0.488	**0.523**	0.506
Coffee	0.516	0.512	0.520	0.518	0.519	**0.526**	0.517
Corn	0.532	**0.536**	0.518	0.504	0.521	0.526	0.513
Cotton	0.498	0.514	0.497	**0.519**	0.481	0.498	0.505
Soybean	0.512	**0.516**	0.497	0.505	0.496	0.512	0.495
Sugar	0.524	0.528	0.526	0.506	0.510	**0.531**	0.528
Wheat Chicago	0.512	0.503	0.511	0.517	**0.517**	0.517	0.504
Wheat Kansas	0.501	**0.514**	0.504	0.495	0.502	0.509	0.511
Average	0.524	0.522	0.509	0.511	0.512	0.521	0.516

This table presents out-of-sample success rates for RSM, ERSM, dynamic probit, VLMC (rolling window) and first-order Markov Chain (rolling window) models for each commodity indices. The 52 weeks in-sample window is selected, as for most models, it generates the highest success rates compared to other in-sample periods. For each commodity, the model exhibiting the highest success rate is shown in bold.

Table 5: Evaluation of trading strategies based on different models and in-sample periods.

	26	52	104	156	208	260	520	780
Buy-and-Hold								
Mean	0.031	0.025	0.019	0.004	0.004	0.012	0.000	0.014
NW-t	1.174	0.946	0.715	0.183	0.178	0.559	0.021	0.600
Sharpe	0.214	0.168	0.126	0.031	0.030	0.090	0.003	0.105
Cum. Return	2.662	1.924	1.428	0.773	0.784	1.134	0.730	1.186
Drawdown	0.624	0.661	0.643	0.662	0.680	0.679	0.560	0.530
RSM								
Mean	0.069	0.078	0.049	0.026	0.011	0.019	0.019	0.036
NW-t	3.938***	4.446***	2.625***	1.422	0.643	1.175	1.317	2.535***
Sharpe	0.616	0.720	0.449	0.239	0.109	0.188	0.221	0.432
Cum. Return	20.062	29.374	7.328	2.494	1.296	1.805	1.809	2.962
Drawdown	0.336	0.357	0.366	0.499	0.608	0.372	0.355	0.227
ERSM								
Mean	0.070	0.067	0.068	0.053	0.035	0.027	0.019	0.024
NW-t	3.936***	3.978***	3.744***	3.223***	2.272**	1.718*	1.230	1.574
Sharpe	0.634	0.596	0.628	0.525	0.359	0.285	0.200	0.276
Cum. Return	21.301	17.131	17.187	8.599	3.766	2.663	1.725	1.952
Drawdown	0.285	0.407	0.278	0.322	0.297	0.414	0.440	0.260
Probit rolling window								
Mean	0.015	0.042	0.023	0.014	0.021	0.005	0.013	0.019
NW-t	1.009	2.567***	1.520	0.932	1.412	0.365	1.062	1.247
Sharpe	0.148	0.397	0.227	0.142	0.212	0.057	0.145	0.215
Cum. Return	1.590	5.476	2.293	1.518	2.025	1.038	1.408	1.644
Drawdown	0.459	0.402	0.463	0.456	0.501	0.585	0.262	0.362
Probit recursive window								
Probit-rec	0.028	0.029	0.034	0.031	0.029	0.024	0.024	0.027
NW-t	1.908*	1.944*	2.300**	2.174**	2.222**	1.852*	1.701*	1.874*
Sharpe	0.289	0.285	0.335	0.318	0.308	0.266	0.268	0.324
Cum. Return	3.079	3.050	3.818	3.244	2.911	2.346	2.124	2.200
Drawdown	0.461	0.461	0.426	0.377	0.399	0.385	0.261	0.296
VLMC								
Mean	0.052	0.052	0.033	0.001	0.006	0.000	0.014	0.010
NW-t	3.105***	3.310***	1.945*	0.041	0.450	0.006	1.057	0.707
Sharpe	0.519	0.540	0.335	0.006	0.065	0.001	0.173	0.130
Cum. Return	9.227	9.304	3.705	0.845	1.083	0.835	1.489	1.265
Drawdown	0.306	0.328	0.418	0.672	0.337	0.561	0.298	0.374
bootstrapped VLMC								
Mean	0.053	0.077	0.049	0.019	-0.004	0.019	-0.007	0.001
NW-t	3.056***	4.490***	2.679***	1.249	-0.275	1.411	-0.524	0.107
Sharpe	0.494	0.753	0.471	0.200	-0.043	0.217	-0.084	0.019
Cum. Return	9.408	29.334	7.489	1.919	0.710	1.940	0.683	0.952
Drawdown	0.342	0.321	0.339	0.493	0.569	0.362	0.523	0.406
First order MC								
Mean	0.045	0.049	0.036	0.026	0.012	0.011	0.006	0.000
NW-t	2.713***	2.950***	2.198**	1.555	0.706	0.714	0.440	0.003
Sharpe	0.440	0.483	0.358	0.259	0.119	0.117	0.070	0.001
Cum. Return	6.668	7.752	4.211	2.541	1.348	1.329	1.093	0.898
Drawdown	0.298	0.415	0.326	0.360	0.413	0.418	0.380	0.453

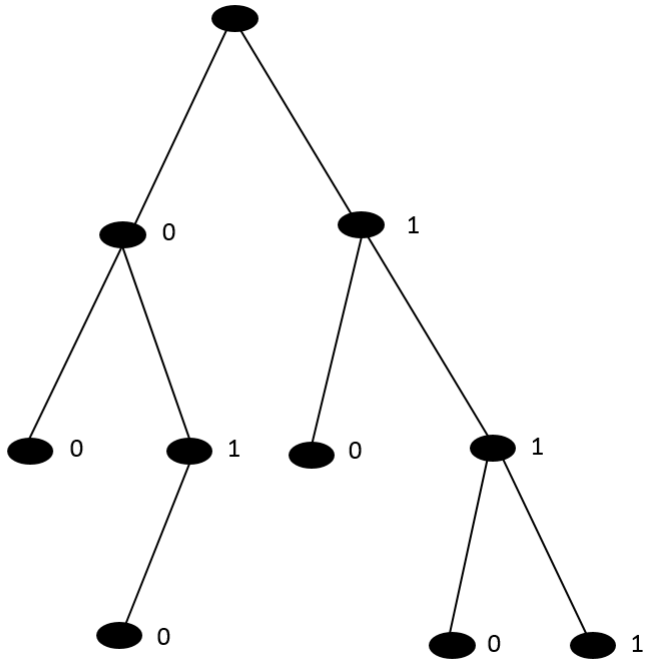
This table summarises the performance statistics of different strategies using an equally-weighted portfolio construction scheme. We employ a simple buy-and-hold strategy as the benchmark and strategies based on RSM, ERSM, dynamic probit, VLMC and first-order Markov Chain models. We test these strategies using in-sample windows of 26, 52, 104 (2 years), 156 (3 years), 208 (4 years), 260 (5 years), 520 (10 years) and 780 weeks (15 years). The statistics reported include the annualised mean returns, the Newey–West t-statistics (*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$), the Sharpe ratio, the cumulative returns and the maximum drawdown. Further information on the strategy evaluation methodology can be seen in Appendix C. The entire sample period spans from January 1971 to August 2018.

Table 6: Superior Set Models selected by MCS Procedure

Rank	Model	P-value	Loss	Mean	NW-t	DM.SPA0.1	DM.SPA50.05
1	VLMC-BS 52	1.000	0.479	0.051	4.490	45	40
2	RSM 52	1.000	0.479	0.056	4.446	44	41
3	ERSM 104	1.000	0.482	0.046	3.744	35	30
4	MC 104	1.000	0.483	0.030	2.198	28	22
5	RSM 104	1.000	0.483	0.044	2.625	30	22
6	ERSM 156	1.000	0.483	0.049	3.223	33	25
7	RSM 780	1.000	0.485	0.036	2.535	20	14
8	RSM 208	1.000	0.485	0.027	0.643	21	18
9	ERSM 52	1.000	0.484	0.045	3.978	27	21
10	ERSM 780	1.000	0.486	0.024	1.574	17	12
11	RSM 260	0.999	0.486	0.020	1.175	19	13
12	ERSM 520	0.999	0.486	0.017	1.230	17	11
13	VLMC-BS 104	0.984	0.486	0.039	2.679	18	10
14	VLMC 780	0.984	0.488	0.010	0.707	7	4
15	ERSM 208	0.981	0.485	0.036	2.272	20	14
16	RSM 520	0.977	0.488	0.023	1.317	10	5
17	RSM 156	0.963	0.486	0.036	1.422	15	8
18	ERSM 260	0.880	0.487	0.031	1.718	15	8
19	VLMC-BS 208	0.835	0.489	0.004	-0.275	5	4
20	Probit recursive 104	0.830	0.490	0.030	2.300	4	2
21	VLMC 260	0.830	0.490	0.002	0.006	4	2
22	VLMC 208	0.830	0.489	0.002	0.450	5	3
23	VLMC 52	0.830	0.486	0.038	3.310	16	9
24	MC 52	0.733	0.486	0.033	2.950	15	8
25	VLMC 104	0.723	0.488	0.030	1.945	8	4
26	MC 156	0.664	0.489	0.021	1.555	6	4
27	ERSM 26	0.664	0.488	0.038	3.936	6	4
28	RSM 26	0.573	0.488	0.047	3.938	7	5
29	MC 780	0.565	0.491	0.000	0.003	2	1
30	MC 260	0.557	0.490	0.008	0.714	4	2
31	VLMC-BS 260	0.550	0.491	0.013	1.411	3	1
32	VLMC-BS 156	0.510	0.490	0.029	1.249	4	2
33	VLMC-BS 26	0.481	0.489	0.033	3.056	6	4
34	VLMC 520	0.425	0.492	0.002	1.057	1	1
35	VLMC-BS 780	0.294	0.493	0.001	0.107	1	0
36	Probit rolling 104	0.284	0.491	0.009	1.520	2	2
37	MC 208	0.276	0.491	0.009	0.706	2	1
38	Probit rolling 520	0.269	0.493	0.019	1.062	1	0
39	Probit rolling 208	0.230	0.492	0.012	1.412	2	1
40	Probit rolling 260	0.224	0.492	0.017	0.365	2	1

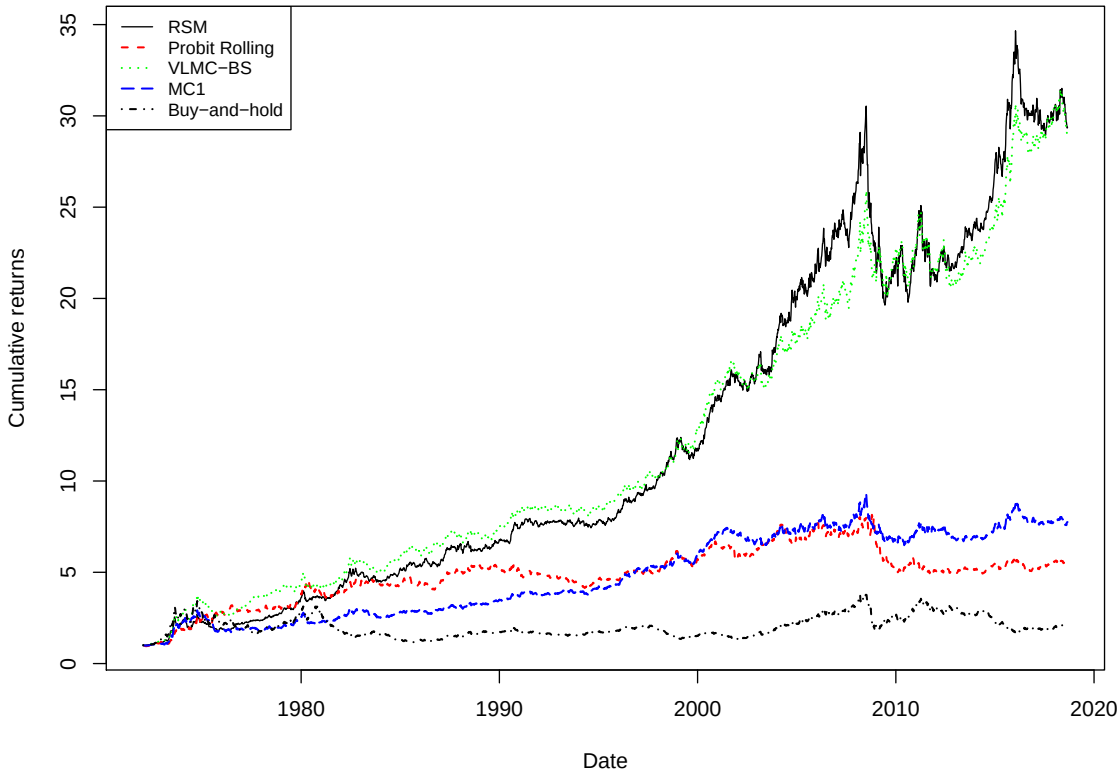
This table reports the results of the MCS procedure for the 50 candidate models with a bootstrap of 5000 times. 40 models are selected for the Superior Set Models (SSM) based on the EPA with $\alpha = 0.2$ and a confidence level of $1 - \alpha = 0.8$. The SSMs are then ranked based the p-value of the EPA hypothesis test. The average losses, the strategy mean returns and their Newey–West t-statistics are reported. The last two columns (DM.SPA) present the number of superior predictive accuracy (SPA) for each model compared to the rest models using Diebold-Mariano test (Diebold and Mariano, 1995). Results are based on one-sided test with significance levels of 0.1 and 0.05.

Figure 1: An example of a VLMC context tree.



This Figure gives a simple example of a context tree. The context tree has five terminal nodes with a maximum order of three. If no terminal nodes are pruned, then such a VLMC model is equivalent to a full Markov Chain of order three.

Figure 2: Cumulative return comparison across different strategies using the 52 weeks in-sample window.



This Figure reports the cumulative return series of the RSM, dynamic probit (rolling window), bootstrapped VLMC and the first-order MC model strategies. A simple buy-and-hold portfolio that is passively long in the 24 commodity indices is adopt as the benchmark. The 52 weeks in-sample window is used for all strategies. The investment horizon starts in January 1972 and ends in August 2018.

Appendices

A. The VLMC model specification

In this appendix, we show the structure and specification of the VLMC model using an example of a 52 weeks in-sample period. We fit VLMC models for each of the 24 futures indices with different cutoff values K ranging from 0.5-3 using a step value of 0.02[17]. The cutoff value for each instrument K, is selected when its AIC is minimised. As mentioned previously, when K increases, more contexts are pruned. This process continues until K is large enough to reduce the context to 1, indicating that the underlying series is independent.

Figure A.1 reports the AIC as a function of cutoff K for each of future instrument. The cutoff K is chosen based on a minimised AIC value. For all futures indices, we find the optimal cutoff K, before the K exceeds 2.3, where the number of contexts is reduced to 1. Further details of the key statistics of these VLMC models are shown in the left panel of Table A.1. The cutoff K for most of the indices are below 2, except for corn ($K = 2.14$) and Wheat traded in Chicago ($K = 2.24$). This yields VLMC models with orders from 5 to 11 and number of contexts from 7 to 106.

In order to obtain more robust results for the cutoff K, we employ the bootstrap simulation to search for the optimal cutoff as in Bühlmann (2000) and Mächler and Bühlmann (2004). We chose $K_0 = 0.3$ so that the context tree is initially "large". The choice of K_0 has only a minor impact on final accuracy and does not affect the optimal K as shown in Mächler and Bühlmann (2004). We also try different K_0, but the results barely change. A subsequent grid search of K from 0.4 to 2.5 using steps of 0.02 is performed in order to get the best VLMC model. The upper limit of 2.3 is selected as its χ^2 probability is equal to 97.5%, where no terminal node is further pruned and the number of contexts is reduced to one.

The right panel of Table A.1 reports detailed statistics of VLMC models with various K values selected via bootstrap. The K value is generally higher than the

[17]The step value can also be a value smaller than 0.02 to make the estimation more accurate. We choose 0.02 however, as there is a trade-off between computation speed and model accuracy.

cutoff K selected without bootstrap. As a consequence, the order and context are relatively smaller, returning structurally simpler models. In some cases, e.g., Brent oil futures, the smallest AIC occurs when there is only one context left, indicating serial independence. In that case, the estimated conditional probability $\hat{P}_t$ of such a VLMC model is always equal to its in-sample probability P_t.

B. Success Rates of Individual Commodities Using Alternative In-sample Windows

This appendix presents the out-of-sample success rates of different models adopted on individual commodities with longer in-sample windows including 104, 156 and 260 weeks. Table B.1, B.2 and B.3 show the results for each commodity based on the RSM, ERSM, dynamic probit, VLMC, VLMC bootstrap and first order Markov Chain models. Comparing these results to the ones in Table 4, we witness that in general the predictive power reduces as the in-sample windows increases. The ERSM is less affected by the increasing in-sample windows as it assigns more weights to the recent past. Therefore, ERSM with long-term in-sample window is essentially indifferent from the one with short-term in-sample window.

For results using a 104 weeks in-sample window as shown in Table B.1, the average success rates for all the models are still significantly higher than 0.5 at 1% level. The ERSM and RSM are the two best models, followed by the VLMC-BS and first order Markov Chain. When a 156 weeks in-sample window is used, the ERSM model still outperforms the rest with the highest number of best performed models across commodities. Finally, when using the 260 weeks in-sample window, the VLMC class models become the worst compared to the dynamic probit models and the RSM/ERSM.

C. Strategy Evaluation

We evaluate the candidate trading strategies by considering both return and risk context. The return measures include average returns, minimum/maximum returns and cumulative net profits. The risk related measures consist of standard deviation, maximum drawdowns and Sharpe Ratio (reward-to-risk ratio). Let R_t denote the return of a strategy at month t ranging from m_1 to m_n, the evaluation measures are calculated as follows:

1. The annualised average return

$$AR \stackrel{\text{def}}{=} \frac{1}{n} \sum_{t=m_1}^{m_n} R_t \tag{16}$$

2. The cumulative return

$$CR \stackrel{\text{def}}{=} \prod_{t=m_1}^{t=m_n} (1 + R_t) \tag{17}$$

3. The annualised volatility/standard deviation

$$SD \stackrel{\text{def}}{=} \sqrt{\frac{1}{n} \sum_{t=m_1}^{m_n} (R_t - AR)^2} \tag{18}$$

4. The gross Sharpe Ratio, annualised

$$SR \stackrel{\text{def}}{=} \frac{AR}{SD} \tag{19}$$

5. The maximum drawdown MDD_t measures the maximum historical decline over the investment horizon. The maximum value from an arbitrary peak of the cumulative profit to any subsequent cumulative profit from time 0 to time T is calculated. The formula of maximum drawdown can be expressed as:

$$MDD_t = \frac{\max_{T \in (0,t)} \{0, \max CR_T - CR_t\}}{\max_{T \in (0,t)} CR_T - 1} \tag{20}$$

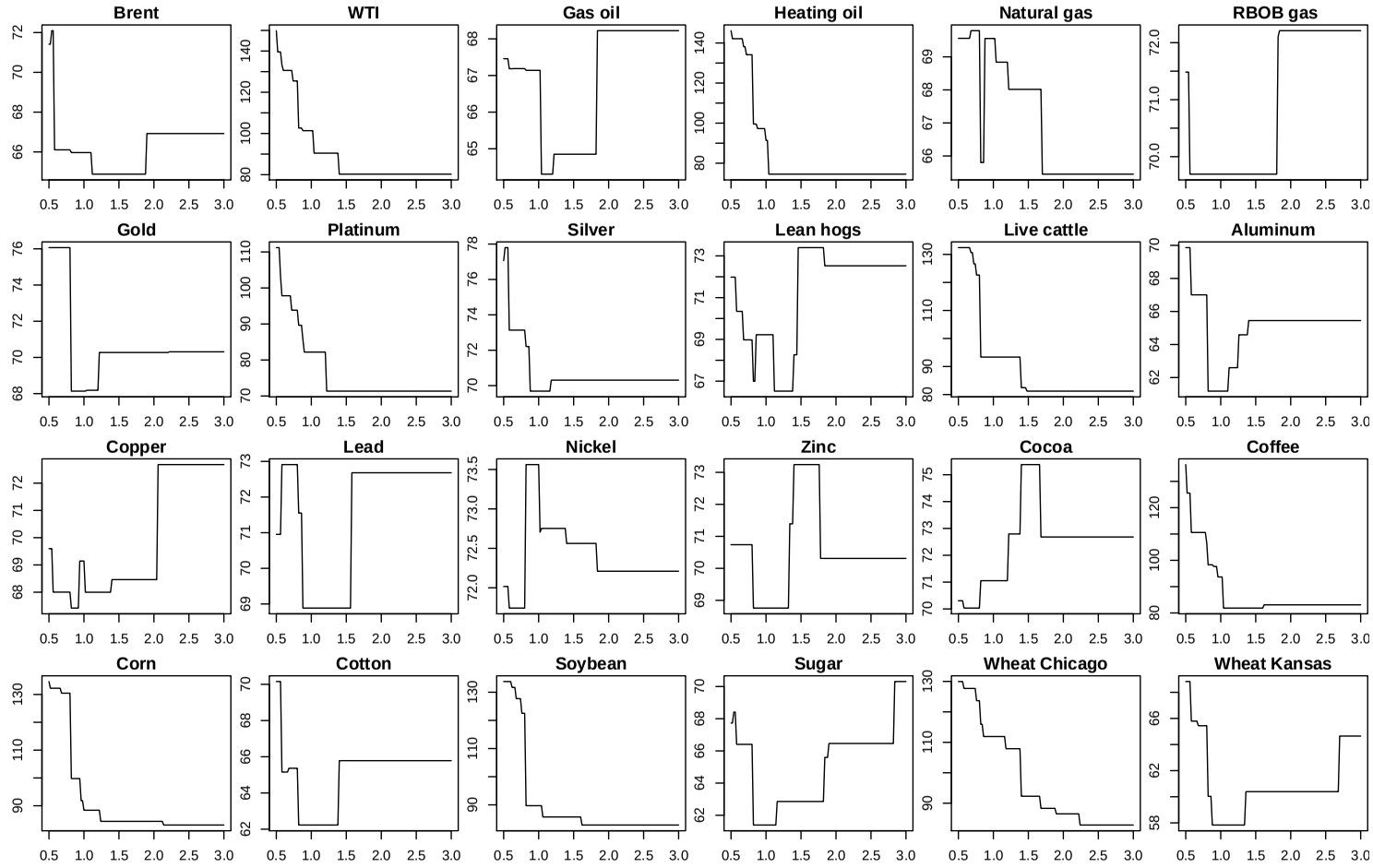
where CR_t denotes the cumulative return at time t. $\max_{T\in(0,t)} CR_T - 1$ is the highest cumulative net profit from time 0 to time T.

Table A.1: Statistics of selected VLMC models.

	Without Bootstrap				Bootstrap			
	Cutoff	Order	Context	AIC	Cutoff	Order	Context	AIC
Brent	1.5	7	18	356.05	2.34	0	1	355.16
WTI	1.4	7	22	411.92	1.45	7	22	411.92
Gas oil	1.12	8	39	334.41	1.58	8	30	334.15
Heating oil	1.04	8	35	435.20	2.08	6	11	373.97
Natural gas	1.7	8	11	357.57	2.48	0	1	358.23
RBOB gas	1.18	8	47	349.09	1.32	8	44	348.23
Gold	0.92	7	60	520.34	0.99	7	54	500.26
Platinum	1.22	10	66	519.25	1.45	6	18	392.24
Silver	1.02	7	36	376.31	1.99	0	1	360.58
Lean hogs	1.25	8	30	350.58	2.5	3	4	354.75
Live cattle	1.48	6	20	388.77	0.95	10	77	537.22
Aluminum	0.96	11	49	483.96	2.06	6	8	380.34
Copper	0.87	8	40	364.40	1.98	6	7	359.10
Lead	1.22	9	45	468.28	0.62	11	123	668.61
Nickel	0.69	11	86	374.06	2.48	4	5	356.72
Zinc	1.07	7	42	457.26	2.08	6	12	384.51
Cocoa	0.69	10	106	622.53	1.86	6	20	416.01
Coffee	1.32	9	46	451.78	1.45	8	36	425.08
Corn	2.14	7	13	394.71	2.08	7	13	394.71
Cotton	1.1	10	33	358.03	1.48	6	11	357.84
Soybean	1.62	8	24	415.24	2.44	0	1	373.22
Sugar	0.98	10	69	537.76	2.36	5	6	380.16
Wheat Chicago	2.24	5	7	367.44	2.44	5	7	367.44
Wheat Kansas	1.11	8	24	334.79	2.44	0	1	356.31

This Table reports the cutoff K, order, number of contexts, and AIC of the best VLMC models for different futures indices. The fitted models are run based on the in-sample data of 52 weeks. The left panel reports the statistics of VLMC models without bootstrap, whereas the right panel reports the statistics via bootstrap.

Figure A.1: AIC as a function of cutoff K for different commodity futures.



The Figure reports the AIC as a function of VLMC cutoff K using the weekly returns of the 24 commodity futures indices. The VLMC models are performed based on the in-sample period of 52 weeks. The vertical and horizontal axes show the AIC and cutoff K, respectively.

Table B.1: Out-of-sample comparison of success rates for individual indices (104 weeks in-sample window).

	RSM	ERSM	Probit-rw	Probit-rec	VLMC	VLMC-BS	MC1
Brent	0.539	0.528	0.533	**0.545**	0.486	0.541	0.530
WTI	0.516	**0.517**	0.511	0.507	0.516	0.508	0.501
Gas oil	**0.563**	0.544	0.549	0.556	0.513	0.555	0.538
Heating oil	0.526	0.518	0.522	0.508	0.520	**0.529**	0.508
Natural gas	0.502	0.515	0.490	0.515	**0.519**	0.513	0.501
RBOB gas	0.518	**0.522**	0.506	0.507	0.511	0.519	0.504
Gold	0.533	0.527	0.515	0.501	0.532	0.537	**0.542**
Platinum	0.548	0.547	0.528	0.529	0.548	**0.559**	0.544
Silver	0.512	0.521	0.506	0.492	0.518	0.507	**0.542**
Lean hogs	0.504	**0.510**	0.490	0.500	0.502	0.497	**0.510**
Live cattle	0.503	0.507	0.501	**0.512**	0.502	0.508	0.511
Aluminum	0.515	**0.539**	0.515	0.512	0.509	0.532	0.517
Copper	0.523	**0.528**	0.525	0.514	0.517	0.525	0.513
Lead	**0.544**	0.538	0.516	0.520	0.520	0.524	0.529
Nickel	0.496	**0.521**	0.507	0.506	0.479	0.496	0.507
Zinc	0.512	**0.526**	0.509	0.498	0.493	0.507	0.523
Cocoa	0.525	**0.535**	0.523	0.506	0.509	0.520	0.525
Coffee	0.517	0.517	0.519	0.519	**0.527**	0.516	0.518
Corn	0.516	**0.529**	0.506	0.502	0.503	0.507	0.499
Cotton	0.504	0.504	0.503	**0.518**	0.485	0.481	0.497
Soybean	0.512	**0.522**	0.493	0.503	0.511	0.512	0.510
Sugar	0.511	0.524	0.513	0.508	0.511	0.506	**0.527**
Wheat Chicago	0.514	0.516	0.508	0.518	**0.522**	0.519	0.510
Wheat Kansas	0.515	**0.511**	0.499	0.490	0.484	0.487	0.511
Average	0.519	0.524	0.512	0.512	0.510	0.517	0.517

This table presents out-of-sample success rates for RSM, ERSM, dynamic probit, VLMC (rolling window) and first-order Markov Chain (rolling window) models for each commodity indices with an in-sample window of 104 weeks. For each commodity, the model exhibiting the highest success rate is shown in bold.

Table B.2: Out-of-sample comparison of success rates for individual indices (156 weeks in-sample window).

	RSM	ERSM	Probit-roll	Probit-rec	VLMC	VLMC-BS	MC1
Brent	0.532	0.523	0.539	**0.541**	0.495	0.529	0.521
WTI	0.509	**0.519**	0.510	0.509	0.496	0.512	0.508
Gas oil	0.546	**0.560**	0.541	0.549	0.521	0.530	0.548
Heating oil	0.500	**0.519**	0.503	0.509	0.509	0.506	0.505
Natural gas	0.524	**0.526**	0.493	0.515	0.505	0.523	0.500
RBOB gas	0.487	**0.514**	0.497	0.513	0.498	0.505	0.491
Gold	0.547	0.533	0.516	0.504	0.512	0.516	**0.553**
Platinum	0.541	**0.548**	0.536	0.528	0.547	0.538	0.542
Silver	0.507	0.519	0.514	0.493	0.506	0.504	**0.530**
Lean hogs	0.500	**0.509**	0.485	0.500	0.497	0.501	0.495
Live cattle	0.487	0.505	0.504	0.509	0.485	0.493	**0.516**
Aluminum	**0.530**	0.529	0.515	0.513	0.529	0.501	0.510
Copper	0.524	**0.529**	0.516	0.515	0.511	0.518	0.507
Lead	0.519	**0.532**	0.501	0.521	0.518	0.521	0.511
Nickel	0.508	0.518	**0.521**	0.511	0.459	0.484	0.511
Zinc	0.506	**0.523**	0.514	0.500	0.502	0.498	0.507
Cocoa	0.521	**0.532**	0.517	0.507	0.521	0.521	0.519
Coffee	0.514	0.511	0.509	**0.517**	0.504	0.487	0.505
Corn	0.498	**0.524**	0.501	0.503	0.487	0.493	0.501
Cotton	0.516	0.508	0.501	**0.520**	0.497	0.515	0.499
Soybean	0.501	0.504	0.481	0.502	0.505	**0.506**	0.499
Sugar	0.520	**0.525**	0.512	0.506	0.513	0.508	0.523
Wheat Chicago	0.512	0.512	0.513	**0.517**	0.510	0.513	0.506
Wheat Kansas	0.506	**0.521**	0.485	0.495	0.502	0.507	0.515
Average	0.515	0.523	0.509	0.512	0.505	0.510	0.513

This table presents out-of-sample success rates for RSM, ERSM, dynamic probit, VLMC (rolling window) and first-order Markov Chain (rolling window) models for each commodity indices with an in-sample window of 156 weeks. For each commodity, the model exhibiting the highest success rate is shown in bold.

Table B.3: Out-of-sample comparison of success rates for individual indices (260 weeks in-sample window).

	RSM	ERSM	Probit-roll	Probit-rec	VLMC	VLMC-BS	MC1
Brent	0.497	0.531	0.525	**0.534**	0.494	0.506	0.500
WTI	0.522	**0.531**	0.496	0.508	0.496	0.492	0.511
Gas oil	0.511	**0.539**	0.520	0.536	0.506	0.518	0.527
Heating oil	0.507	0.511	0.514	0.512	**0.515**	**0.515**	0.500
Natural gas	**0.540**	0.537	0.520	0.526	0.504	**0.540**	0.509
RBOB gas	0.496	0.503	0.500	0.510	0.490	**0.513**	0.489
Gold	**0.542**	0.535	0.513	0.501	0.522	0.531	0.536
Platinum	0.526	0.537	0.532	0.523	0.522	0.530	**0.544**
Silver	0.511	0.510	0.506	0.495	**0.533**	0.506	0.523
Lean hogs	0.509	0.504	**0.511**	0.504	0.499	0.492	0.483
Live cattle	0.515	0.495	0.488	0.507	0.516	**0.528**	0.519
Aluminum	0.523	0.520	**0.530**	0.513	0.505	0.502	0.516
Copper	**0.518**	0.517	0.510	0.516	0.501	0.502	0.502
Lead	0.507	**0.531**	0.504	0.515	0.496	0.492	0.507
Nickel	0.516	0.508	**0.524**	0.513	0.496	0.499	0.492
Zinc	0.505	0.501	**0.511**	0.496	0.479	0.481	0.482
Cocoa	0.531	0.514	0.499	0.507	**0.537**	0.525	0.515
Coffee	0.504	0.511	0.510	**0.523**	0.497	0.470	0.505
Corn	0.498	**0.516**	0.490	0.500	0.499	0.507	0.506
Cotton	0.503	0.507	0.511	**0.518**	0.501	0.517	**0.518**
Soybean	0.498	**0.508**	0.504	0.506	0.503	0.506	0.489
Sugar	0.510	0.516	**0.520**	0.504	0.504	0.507	0.518
Wheat Chicago	0.510	0.510	0.509	0.519	**0.526**	0.517	0.511
Wheat Kansas	0.501	**0.524**	0.518	0.504	0.497	0.494	0.499
Average	0.512	0.517	0.511	0.512	0.506	0.508	0.508

This table presents out-of-sample success rates for RSM, ERSM, dynamic probit, VLMC (rolling window) and first-order Markov Chain (rolling window) models for each commodity indices with an in-sample window of 260 weeks. For each commodity, the model exhibiting the highest success rate is shown in bold.