Microstructure in the Machine Age

David Easley, Marcos López de Prado, Maureen O'Hara, and Zhibai Zhang*

February 2019

We thank Lee Cohn, Michael Lewis, Michael Lock and Yaxiong Zeng for useful discussion and help with preparing some of the data used in this study. We also thank seminar participants at Cornell University, including Pamela Moulton and Shawn Mankad; and at the Georgia State FinTech Conference, including our discussant at the conference, Haoxiang Zhu.

*Easley is in the Departments of Economics and Information Science, Cornell University and the University of Technology Sydney (UTS); López de Prado is with AQR Capital Management, and the Department of Operations Research and Information Engineering at Cornell University; O'Hara is in the Johnson College of Business, Cornell University and the University of Technology Sydney (UTS); Zhang is with the Tandon School of Engineering, New York University.

Microstructure in the Machine Age

Abstract

Understanding modern market microstructure phenomena requires large amounts of data and advanced mathematical tools. In this paper, we demonstrate how a machine learning algorithm can be applied to microstructural research. We find that simple microstructure measures designed to reflect frictions in a simpler market continue to provide insights into the process of price adjustment. We find that some of these microstructure features with apparent high explanatory power can exhibit low predictive power, and vice versa. We also find that some microstructure-based measures are useful for out-of-sample prediction of various market statistics, leading to questions about the efficiency of markets. Our results are derived using 87 of the most liquid futures contracts across all asset classes.

Keywords: Market microstructure, machine learning, features importance, MDI, MDA, futures, econometrics.

JEL Classification: C02, D52, D53, G14, E44.
AMS Classification: 91G10, 91G60, 91G70, 62C, 60E

Microstructure in the Machine Age

1. Introduction

One might have expected as markets became faster, market data became more copious, and technology superseded human participants, that the microstructure of markets would play an ever-decreasing role in explaining market behavior. The opposite is true. When time scales shrink to nanoseconds, how the market is structured turns out to be critical in predicting where the market is going. And when data explodes to mammoth dimensions, being able to characterize what variables related to market frictions can and should matter for market behavior, a particular focus of microstructure research, takes on even more significance. Yet, despite this continued importance, microstructure research faces some daunting challenges in this new era.[1] The ubiquity of computerized-trading, abetted by the rise of big data, has increased the complexity of trading strategies far beyond what is envisioned in simple microstructure models. Similarly, the empirical measures that fill the microstructure "tool box" were constructed based on simple with-in market relationships that may no longer hold in the high frequency world of cross-market trading. The problem, simply put, is that microstructure needs to evolve.

In this paper, we demonstrate how machine learning techniques can play an important role in that evolution. Much as microstructure research is often used to predict how trading will affect price and liquidity dynamics, machine learning can improve those predictions given complex data and computational constraints. Using a random forest machine learning algorithm, we investigate how well some standard empirical microstructure measures (termed "features" in machine learning parlance) predict variables of interest to market participants. Our focus is on a set of variables

[1] For more discussion, see O'Hara [2015].

typically used in electronic market making, dynamic market hedging strategies, and volatility estimation. Our purpose here is not to provide an exhaustive examination of market data predictability but rather to illustrate how machine learning can bring new insights to microstructure research by showing what features actually work for out-of-sample predictability. In doing so, we also provide clear evidence of the value of some extant microstructure variables for understanding the new dynamics of market behavior. It is worth emphasizing that machine learning algorithms are often highly non-parametric and do not pre-specify a functional form. The non-parametricity should not be viewed as a drawback, as the algorithms are designed to be adaptive so that they can extract patterns in data that parametric models may not recognize. As a result, machine learning algorithms often provide higher predictive power and therefore are better candidates for our investigation of predictability based on microstructure variables.

Our analysis draws on three generations of market microstructure models to provide specific measures as inputs to our machine learning investigation. These variables include the Roll measure, the Roll Impact, a volatility measure, Kyle's λ, the Amihud measure, and VPIN (the volume synchronized probability of informed trading). We focus on predicting six important outcomes of market price dynamics using a one-week forecast horizon: the sign of change of the bid-ask spread; sign of change in realized volatility; sign of change in Jacques-Bera statistic; sign of change in sequential correlation of realized returns; sign of changes in absolute skewness of returns; sign of changes in kurtosis of realized returns. We evaluate the importance of each feature using Mean-Decreased Impurity (an in-sample measure) and Mean-Decreased Accuracy (an out-of-sample measure) methods. We use five years of tick data from the 87 most liquid futures traded globally (including indices, currencies, commodities, short rates, and fixed income instruments). This extensive sample, one of the largest ever used in a microstructure analysis, epitomizes the big

data that can be brought to bear in machine learning analyses. This scale allows us to establish the validity and accuracy of our findings generally, and not merely for a specific contract or asset class.

Our research provides a number of results. As expected, we find that the various microstructure measures show different importance for in-sample and out-of-sample estimation, illustrating how variables that may have explanatory power in-sample need not have predictive power out-of-sample. Consistent with previous studies, all of the measures appear to have in-sample explanatory power. Across the six predicted variables, the Amihud measure, VIX and VPIN have the best performance in-sample, while VPIN has the best out-of-sample performance. For example, predicting the sign of the change in the bid-ask spread, in-sample results show that Amihud and VPIN consistently have the largest importance across all window sizes, whereas out-of-sample results show that VPIN predominates. Indeed, out-of-sample prediction results show that VPIN is the most important predictor for five variables, with the Roll measure dominating for the sixth (predicting the sign of the change in sequential correlation). We interpret these results as showing that simple measures designed to reflect market frictions still work in modern, complex markets dominated by machine-based trading. These results demonstrate not only the importance of particular microstructure-related variables, but also the possibility of successful prediction of future market dynamics. As we discuss, such predictions have wide applicability for areas such as risk management, dynamic trading strategies, and electronic market making.

Our paper joins a growing literature examining the implications of machine learning and big data for economic research. Varian [2014], Abadie and Kasy [2017], and Mullainathan and Spiess [2017] provide excellent discussions of how machine learning can be applied to analyze economic problems involving big data, while recent applications of such techniques can be found

in Bajari, Nekipelov, Ryan and Yang [2015] and Cavallo and Rigobon [2016]. In the finance area, Chinco, Clarke-Joseph and Ye [2018] apply LASSO techniques to make 1-minute ahead equity return forecasts; Rossi [2018] uses boosted regression trees to forecast stock returns and volatility; Krauss, Do, and Huck [2017] use machine learning for statistical arbitrage on the S&P 500; and López de Prado [2018] provides extensive analyses of financial machine learning techniques and applications. Gu, Kelly and Xiu [2018] applied multiple machine learning regression algorithms in asset pricing and find that non-linear regression methods can give rise to better R-squareds than econometric models. Our work contributes to this literature by showing how supervised machine learning techniques combined with metrics suggested by microstructure theories can help identify important market variables irrespective of functional form. We believe that machine learning's decoupling of the search for variables from the search for specification will be important for the development of microstructure research.

This paper is organized as follows. In the next section we set out the variables we are interested in predicting and the microstructure variables we use as inputs in our analysis. Section 3 provides a brief introduction to the random forest classification method and feature importance measures. We discuss two such measures: Mean Decreased Impurity (MDI) and Mean Decreased Accuracy (MDA). We also explain how we categorize realized outcomes in terms of binary labels. In Section 4 we discuss the data we use, how we transform the data into units of analysis called bars, and the microstructure variable definitions we use in the analysis. Section 5 then presents our empirical results and investigates their robustness with respect to various backward window sizes, alternative hyper-parameter configurations, and different bar types. Section 6 concludes by discussing the implications of our results for trading strategies, considers what we learn about

explanatory and predictive roles of microstructure variables, and suggests an agenda for future microstructure research in the machine age.

2. Microstructure variables and market movements

Microstructure models provide variables that indirectly measure the observable implications of market frictions. To the extent that these measures are successful they should predict the future values or movements in market metrics such as bid-ask spreads, volatility, and other variables related to the shape of the distribution of returns. Some models (which we will term "first generation") use price data for this task. Examples here are the Roll [1984] measure which uses price sequences to predict effective bid-ask spreads, Beckers' [1983] volatility estimation based on high-low prices, and the Corwin and Schultz [2011] bid-ask spread estimator. Second generation models focus on price and volume data, generating metrics such as the Kyle [1985] lambda, the Amihud [2002] measure, and Hasbrouck's [2009] lambda. Third generation models use trade data, inspiring metrics such PIN, the probability of informed trading (Easley et al [1996]) and VPIN, the volume-synchronized probability of informed trading (Easley et al [2011]). In our analysis, we evaluate the predictive power of measures representative of these three generations of microstructural models.

Being able to forecast future developments in the price process and liquidity has obvious importance, but less apparent is how well these standard microstructure measures work in current markets. The models that produce these measures are relatively simple and were designed at a time when markets were less complex than they are now. Those models do not provide much guidance about functional forms describing the relationship between any of these measures and price or liquidity dynamics. So imposing a particular functional form for this relationship, even a

flexible one, and applying standard econometric techniques to estimate it could potentially obscure any relationship.[2]

Our interest is in evaluating predictability using various microstructure variables. We begin with data about our microstructure variables (such as illiquidity, Kyle's lambda or VPIN) and data about the market measures (such as bid-ask spreads, volatility and the like) we are interested in predicting. However, unlike the standard approach in econometrics, we do not attempt to pre-specify an underlying data generating process, and so we do not attempt to estimate parameters of a model relating our microstructure measures to market measures. Our primary interest is in understanding which microstructure variables are useful for prediction and which ones are not useful. We are agnostic about the mechanism relating the variables in our data set to each other, as attempting to specify a mechanism, no matter how complex its structure or underlying probability space, is unnecessarily limiting for our data-exploratory purposes. We believe that this machine learning point of view is more powerful for the questions we want to ask; although we do recognize that for other interesting questions more closely related to developing an understanding of why one measure is a better predictor than another is, specifying a data generating process and applying standard econometric tools may be more productive.

Thus, we use machine learning to investigate the efficacy of a set of microstructure measures for forecasting a set of variables of wide interest in the market. We discuss in detail in Section 3 how the random forest algorithm we use works, but it is important to stress that we use the algorithm to predict the sign of changes in variables, rather than to provide actual point predictions.

[2] As Mullainathan and Spiess [2017] explain, econometric techniques are well suited for variance adjudication; however they often provide suboptimal forecasts. The reason is that the best forecast estimators may not be BLUE (best linear unbiased estimator). Unbiasedness is undoubtedly a useful property when the model is properly specified, however it may be a hindrance when important explanatory variables are missing or when the interaction between variables is not correctly modeled.

While this might seem of limited importance, we explain below why this is not the case and discuss how for our candidate variables such forecasts can be used in practice.[3]

1. *Sign of change of the bid-ask spread.*

When we expect the bid-ask spread to widen, an execution algorithm could use that expectation to increase the volume participation, thereby increasing the portion of the executed order before an increase in transaction costs materializes. Conversely, when we expect the bid-ask spread to narrow, an execution algorithm could use that expectation to decrease the volume participation and so execute a larger portion of the order after the fall in transaction costs takes place. The magnitude of the change in volume participation would be a function of our confidence in the forecast's accuracy.

2. *Sign of change in realized volatility.*

When we expect realized volatility to increase, an execution algorithm could use that expectation to increase the volume participation, in order to reduce the uncertainty of the average fill price (market risk). It is not necessarily true that we would like to decrease the volume participation if we expect a decrease in realized volatility, because by the time the volatility has decreased, prices may have drifted away from our target. In general, we would like to increase the volume participation rate if we forecast an increase in realized volatility, and reduce the volume participation rate after a decrease in realized volatility has already materialized.

3. *Sign of change in Jarque-Bera statistic.*

The Jarque-Bera statistic tests for the null hypothesis that observations are drawn from a Normal distribution. This is relevant for risk management purposes, as many risk models assume

[3] We consider only positive and negative changes. One could, and for an investment or trading strategy probably would want to, consider a finer partition of the set of changes at least taking into account a third category in which the change, either positive or negative, is small.

Normality of returns. A higher probability of non-Normal returns reduces our confidence in those models. For example, a risk manager may want to reduce the significance level (false positive rate, type I error probability) of his Gaussian models when returns are expected to be non-Normal.

4. Sign of change in kurtosis / Sign of change in absolute skewness of returns.

The Jarque-Bera statistic uses skewness and kurtosis to test for Normality of observations. This test implies a trade-off between skewness and kurtosis, in the sense that the test may not reject the null hypothesis of Normality when an increase in skewness is offset with a decrease in kurtosis. However, offsetting skewness with kurtosis is not without economic meaning. Because skewness is an odd moment, it deforms the Normal distribution by shifting its probability towards one side. One possible reason for this deformation is the presence of informed traders, who push prices in an attempt to fill orders before a piece of news is widely known. In contrast, because kurtosis is an even moment, it deforms the Normal distribution by shifting its probability symmetrically towards extreme events. One possible explanation for this deformation is a reduction of liquidity, as market makers reduce the size of their quotes in anticipation of a news release, hence increasing the likelihood of extreme outcomes on either side. From an execution and portfolio management perspective, it is important to differentiate between these two causes of non-Normality, and to forecast them separately.

5. Sign of change in sequential correlation of realized returns.

Another common assumption of risk models (for example in value-at-risk approaches) is that returns are serially uncorrelated. When returns are serially correlated, trends occur with a higher frequency than would be otherwise expected. This leads to greater potential drawdowns and time underwater. Like in the non-Normal case, a higher probability of serially correlated returns reduces our confidence in models that assume an uncorrelated structure. It would therefore be rational to

reduce the significance level of this kind of risk model when returns are expected to be serially correlated.

3. The random forest classification algorithm and feature importance measures

In this section, we introduce the random forest classification algorithm and how we use it to evaluate the predictive power of a set of explanatory variables. In machine learning, classification is the practice of using explanatory variables to predict a categorical/discrete target variable. It is analogous to regression in that both are fitted by minimizing an error function built on the explanatory and target variables in the training data set. However, in our machine learning problem the target variable is discrete (e.g., "yes" or "no"), and so the error functions popular for regression (e.g. mean-squared-error) are not viable. Instead, useful error functions include measures such as cross-entropy and information gain. We refer to the explanatory variables as features and the endogenous variable as a finite set of labels.

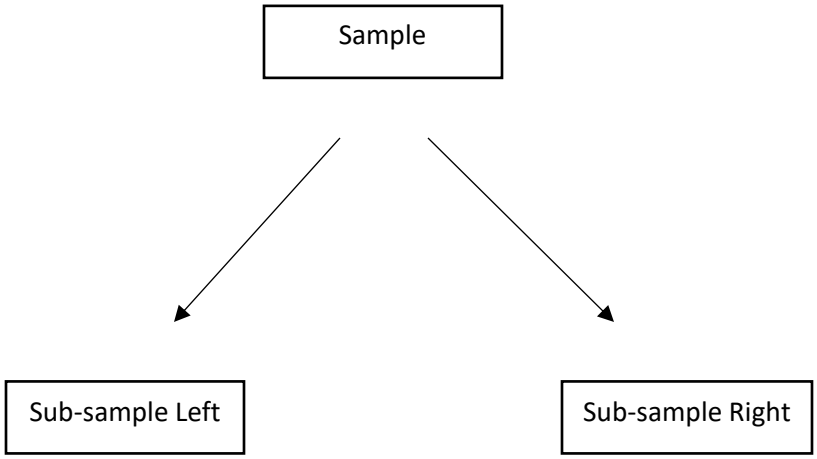
Among machine learning classification methods, random forest is one of the most robust and widely used algorithms. As Varian [2014] notes, "this method produces surprisingly good out-of-sample fits, particularly with highly nonlinear data."[4] It consists of a number of individual classifiers called decision tree algorithms and uses the mean of these trees' classifications as its prediction. As the number of low-correlated trees increases, the variance of the forest's forecasting error becomes smaller, hence reducing the chance of the algorithm overfitting the data.

Our machine learning algorithm is applied one futures contract at a time. So it operates on a data set, or sample, $\{(x_t, y_t)\}_{t=1}^{T}$, consisting of observations of features (x's) and a label (y) for the selected contract, with t indexing T observations.[5] The first step in creating a random forest is to

[4] See Varian [2014] pg. 14. This article provides a description of the random forest technique, as does López de Prado [2018], Chapter 6.

[5] We discuss creation of the sample in subsequent sections and the creation of a forest of trees later in this section.

build a decision tree by splitting the sample into two subsamples, and then splitting each of these subsamples into two subsamples, and so on. Graphically, the decision tree consists of numerous sequential splits, each of which takes the following form:



To create the split we first compute for each feature the information gain that would be created by splitting the sample using that feature. For any split of a sample S at node n in the tree into two subsamples, L and R, this information gain is

$$IG(S,n) = I(S) - \frac{N_L}{N_S} I(L) - \frac{N_R}{N_S} I(R)$$

where we use the Gini Index I(S)=$\sum_i \ p_i(1-p_i)$ as our purity measure for any data set S; p_i is the fraction of labels of the ith class in data set S; and N_S , N_L and N_R are the number of data points in the sample, the Left subsample and the Right subsample. The information gain from using a particular feature to split the sample is then defined to be the maximal gain that can be obtained by choosing a value of the feature and splitting the sample such that all data points with smaller values of that feature are in the Left subsample and those with larger values of that feature are in the Right subsample. Intuitively, the information gain is maximal when a feature is

able to split the data into two pure subsets (subsets with a single label). If the data could be split using the selected feature so that each subsample was pure (contained only increases or only decreases in the label) then the information gain would take on its maximum possible value I(S); while a less pure split will produce a smaller gain.

The actual split of sample S at node n is the one that maximizes the information gain over the choice of features used to create the split. This procedure is repeated for each subsample, and for each of the new subsamples created by any split until either a predetermined stopping criterion is reached or until no additional splits yield any information gain. We allow our trees to grow without bound; we consider the effect of bounds in the Robustness section. Although the information gain from alternative splits has to be computed many times, this approach is computationally tractable because each split is done using a greedy algorithm---there is no attempt to choose the split by looking ahead to implications of the current split for possible future information gains.

For each contract, and any sample for that contract, we create a random forest by modifying the simple procedure above in two ways. First, we create multiple decision trees and assign each tree a bootstrapped sample from our underlying sample. The averaging produced by bootstrapping reduces variance that could otherwise result from fitting a single tree to noise in the data set. Second, at each node in each tree we consider only two randomly selected features as candidates to determine the optimal split.[6] This second modification is done to take into account the possibility that one feature dominates splits even if a second highly correlated feature would produce similar,

[6] The number of features to consider is a parameter that can be adjusted. We use the standard rule of selecting $\text{int}(\sqrt{6}) = 2$ features where 6 is the total number of features we consider.

but slightly smaller information gains. Without restricting attention to randomly selected sets of features, we would attribute too much importance to the first feature relative to the second one. For any contract, and data set for that contract, we create 100 trees in our random forest. Finally, given any feature vector the prediction made by our random forest for the sign of the label is determined by majority vote across the trees in the forest.[7]

For decades, researchers have recognized the prevalence of hierarchical relationships in economic and financial systems. As Nobel laureate Simon [1962] put it, "the central theme that runs through my remarks is that complexity frequently takes the form of hierarchy, and that hierarchic systems have some common properties that are independent of their specific content." How exactly to measure the contribution of features to the hierarchal structure of the random forest this feature importance is a critical issue. In our analysis, we use two standard measures of feature importance – mean decreased impurity (MDI) and mean decreased accuracy (MDA).[8]

(1) Mean Decreased Impurity (MDI) based feature importance

MDI feature importance evaluates the information gain of each feature in all trees, weights them with the number of samples of each split, sums and then normalizes the score to be one in total. The importance of a feature is its contribution to the building of trees as quantified by the information gain on the splits. Given some data set the MDI for feature i in that data set is

$$MDI(i) = (1/100)\sum_{N} \sum_{n \in N: v(s_n)=i} p(t) IG(s_n, n)$$

where $v(s_n)$ is the feature used in the split of s_n with s_0 being the initial data set.

[7] For more detail on the creation of a random forest for financial data see Lopez de Prado [2018].
[8] The interested reader can find a detailed explanation of these techniques in López de Prado [2018], Chapter 8.

(2) Mean Decreased Accuracy (MDA) based feature importance

It is worth pointing out that MDI is an in-sample method, as it is derived from the same information used to fit the trees. This makes it similar to a *p*-value in regression analysis. In contrast, MDA evaluates feature importance out-of-sample, and, unlike MDI, it can be used with any classifier. The MDA procedure computes feature importance as follows: (a) the data set is split into non-overlapping training and testing sets; (b) a classifier is trained on the training set using all features; (c) predictions are made on the test set, and a performance measure (e.g. accuracy) is recorded as p_0; (d) values of one of the features, i, in the test set are randomly shuffled and predictions are re-made on the test set; (e) the performance associated with the shuffling of feature i is recorded as p_i. The MDA feature importance of i in the given data set is then

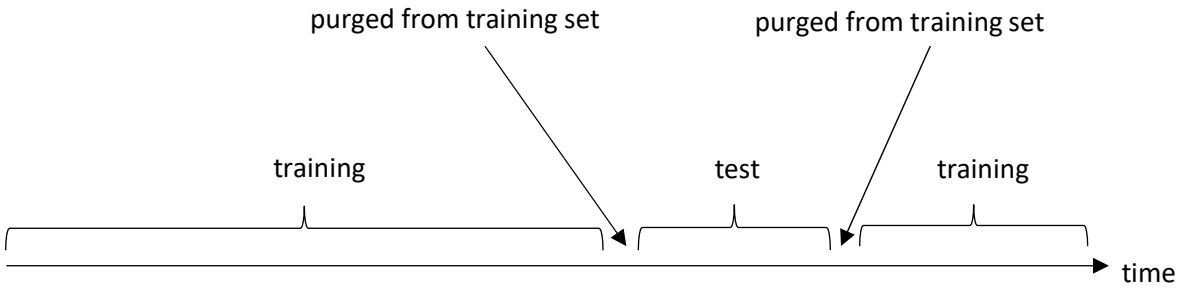
$$MDA(i) = \frac{p_0 - p_i}{p_0}$$

Thus, MDA's feature importance is determined by how the out-of-sample prediction worsens because of shuffling the values of a particular feature. The more deterioration there is in performance, the more important is this feature.

Finally, we turn to the issue of prediction accuracy. We define accuracy to be the number of correct predictions divided by the total number of predictions generated from a given data set split into a training set and a test set.[9] If we applied this idea once to our data set for a futures contract we would not use the information in the data set efficiently as we would lose the opportunity to train the random forest on all of the data. In particular, the data held out as a test set is not used in training. To use all of the data while still computing out-of-sample predictions we

[9] There are other measures of accuracy, in particular, ones that treat true and false positives or negatives asymmetrically. To make use of our approach in an investment strategy such alternative measures (such as precision, recall or F1 score) may be more useful than the simple accuracy measure we use here. Our goal is more generic and not tied to a particular investment strategy so we treat errors symmetrically.

apply a 10-fold purged Cross-Validation method to train the forest and compute accuracy Specific details of this approach can be found in López de Prado [2018], but summarily, we: (i) partition the entire data set into 10 intervals; (ii) take one as a test set; (iii) purge approximately one-week of data from the training set to remove observations that could contained leaked information from the test set (see the figure below); (iv) train the algorithm on the remaining data; and (v) make a prediction on the test set. This procedure is repeated 10 times (once for each test set) so the entire period is tested. Accuracy is computed across all the predictions from the 10 test sets.



4. Data

In this section, we turn to the data and the specific definitions of the labels and features we use in our analysis. We also address a variety of implementation issues. Our analysis uses dollar-volume bars, so we set out how we use tick data to form bars across the various contracts in our sample. Because we use futures data, our data has to "roll" across contract expirations to create a continuous price sequence. We describe how we effectuate that transition using a process akin to creating an ETF on the contract. Finally, we discuss measurement issues connected with viewing microstructure variables in volume bars as opposed to time-based units.

Our analysis is done on the 87 most liquid futures contracts traded globally, with specific details of each contract given in Table A.1 in the Appendix. We use these futures contracts for two

reasons. First, we are able to examine the universe of active futures, so there is no issue of selecting a sample out of some larger collection of financial assets. Second, we have complete trade data about the trade of these assets. Our sample period begins on July 2, 2012 and ends on October 2, 2017. Tick level data is available for most of these contracts over a longer period, but we are interested in VIX as a feature and the futures contract on VIX (ticker UX1) only began trading in July 2012. We note that two commodity contracts in our sample (IK1 and BTS1) have shorter sample periods beginning in October 2015.

A. Creating dollar-volume bars

We obtain tick level trade data for each futures contract and aggregate the data into intervals, or bars, based on dollar volume. Aggregating data into bars variously defined over time or volume increments is standard practice in industry and in academic research (see, for example, Engle and Lange [2001]; Easley et al. [2012]; Chakrabarty et al. [2012]; Easley et al. [2016]; Low et al. [2018]). Barardehi, Bernhardt and Davies [2019] also propose a trade time approach in their measurement of liquidity and show that it works better than a clock time approach. Easley and O'Hara [1992] demonstrated that the time between trades should be correlated with the existence of new information, providing our basis for looking at trade time (volume) instead of clock time. Information arrival results in patterns in volume, essentially akin to intra-day seasonalities.[10] By drawing a sample whenever the market exchanges a constant volume, we attempt to mimic the arrival to the market of news of comparable relevance. We use dollar-volume to allow comparability across the 87 contracts in our sample. Also, López de Prado [2018] presents evidence that the sampling frequency of dollar-volume bars is more stable than the sampling frequency of time bars or volume bars. One reason for this stability is that dollar-volume bars take

[10] Futures often trade over a 23.75 hour day and volume patterns are very pronounced.

into account price fluctuations, hence normalizing the dollar-value transacted across different time periods.

The τth bar is formed at tick t when

$$\sum_{j=t_0^\tau}^{t} p_j V_j \geq L,$$

where t_0^τ is the index of the first tick in the τth bar, p_j is the trade price at tick j, V_j is the trade volume at tick j, and L is a pre-determined threshold that gives roughly 50 bars per day for the year 2016.[11] Note that because average daily trading volume differs across contracts, the dollar volume in each bar will differ across specific futures contracts, but the average daily number of bars will not (in 2016). For each individual contract, on an active day, bars will fill faster and there can be more than 50 bars in a day; on an inactive day, bars will fill more slowly and there can be fewer than 50 bars in a day.

We compute each microstructure variable in our analysis at each bar τ, by applying a rolling "lookback window" of size W. For example, at bar τ we use bars within the set {τ-W+1, τ-W+2, …, τ-1, τ} to compute the microstructure variables and labels. In our analysis, we consider lookback windows ranging from 25 bars to 2000 bars.

B. Rolling Contracts

Since futures contracts expire, we need to "roll" the contracts (i.e. sell the expiring one and enter the new one) to form a price series as if it were a continuous instrument. To do so, we transform the price of a futures contract to the value of an ETF that perfectly tracks the futures with $1 initial capital.[12] To understand this process, consider the following example.

[11] We chose 2016 because it is the last full year before the end of our sample.
[12] For a detailed discussion of this technique, see López de Prado [2018].

Assume we would like to take a long position in the front contract of the E-mini S&P 500 futures (Bloomberg code: ES1 <Index>), from 01/02/2015 onwards. On 03/20/2015, the front month futures contract is soon to expire and we have to sell it and buy the then second month futures contract, hence "rolling to the next contract". In this rolling process, there is no change in the value of our investment except for the tiny transaction cost. However, there is usually a difference in raw price between the front and second month contracts. If the front month contracts were trading at $2000 while the second one was at $2020, then if we simply switch the price time series from front month to second month there is now a 1% difference. The machine learning algorithm would incorrectly think that there is a sudden jump in price, and consider it as some sort of a signal.

To avoid this problem, we produce a new time series we call the ETF price of the futurres series, which reflects the value of $1 invested in the futures contract assuming one can hold fractional shares. This series starts with 1, and its current value equals the investment's cumulative return (see Table A.2 for an example). When the futures contract rolls, one sells the old contract and invests all the money in the new contract. During this event, there is no change to the investment assuming zero transaction cost, so the ETF price is unaffected by the artificial change in raw price. Figure 1 provides a plot for ES1 Index's cumulative return and ETF price series.

In Appendix A.2 we provide the calculation details for this process. In the following analyses, for each futures contract we use the ETF-based price and the corresponding volume instead of the raw price and volume unless mentioned differently.

C. Features and Labels

As discussed in the previous section, we focus on a few well-known market microstructure variables. These features are all constructed from the bar data described above. One issue that

arises in our construction of these microstructure measures is that they initially were not defined using the same concepts of time periods or bars, or using lookback windows. Therefore, for each measure we have to adapt the original definition to our setting. We call these measures by their original names, but it we note that they are actually our translation of the measure to our setting.

More specifically, we have:

- Roll measure, given by

$$R_\tau = 2\sqrt{|cov(\mathbf{\Delta P_\tau}, \mathbf{\Delta P_{\tau-1}})|},$$
$$\mathbf{\Delta P_\tau} = [\Delta p_{\tau-W}, \Delta p_{\tau-W+1}, \dots, \Delta p_\tau],$$
$$\mathbf{\Delta P_{\tau-1}} = [\Delta p_{\tau-W-1}, \Delta p_{\tau-W}, \dots, \Delta p_{\tau-1}]$$

  where Δp_τ is the change in close price between bars $\tau - 1$ and τ, and W is the lookback window size.

- Roll impact, which is the Roll measure divided by the dollar value traded over a certain period, is

$$\tilde{R}_\tau = \frac{2\sqrt{|cov(\mathbf{\Delta P_\tau}, \mathbf{\Delta P_{\tau-1}})|}}{p_\tau V_\tau}.$$

  We evaluate the denominator at each bar and because of our bar formulation, the denominator varies very little between consecutive bars.

- Kyle's lambda is given by

$$\lambda_\tau = \frac{p_\tau - p_{\tau-W}}{\sum_{i=\tau-W}^{\tau} b_i V_i}$$

  where $b_i = \text{sign}[p_i - p_{i-1}]$, which is computed on bar level, and W is the lookback window size.

- Amihud's lambda is given by

$$\lambda_\tau^A = \frac{1}{W}\sum_{i=\tau-W+1}^{\tau} \frac{|r_i|}{p_i V_i},$$

  where r_i, p_i, V_i are the return, price and volume at bar i, and W is the lookback window size. Our version of Amihud's lambda measured using dollar volume bars is closely related to the Barardehi, Bernhardt and Davies [2019] trade time analogue of the Amihud measure.

- Volume-synchronized probability of informed trading is estimated as

$$VPIN_\tau = \frac{1}{W}\sum_{i=\tau-W+1}^{\tau} \frac{|V_i^S - V_i^B|}{V_i},$$

Where volume is signed using the BVC method, $V_i^B = V_i \mathrm{Z}\left[\frac{\Delta p_i}{\sigma_{\Delta p_i}}\right]$, $V_i^S = V_i - V_i^B$, and W is the look back window size. See Easley et al. [2016] for additional details.

- VIX index. We use VIX's front month futures (Bloomberg code: UX1 <Index>) tick level trade data to represent VIX. For each bar, we take the price of the closest tick of UX1 Index prior to that bar's timestamp as VIX's value.

Table 1 provides a correlation matrix of these variables over our sample period. As is apparent, while some of these variables are highly correlated, others are not, suggesting that they may have very different properties for forecasting purposes. Note that these correlations are all calculated based on dollar-volume bars. For VPIN, calculation in dollar-volume bars is a natural milieu but the other variables were traditionally derived based on fixed time intervals, such as daily bars. A natural concern is that this specification may bias our results against finding significance for these types of variables. As part of our robustness testing, we also calculated all variables using hourly time bars, and re-ran our analysis using this alternative data specification. We discuss these results in Section 5.4, but we note upfront that we find slightly greater accuracy using volume bars instead of time bars and that measures of feature importance are largely unchanged.

For the classification, we are interested in predicting the sign of the change in several important variables. Note that these labels are binary as their value is either positive +1 or negative -1, reflecting that we are forecasting the sign of changes in the relevant variable. In particular, we label observations according to:

- the sign of change in bid-ask spread. The spread is computed via the Corwin-Schultz estimator

$$S_\tau = \frac{2(e^{\alpha_\tau} - 1)}{1 + e^{\alpha_\tau}},$$
$$\alpha_\tau = \frac{\sqrt{2\beta_\tau} - \sqrt{\beta_\tau}}{3 - 2\sqrt{2}} - \sqrt{\frac{\gamma_\tau}{3 - 2\sqrt{2}}},$$

$$\beta_\tau = E\left[\sum_{j=0}^{1}\left[\log\left[\frac{H_{\tau-j}}{L_{\tau-j}}\right]\right]^2\right],$$

$$\gamma_\tau = \left[\log\left[\frac{H_{\tau-1,\tau}}{L_{\tau-1,\tau}}\right]\right]^2,$$

where $H_{\tau-j}$ and $L_{\tau-j}$ are the high and low prices at $\tau - j$, and $H_{\tau-1,\tau}$ and $L_{\tau-1,\tau}$ are the high and low prices over the 2 bars $(\tau - 1, \tau)$. For a given forecasting horizon h, the label is then

$$\text{sign}[S_{\tau+h} - S_\tau],$$

and effectively we are predicting whether the estimated spread will widen or narrow. Note there is a window size variable in computing β_τ.

- the sign of change in realized volatility, or simply

$$\text{sign}[\sigma_{\tau+h} - \sigma_\tau],$$

where σ_τ is the realized volatility of 1-bar returns over a lookback window of size *W*. In this case we are predicting whether the realized volatility will go up or down.

- the sign of change in Jarque-Bera statistics of realized returns

$$\text{sign}\big[JB[r_{\tau+h}] - JB[r_\tau]\big]$$

$$JB[r_\tau] = \frac{W}{6}\left(Skew_\tau^2 + \frac{1}{4}(Kurt_\tau - 3)^2\right),$$

where $Skew_\tau$ is the skewness and $Kurt_\tau$ is the kurtosis of realized returns over the lookback window of size *W*. This label can be viewed as a higher moment generalization of the realized return volatility above.

- the sign of change in the first order autocorrelation of realized returns

$$\text{sign}[AR_{\tau+h} - AR_\tau]$$

$$AR_\tau = \text{corr}[r_\tau, r_{\tau-1}],$$

where the correlation is evaluated over the returns of the past W bars.

- the sign of change in absolute skewness of realized returns

$$\text{sign}[Skew_{\tau+h} - Skew_\tau]$$

- the sign of change in kurtosis of realized returns

$$\text{sign}[Kurt_{\tau+h} - Kurt_\tau]$$

In the current analysis, we fix the forecast horizon h to be 250 bars ahead, which roughly corresponds to a week of trading. In section 5.4 we analyze a forecast horizon of 50 bars and find similar results.

5. Results and analysis

In this section, we first set out the parameters of our random forest classification methodology. We then present the main results of the paper, namely the feature importance of the microstructure variables. This is followed by a sensitivity analysis in which we tune the parameters of the random forest, and by various robustness check, including a comparison between dollar-volume-bar and time-bar results and a comparison of our machine learning results with the results of a logistic regression.

Our analysis uses a standard open-source machine learning software package, Scikit-learn (see Pedregosa et al. [2011]). We begin by specifying the configuration (hyper-parameters, in machine learning parlance) of the random forest machine learning algorithm. For our analysis, we choose the default values for the random forest's hyper-parameters[13]

- number of trees (*n_estimators*) = 100
- maximal features per split (*max_features*) = int($\sqrt{6}$) = 2
- sample weight (*class_weight*) = inverse of total number of samples in the sample's class ('*balanced*')

The number of trees is a parameter that controls how many decision trees the random forest contains. The maximal features here is square root of the total number of features, a common choice for random forest. Sample weight is the weight one assigns to each sample in the training class, and we use a balanced approach to reduce the bias that can come from label imbalance. We report results from an unregularized random forest (i.e. one in which the decision trees are allowed to grow without limit). In Section 5.4 we rerun the analysis using a regularized random forest to

[13] Scikit-learn's corresponding notations are in parenthesis.

check that our results are stable and robust and to allay fears that the original random forest is overfit.

5.1 MDI results

We first examine feature importance using MDI. As a reminder, MDI is an in-sample method that is based on the explanatory power of each feature and gives rise to normalized values for feature importance (all positive and sum to one). Table 2, Panels A through E, reports the MDI feature importance for each of the six predicted variables we consider in our analysis. Every row corresponds to a specific backward window size, as indicated by the first column. Each entry is formulated as "mean MDI feature importance score" ± "MDI feature importance score standard deviation", where the mean and standard deviation are evaluated across all 87 instruments. The highest importance is bolded for each window size.

To provide some intuition for how features contribute to in-sample explanation of features we provide in Figure 2 a scatter plot of predicted changes in spread for the ES1 index as a function of the VPIN and Roll measures. The random forest assigns a predicted increase or decrease in spread given any list of all of the features. Plotting this assignment against feature vectors produces a plot in R^6 which we project to R^2 in Figure 2. The right angle shape in that figure indicates a cut for spread predictions at just below 0.002 for the Roll measure and just above 0.05 for VPIN. The random forest predicts decreases in spread in the north-west quadrant and increases elsewhere.[14]

For bid-ask spread estimation, Panel A shows that the Amihud measure has the highest feature importance, followed by the VPIN metric. Feature importance increases with the window

[14] This appears noisy in the figure because we are conditioning on only two of the six features. Otherwise, there would be sharper regions.

size for Amihud, Kyle and VPIN, but not for the Roll measures or for VIX. The differential (and lower) performance of VIX relative to VPIN refutes the notion that VPIN is simply picking up volatility effects. The Amihud measure is also the most important for absolute skewness prediction (Panel E).

Panel B provides feature importance results for volatility prediction. Here we find mixed results depending on the window size. For both the shortest (25) and longest (greater than or equal to 1000) bars, VPIN dominates. Amihud is the most important if measured over 250-500 bar window, while VIX prevails for the 50 bar window (although VIX and VPIN are very similar for the 25 bar window as well). Feature importance for predicting the Jacques-Bera test in Panel C also shows mixed results. Overall, Amihud is most important, but for some windows VIX and VPIN predominate. The Amihud measure also does well when using longer window sizes for sequential correlation prediction (Panel D), while VIX dominates for shorter windows. Interestingly, the Roll measures, which might have been expected to do well with correlation change predictions, do not fare well. The results for kurtosis prediction again favor Amihud for long windows, but VIX and VPIN for shorter widows.

Overall, the data suggest that measured by in-sample performance the Amihud measure does best, with VPIN and VIX also having strong feature importance. The Kyle lambda and Roll measures are never the most important measure for predicting any of the six variables. However, all of the measures have similar MDI results for most of the variables. Perhaps most importantly, these measures all provide significant in-sample explanatory power even though they are simple measures designed for a simpler world.[15]

[15] It should be noted however that at each split in the trees we consider only two features. A feature that was never useful would have an MDI of zero, but one that sometimes is better than the single alternative it is compared with will have a non-zero MDI.

5.2 MDA results

We next turn to evaluating MDA feature importance. Table 3 summarizes the results of MDA feature importance for each predicted variable. In contrast with MDI, MDA is an out-of-sample method that captures the predictive power of each feature. Accordingly, MDA's outputs are not guaranteed to be positive (some features may actually be detrimental for forecasting purposes), nor are they normalized. As can be seen in the table, there are several entries with negative yet close to zero scores, and the interpretation is that they contribute little to the out-of-sample prediction despite the explanatory power they might have in sample. Every row corresponds to a specific lookback window as indicated by the first column. Each cell is formulated as "mean MDA feature importance score" ± "MDA feature importance score standard deviation", where the mean and standard deviation are evaluated across all 87 instruments. The highest importance is bolded for each window size. The last column summarizes the out-of-sample prediction accuracy averaged across all instruments.

For bid-ask spread prediction, VPIN has the highest feature importance for every widow size and it has the highest importance for 5 or 6 window sizes for kurtosis prediction and the Jarque-Bera test prediction. The Roll Measure dominates for sequential correlation prediction. For realized volatility prediction, the Roll measure is better for shorter windows, with VPIN a close second. Over longer lookback windows, however, VPIN again provides greater feature importance for realized volatility prediction while the Roll measure generally contributed little to out-of-sample prediction. Interestingly, VIX has little out-of-sample prediction power regardless of the window size. Finally, for absolute skewness prediction, the feature importance results are mixed, with Kyle lambda, VPIN, Roll Impact, and Roll Measure each having greater importance for specific window sizes.

We interpret these results as providing support for the predictive power of microstructure measures that reflect frictions in the market. VPIN is generally the most important among these features at predicting variables that should be influenced by the presence of information-based trade: spread and measures of fat tails in the distribution of returns. The Roll measure is created from correlation in price changes and so it is not surprising that this measure has some explanatory power for serial correlation in returns. Finally, although we include VIX in our set of features, it is not intended to reflect microstructure frictions and so it is not surprising that it has little explanatory power for the variables we attempt to predict.

The overall accuracy levels in Table 3 suggest that our machine learning algorithm is capturing something of value. For binary financial time series classification, a classifier often gives accuracy around 0.5. This standard inability to do better than random guessing is consistent with the efficient market hypothesis: for liquid markets, the market is efficient most of the time and acts like a random walk. So anything above 0.5 can be viewed as capturing a potential inefficiency of the market and so is a positive result. With the exception of the bid-ask spread estimation, our out-of-sample accuracy levels reach highs ranging from 0.54 to 0.61 (depending on the lookback windows) which by financial machine learning standards is very good.[16] The bid-ask spread accuracy is not as good. We conjecture that this is due to the lack of an observable bid-ask spread in futures; we impute one using the Corwin-Schultz estimator. It may be that the errors in the technique itself as applied to futures make estimation via the random forest methodology ineffective. Alternatively, it may be that the dollar-volume bar approach taken here is not well

[16] For example, see Krauss et al [2017] who in a similar binary classification problem obtain accuracy levels between 0.50 and 0.55.

suited to this particular estimation. We investigate this possibility in Section 5.4 where we re-run our analysis using time bars.

5.3 Why are the MDI and MDA results so different?

Our finding that microstructural features with good explanatory power can have poor predictive power, and vice versa, may be surprising at first. The reason is, in the MDI feature importance analysis, each tree is fit on the entire sample, and the inference is conducted on the output of that fit. In the MDI approach, the trees are not exposed to out-of-sample, never-seen-before data points. As a result, MDI explains the past, even if each label was determined after the associated feature was observed. This is not dissimilar to the way inference is conducted in standard econometric approaches: A particular functional form is fit on an entire sample, and the estimated coefficients are subjected to a number of hypothesis tests. In a sense, MDI is an econometric-like feature important analysis, analogous to p-values of estimated betas. In an MDA analysis, the trees are not fit on the entirety of the data. Instead, each tree is fit on a fraction of the data, and after the fit has taken place, the tree is exposed to a never-seen-before sample. This type of K-fold cross-validation analysis, although commonplace in the machine learning literature, is less common in the market microstructure literature.

That MDI and MDA have such different results on microstructural features should give researchers pause. Most of the empirical research on market microstructure has been built on in-sample, MDI-like methods, absent of systematic cross-validation. When in-sample analyses are overfit to the entire sample, some features appear to be more important than they truly are for out-of-sample prediction. It is essential to recognize that an econometric forecasting specification, when fitted on the entire sample, leads to in-sample (MDI-like) results that may be overfit. In other

words, even though the regression's specification attempts to forecast a variable, the resulting inference can be useless for forecasting purposes.

5.4 Sensitivity to hyper-parameters, time periods and forecast windows

As the random forest algorithm is highly non-parametric and can be tuned easily, one should ask about the stability of the results above with respect to tuning of the model parameters. After all, if the feature importance changes drastically when a random forest is constructed differently, then the results are not consistent. For this reason, we conduct multiple sensitivity tests for the feature's importance. All of our tests confirm that the feature importance score is consistent across different parameters, models and time.

First, we tune two different model parameters intrinsic to all tree-based machine learning algorithms: maximal depth and minimal weight fraction per leaf. Changing these parameters transforms our unregularized random forest into a regularized random forest, and this allows us to check for consistency of our results. The first parameter sets a depth threshold (the maximal number of sequential splits) for all decision trees that compose a random forest. For instance, if we set the maximal depth to be 5, then each tree cannot go beyond 5 sequential splits.[17] After adding this parameter to the random forest, we compute the MDA feature importance correlation between the original random forest and the regularized random forest across all 87 instruments. As shown in Table 4, correlation coefficients for every predicted variable and window size are virtually one. This indicates that the feature importance results are consistent and robust to changes in the tree depth hyper-parameter.

[17] In scikit-learn library, the parameter is controlled by argument "*max_depth*", and we set *max_depth* = 5.

The second parameter, which controls the least sample fraction on a leaf required to stop splitting, has a similar functionality.[18] A leaf is the name given to the node at the end of each branch or split. Restricting the minimal weight fraction per leaf essentially limits how big the tree can grow and thus limits the chances of overfitting. Again, we compute the MDA feature importance correlation between the unregularized and the regularized random forest. These results are given in Table 5. Just like the case in Table 4, correlation coefficients for every predicted variable and window size are close to one, which further confirms the robustness of our feature importance analysis.

Since our data are time series, another question is whether the feature importance is stationary across time. For instance, is it possible that VPIN is good at predicting realized kurtosis when the market is volatile, and not otherwise? To answer this, we run the MDA test presented in 5.2 on a yearly basis. More specifically, we split the data set into 5 parts in chronological order. Since the total length of data in time is 5 years, each part covers a year long period. For simplicity, we only show the results for 250 bar lookback window in Table 6. It is evident that the feature importance, especially the ranking does not vary much across time, indicating that the feature importance is stationary. A related question is whether the feature importance is stationary across different instruments. This is partially proven by the small standard deviations in the feature importance shown in Tables 2 and 3. In addition, we include a list of feature importance for kurtosis prediction per instrument, with lookback window fixed at 250 bars in Table A.3.

Finally, we ask about stability of our results with respect to the forecast window. All of our results are for a 250 bar forecast window and it is worth asking how feature importance changes as the window size varies. Table 9 provides the correlation in MDA feature importance results

[18] In scikit-learn library, the parameter is controlled by argument "*min_weight_fraction_leaf*". .

between a 250 bar forecast window and a 50 bar forecast window. As the table shows, most entries are high (particularly for short lookback windows) which indicates that the results are reasonably stable across different forecast scales.

5.5 Dollar-volume-bar versus time-bar accuracy

All of the analysis above is based on a dollar-volume bar formulation. These bars have the desirable feature of aligning the sampling of data with the arrival of information, which seems an appropriate property for the high-frequency world characterizing futures trading. There are, of course, other bar types that could be used, and in this subsection we compare the accuracy using dollar-volume bars with the accuracy that results from using another popular bar method, namely the time-bar method. A time-bar is formed when the difference between the close tick and open tick's timestamps exceeds a predefined value. In particular, we formulate hourly time-bars for all the futures instruments and apply the same cross-validation with the same random forest configuration. The results of out-of-sample prediction accuracy averaged over all instruments are given in Table 7.

As shown in the table, the accuracy results are very close for the two metrics. For four of the metrics, dollar-volume bars have higher accuracy, while for the Jarque-Bera test and bid-ask spread time bars are slightly more accurate. This similarity is important for allaying fears that variables originally calculated over fixed time intervals may be distorted when cast in a volume-based metric. Additionally, we find that even though overall the time-bar formulation has slightly lower accuracy, in many cases it gives rise to similar feature importance ranking with dollar-volume bars. For example, in Figure 3 we present the MDA feature importance for both bar methods for kurtosis prediction using a window size of 50 bars. The similarity in feature importance ranking is evident. Finally, Table 7 shows that regardless of how we measure bars,

the accuracy of out-of-sample bid-ask spread prediction is low. Thus, prediction difficulties here are not due to bar measurement issues. As noted earlier, we believe a more compelling explanation lies in the construction of this variable.

5.5 Logistic Regression

Next, we consider a different classification model, namely, logistic regression for a model-based sensitivity test. A logistic regression models the logarithm of the odds of our two labels with a linear functional form. When the classification label is binary denoted as {0,1}, the prediction probability for the two classes is given by

$$p(0|\vec{x}) = \frac{1}{1 + e^{-\vec{w}\cdot\vec{x}}}, \;\; p(1|\vec{x}) = 1 - \;\; p(0|\vec{x}),$$

where $\vec{x}$ is the feature vector and the coefficient vector $\vec{w}$ is obtained through a regularized maximum likelihood fit. For each sample, the prediction is the class label with the higher prediction probability. Logistic regression is another commonly used approach because of its simplicity and parametricity. It does not have an in-sample feature importance analysis as MDI. Nonetheless, we can apply MDA feature importance and compare it to the random forest result. In Table 8 we present the MDA feature importance correlation between logistic regression and random forest.[19] In general, the correlation between two algorithms is high (greater than 0.6), although when the lookback window size is large the correlation declines. Prediction accuracy with the logistic is also similar to what we obtain with machine learning with the logistic approach typically being slightly more accurate (see Appendix A.4).

To shed more light on the difference between the logistic regression and random forest approaches we provide a scatter plot in Figure 4 for the logistic regression's predicted bid-ask

[19] Detailed results from the logistic approach are provided in the Appendix, Table A.4.

spread as function of the two most important out-of-sample features, VPIN and the Roll Measure. Figure 2 provided a scatter plot of predictions for the random forest approach as a function of these two variables. The two plots are similar with the main difference being in the shape of the decision boundary. Both plots illustrate predictions that are in line with our intuition, e.g. higher VPIN leads to spread increasing.

We view the similarity of the results obtained with these two quite different approaches as further evidence that the microstructure frictions our features attempt to measure are real and that they have implications for the process of price adjustment. There is no apparent reason for why the logistic model with log odds given by a linear function of our microstructure features should fit the data reasonably well, but apparently it does as overall the prediction results are at least as strong as those we obtain with the hierarchical random forest approach. Of course, without having first done the random forest analysis we would not have known that the logistic model offers a good specification. In other words, the random forest sets a non-parametric benchmark that a classical model can beat by injecting structural information into the forecasting problem. This exemplifies our view that machine learning algorithms do not replace classical methods, but rather complement the use of those classical methods by de-coupling the search for specification from the search for important variables.

6. **Conclusion and future directions**

In this study we have attempted to shed light on the importance of various microstructure features for explanatory and forecasting purposes. The six variables we wish to explain and predict are highly relevant to market makers and portfolio managers: Bid-ask spread, realized volatility, Normality, skewness, kurtosis and serial correlation. We apply machine learning methods in order to capture the complexity inherent to high-frequency data, without concerning ourselves at this

point with determining a parametric structure to characterize the complex relationship between variables. We provide clear evidence that some extant microstructure variables have value for predicting the new dynamics of market behavior. At that same time, however, we find that other popular microstructure variables can have high explanatory power (in-sample), and yet fail to provide forecasting power (out-of-sample).

We believe these findings have important implications for future microstructure research. Foremost among these implications is good news: our results clearly show that market frictions continue to play an important role in affecting market dynamics and that extant microstructure measures capture (to varying extents) these dynamic effects. Thus, despite the complexity of current markets, frictions such as asymmetric information, or illiquidity arising from constraints on market maker risk bearing, or endogenous patterns arising from algorithms programmed to hide in particular market structures, all continue to affect price dynamics as predicted by microstructure research. More good news is that the efficacy of these microstructure variables in capturing these effects appears to be remarkably robust. Our out-of-sample forecasting results are virtually the same whether we use time clocks or volume clocks, shorter samples or longer, regularized or unregularized forests, even simple logistic models versus hierarchal machine learning – the rankings of which variables matter most stay the same. These findings should be helpful in thinking about the type of models (and measures) we need to work on to capture better market dynamics.

There are other implications to consider as well. Since most empirical research in the market microstructure literature follows an in-sample procedure, without out-of-sample cross-validation, it is possible that some established empirical results are artificial. To determine this, however, requires more extensive study and new empirical analytics. The machine learning

approach taken here is one such direction, but there are many new approaches that seem well-suited to analyses of complex market structures.

At a more fundamental level, the high out-of-sample accuracy we have achieved appears to indicate that markets are less efficient than is generally believed. For microstructure researchers, efficiency has long been a problematic concept; over short intervals, prices are not random walks, and even the concept of a price is tricky given that it may differ depending on whether you want large or small amounts, are a buyer or a seller, etc. Our findings here, however, are more concrete and troubling. Using machine learning techniques, successful forecasting of price process dynamics using simply past data on market microstructure features is both feasible and accurate. From a practical perspective, this suggests increased research on ways to exploit this information in profitable trading strategies. From a broader perspective, these results highlight the changing role played by trading and trading strategies in affecting asset price dynamics. Recognizing these trading dynamics may be particularly useful for asset pricing research.

Finally, we suggest a fruitful direction for future machine learning microstructure research. In particular, while our research here draws on extensive data involving what is essentially the entire futures market, our analysis looks at the within-market effects of our microstructure variables. That is, we look at how the various microstructure measures perform on each individual futures contract and then aggregate across all 87 contracts to find our results. Yet, as noted in the introduction, cross-market activity (and particularly cross-market market making) is now the norm, suggesting that there should be important cross effects of measures such a VPIN or Amihud in one market or other markets. Such an investigation seems both promising and only really feasible using modern machine learning techniques, and we hope to address this in future research.

References

Abadie, A. and M. Kasy (2017). The risk of machine learning. Working Paper.

Amihud, A., 2002, Illiquidity and stock returns: cross-section and time-series effects, Journal of Financial Markets, 5(1), 35-56.

Barardehi, V., D. Bernhardt and R. Davies, 2019, "Trade-Time Measures of Liquidity," Review of Financial Studies, 32(1), 126-179.

Bostic, W., Jarmin, R., and B. Moyer, "Modernizing Federal Economic Statistics, American Economic Review, 106(5), 161-164.

Cavallo, A. and R. Rigobon, 2016, "The Billion Prices Project: Using Online Prices for Measurement and Research," Journal of Economic Perspectives, 30(4), 171-198.

Chinco, A., A. Clark-Joseph, and M. Ye, 2018, "Sparse Signals in the Cross-section of Returns," Journal of Finance, forthcoming.

Corwin, S. and P. Schultz, 2012, A Simple Way to Estimate Bid-Ask Spreads from Daily High and Low Prices," Journal of Finance, 67(2), 719-759.

Easley, D., N. Kiefer, M. O'Hara, and J. Paperman, 1996, "Liquidity, Information, and Infrequently Traded Stocks", Journal of Finance, September 1996.

Easley, D., M. Lopez de Prado, and M. O'Hara, 2012, "The Volume Clock: Insights into the High Frequency Paradigm," Journal of Portfolio Management, (with D. Easley and M. Lopez de Prado), Winter.

Easley, D., M. Lopez de Prado, and M. O'Hara, 2013, "Flow Toxicity and Liquidity in a High Frequency World," Review of Financial Studies, 25(5), 1457-1493.

Easley, D., M. Lopez de Prado, and M. O'Hara, 2015, "Optimal Execution Horizon," Mathematical Finance, 25(3), July, 640-672.

Easley, D., M. Lopez de Prado, and M. O'Hara, 2016, "Discerning Information from Trade Data," Journal of Financial Economics, 120 (2), May, 269 – 286.

Easley, D. and M. O'Hara, 1992, "Time and the Process of Security Price Adjustment," Journal of Finance, 47 (2), June, 577-607.

Hasbrouck, J., 2009, "Trading Costs and Returns for U.S. Equities: Estimating Effective Costs from Daily Data," Journal of Finance, 63(3), 1445 – 1474.

S. Gu, B. Kelly, D. Xiu, "Empirical Asset Pricing via Machine Learning," Chicago Booth Research Paper No. 18-04

C. Krauss, X. A. Do and N. Huck, "Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500", European Journal of Operational Research, Volume 259, Issue 2, 1 June 2017, Pages 689-702

Kyle, A. P., 1985, "Continuous Auctions and Insider Trading," Econometrica, 53(6) 1315-1335

Lopez de Prado, M., 2018, Advances in Financial Machine Learning, (Wiley, NY)

Low, R., Li, T. and T. Marsh, 2018, BV-VPIN: Measuring the impact of order flow toxicity and liquidity on international equities markets, Working paper.

Mullainathan, S. and J. Spiess, 2017, "Machine Learning: An Applied Econometric Approach," Journal of Economic Perspectives, 31(2), 87-106

O'Hara, M. 2015, "High Frequency Market Microstructure", Journal of Financial Economics, *116 (2), 257-270.*

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011), 117 "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research 12, 2825-2830.

Roll, R., 1984, A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market, Journal of Finance, 39(4), 1127-1139.

Rossi, A., 2018, "Predicting Stock Market Returns with Machine Learning," Working paper, University of Maryland.

Simon, H. A., 1962, "The architecture of complexity." Proceedings of the American Philosophical Society, 106(6), 467-482

Varian, H. R., 2014, "Big Data: New Tricks for Econometrics", Journal of Economic Perspectives, 28(2), Spring, 3-28.

Table 1 Correlation Matrix of Microstructure Variables

	Roll	Roll_impact	Kyle lambda	Amihud	VPIN	UX (VIX)
Roll	1.0000	0.9275	0.0001	0.3441	0.0190	0.1574
Roll_impact	0.9275	1.0000	0.0001	0.3255	0.0141	0.1506
Kyle lambda	0.0001	0.0001	1.0000	0.0002	-0.0001	0.0008
Amihud	0.3441	0.3255	0.0002	1.0000	0.0776	0.2971
VPIN	0.0190	0.0141	-0.0001	0.0776	1.0000	0.0320
UX (VIX)	0.1574	0.1506	0.0008	0.2971	0.0320	1.0000

Table 2 MDI feature importance

Panel A: MDI feature importance for bid-ask spread prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN
25 bars	**0.181** ± 0.001	0.17 ± 0.001	0.155 ± 0.002	0.15 ± 0.002	0.165 ± 0.002	0.179 ± 0.001
50 bars	**0.184** ± 0.001	0.17 ± 0.001	0.153 ± 0.002	0.15 ± 0.002	0.167 ± 0.001	0.177 ± 0.0
250 bars	**0.194** ± 0.001	0.171 ± 0.001	0.15 ± 0.002	0.148 ± 0.002	0.157 ± 0.001	0.18 ± 0.001
500 bars	**0.197** ± 0.001	0.171 ± 0.001	0.149 ± 0.002	0.148 ± 0.002	0.151 ± 0.001	0.184 ± 0.001
1000 bars	**0.198** ± 0.001	0.172 ± 0.001	0.148 ± 0.003	0.148 ± 0.002	0.146 ± 0.001	0.187 ± 0.001
1500 bars	**0.197** ± 0.001	0.173 ± 0.001	0.148 ± 0.003	0.148 ± 0.002	0.145 ± 0.001	0.19 ± 0.001
2000 bars	**0.197** ± 0.001	0.172 ± 0.001	0.147 ± 0.003	0.147 ± 0.002	0.145 ± 0.001	0.192 ± 0.002

Panel B: MDI feature importance for realized volatility prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN
25 bars	0.169 ± 0.003	0.157 ± 0.002	0.149 ± 0.003	0.157 ± 0.003	0.178 ± 0.003	**0.178** ± 0.004
50 bars	0.185 ± 0.001	0.155 ± 0.001	0.135 ± 0.002	0.144 ± 0.002	**0.205** ± 0.001	0.175 ± 0.003
250 bars	**0.223** ± 0.002	0.148 ± 0.001	0.098 ± 0.002	0.119 ± 0.002	0.22 ± 0.002	0.192 ± 0.002
500 bars	**0.234** ± 0.001	0.146 ± 0.001	0.089 ± 0.002	0.114 ± 0.002	0.206 ± 0.001	0.211 ± 0.002
1000 bars	0.234 ± 0.002	0.143 ± 0.002	0.085 ± 0.002	0.117 ± 0.003	0.18 ± 0.002	**0.241** ± 0.004
1500 bars	0.23 ± 0.002	0.143 ± 0.002	0.083 ± 0.002	0.119 ± 0.003	0.168 ± 0.002	**0.257** ± 0.004
2000 bars	0.226 ± 0.002	0.142 ± 0.002	0.082 ± 0.002	0.119 ± 0.003	0.165 ± 0.002	**0.267** ± 0.004

Table 2 (continued)

Panel C: MDI feature importance for Jarque-Bera test prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN
25 bars	**0.184** ± 0.002	0.167 ± 0.002	0.14 ± 0.001	0.143 ± 0.002	0.183 ± 0.002	0.183 ± 0.004
50 bars	0.19 ± 0.001	0.161 ± 0.001	0.13 ± 0.001	0.135 ± 0.001	**0.203** ± 0.001	0.181 ± 0.003
250 bars	0.22 ± 0.001	0.149 ± 0.001	0.098 ± 0.001	0.118 ± 0.002	**0.22** ± 0.002	0.195 ± 0.002
500 bars	**0.233** ± 0.002	0.147 ± 0.001	0.091 ± 0.002	0.117 ± 0.002	0.206 ± 0.001	0.206 ± 0.002
1000 bars	**0.24** ± 0.001	0.147 ± 0.001	0.089 ± 0.002	0.121 ± 0.003	0.181 ± 0.001	0.221 ± 0.002
1500 bars	**0.241** ± 0.002	0.147 ± 0.001	0.088 ± 0.002	0.124 ± 0.003	0.168 ± 0.001	0.232 ± 0.002
2000 bars	0.24 ± 0.002	0.144 ± 0.001	0.088 ± 0.002	0.125 ± 0.003	0.161 ± 0.001	**0.241** ± 0.002

Panel D: MDI feature importance for sequential correlation prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN
25 bars	0.183 ± 0.002	0.163 ± 0.002	0.142 ± 0.002	0.15 ± 0.002	**0.191** ± 0.003	0.171 ± 0.002
50 bars	0.188 ± 0.002	0.159 ± 0.002	0.132 ± 0.002	0.141 ± 0.002	**0.206** ± 0.002	0.174 ± 0.002
250 bars	0.215 ± 0.002	0.146 ± 0.001	0.109 ± 0.002	0.131 ± 0.003	**0.216** ± 0.002	0.184 ± 0.001
500 bars	**0.228** ± 0.001	0.146 ± 0.001	0.1 ± 0.002	0.125 ± 0.003	0.2 ± 0.001	0.201 ± 0.001
1000 bars	**0.233** ± 0.002	0.146 ± 0.001	0.099 ± 0.002	0.133 ± 0.003	0.176 ± 0.001	0.213 ± 0.002
1500 bars	**0.234** ± 0.002	0.143 ± 0.001	0.1 ± 0.002	0.139 ± 0.003	0.163 ± 0.001	0.221 ± 0.002
2000 bars	**0.232** ± 0.002	0.143 ± 0.001	0.1 ± 0.003	0.143 ± 0.003	0.156 ± 0.001	0.224 ± 0.002

Table 2 (continued)

Panel E: MDI feature importance for absolute skewness prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN
25 bars	**0.181** ± 0.002	0.172 ± 0.002	0.141 ± 0.001	0.144 ± 0.001	0.18 ± 0.002	0.181 ± 0.003
50 bars	0.189 ± 0.001	0.169 ± 0.001	0.132 ± 0.001	0.136 ± 0.001	**0.2** ± 0.001	0.174 ± 0.002
250 bars	**0.219** ± 0.001	0.155 ± 0.001	0.101 ± 0.001	0.12 ± 0.002	0.218 ± 0.001	0.187 ± 0.001
500 bars	**0.23** ± 0.001	0.152 ± 0.001	0.096 ± 0.002	0.12 ± 0.002	0.201 ± 0.001	0.201 ± 0.001
1000 bars	**0.237** ± 0.001	0.152 ± 0.001	0.092 ± 0.002	0.124 ± 0.003	0.18 ± 0.001	0.216 ± 0.001
1500 bars	**0.236** ± 0.001	0.153 ± 0.001	0.091 ± 0.002	0.127 ± 0.003	0.168 ± 0.001	0.225 ± 0.002
2000 bars	**0.238** ± 0.001	0.151 ± 0.001	0.09 ± 0.002	0.127 ± 0.002	0.161 ± 0.001	0.232 ± 0.002

Panel F: MDI feature importance for kurtosis prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN
25 bars	0.181 ± 0.002	0.159 ± 0.002	0.135 ± 0.002	0.139 ± 0.002	0.182 ± 0.001	**0.205** ± 0.003
50 bars	0.188 ± 0.001	0.157 ± 0.001	0.127 ± 0.001	0.133 ± 0.001	**0.202** ± 0.001	0.193 ± 0.002
250 bars	0.219 ± 0.001	0.149 ± 0.001	0.097 ± 0.001	0.117 ± 0.002	**0.221** ± 0.001	0.197 ± 0.001
500 bars	**0.233** ± 0.001	0.147 ± 0.001	0.091 ± 0.002	0.117 ± 0.002	0.206 ± 0.001	0.207 ± 0.002
1000 bars	**0.24** ± 0.001	0.146 ± 0.001	0.089 ± 0.002	0.121 ± 0.003	0.182 ± 0.001	0.221 ± 0.002
1500 bars	**0.241** ± 0.002	0.146 ± 0.001	0.088 ± 0.002	0.124 ± 0.003	0.168 ± 0.001	0.232 ± 0.002
2000 bars	0.24 ± 0.002	0.145 ± 0.001	0.088 ± 0.002	0.125 ± 0.003	0.161 ± 0.001	**0.241** ± 0.002

Table 3 MDA feature importance

Panel A: MDA feature importance for bid-ask spread prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
25 bars	0.0033 ± 0.00084	0.0042 ± 0.00068	-0.0031 ± 0.00108	0.0174 ± 0.00154	0.0011 ± 0.00059	**0.0248** ± 0.00142	0.4535
50 bars	0.0042 ± 0.00094	0.0045 ± 0.00078	-0.004 ± 0.00088	0.0126 ± 0.00142	0.0002 ± 0.00056	**0.0167** ± 0.00117	0.4525
250 bars	0.0048 ± 0.00125	0.0018 ± 0.00088	-0.0047 ± 0.00087	0.0031 ± 0.00124	0.0021 ± 0.0009	**0.0161** ± 0.00179	0.4572
500 bars	-0.0001 ± 0.00094	-0.0003 ± 0.00082	-0.003 ± 0.00103	-0.003 ± 0.00097	0.0031 ± 0.00115	**0.0268** ± 0.00214	0.4587
1000 bars	-0.002 ± 0.00106	0.0007 ± 0.00084	-0.002 ± 0.00103	-0.0033 ± 0.00108	0.0039 ± 0.00138	**0.0198** ± 0.00166	0.4546
1500 bars	-0.0023 ± 0.00094	-0.0007 ± 0.00083	-0.002 ± 0.001	-0.0017 ± 0.00109	0.0053 ± 0.00133	**0.015** ± 0.00114	0.4513
2000 bars	-0.0005 ± 0.00093	0.0012 ± 0.00093	0.0002 ± 0.00099	-0.0025 ± 0.00115	0.0058 ± 0.00148	**0.0102** ± 0.0011	0.4498

Panel B: MDA feature importance for realized volatility prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
25 bars	0.0013 ± 0.00289	0.0185 ± 0.00138	0.0237 ± 0.00211	**0.0558** ± 0.00288	-0.0006 ± 0.00095	0.0531 ± 0.00482	0.61
50 bars	0.0057 ± 0.0011	0.0133 ± 0.00125	0.019 ± 0.00155	**0.0435** ± 0.00229	0.0004 ± 0.00093	0.0402 ± 0.00432	0.5813
250 bars	0.0163 ± 0.00304	0.006 ± 0.00163	0.0063 ± 0.00148	**0.025** ± 0.00245	0.0002 ± 0.0022	0.0172 ± 0.0037	0.5493
500 bars	0.0133 ± 0.00373	0.0005 ± 0.00229	-0.0029 ± 0.00149	0.0028 ± 0.00231	-0.002 ± 0.00251	**0.0307** ± 0.0044	0.5399
1000 bars	0.0063 ± 0.00315	0.0024 ± 0.00265	-0.0029 ± 0.00124	-0.0002 ± 0.00251	0.01 ± 0.00288	**0.0477** ± 0.0056	0.5578
1500 bars	0.002 ± 0.00377	0.004 ± 0.00293	-0.0072 ± 0.00187	-0.0034 ± 0.0032	0.0101 ± 0.00311	**0.0513** ± 0.00564	0.559
2000 bars	0.0036 ± 0.00386	0.002 ± 0.00293	-0.0076 ± 0.00191	-0.0111 ± 0.00287	0.0187 ± 0.00418	**0.056** ± 0.00544	0.5668

Panel C: MDA feature importance for Jarque-Bera test prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
25 bars	0.0027 ± 0.00083	0.0219 ± 0.00235	0.0041 ± 0.0008	0.0095 ± 0.00122	0.0001 ± 0.00082	**0.0334** ± 0.00996	0.5406
50 bars	0.0008 ± 0.00091	0.0132 ± 0.00189	0.0032 ± 0.00086	0.0083 ± 0.00101	0.0001 ± 0.00101	**0.0428** ± 0.00589	0.5416
250 bars	0.002 ± 0.00264	-0.0006 ± 0.0018	0.0004 ± 0.00122	0.0086 ± 0.00182	-0.0024 ± 0.00192	**0.0411** ± 0.00424	0.5415
500 bars	-0.0043 ± 0.00349	-0.002 ± 0.00181	-0.0013 ± 0.00145	0.002 ± 0.002	-0.0011 ± 0.00251	**0.0244** ± 0.0039	0.5232
1000 bars	-0.0059 ± 0.00319	-0.003 ± 0.00257	-0.001 ± 0.00167	**-0.0001** ± 0.00245	-0.0049 ± 0.0027	-0.0019 ± 0.00421	0.5066
1500 bars	-0.0051 ± 0.00382	-0.0049 ± 0.00263	-0.0007 ± 0.00182	-0.006 ± 0.0028	-0.0042 ± 0.00273	**0.0026** ± 0.00421	0.5051
2000 bars	-0.0087 ± 0.00331	-0.0042 ± 0.00282	-0.0031 ± 0.00224	-0.005 ± 0.00342	-0.0025 ± 0.00342	**0.0003** ± 0.00428	0.5074

Panel D: MDA feature importance for sequential correlation prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
25 bars	0.0048 ± 0.00112	0.0042 ± 0.00149	0.0158 ± 0.00209	**0.0548** ± 0.00733	0.0012 ± 0.00134	0.0053 ± 0.00141	0.5401
50 bars	0.005 ± 0.00135	0.0012 ± 0.00072	0.0096 ± 0.00147	**0.0433** ± 0.00716	0.0007 ± 0.0011	0.0057 ± 0.00177	0.5357
250 bars	0.0128 ± 0.00268	0.0002 ± 0.00136	0.0112 ± 0.00204	**0.0391** ± 0.00671	-0.0013 ± 0.00214	0.0021 ± 0.0024	0.5394
500 bars	0.0081 ± 0.00278	0.0017 ± 0.00186	0.0069 ± 0.00186	**0.023** ± 0.0046	0.0021 ± 0.00254	0.0045 ± 0.00289	0.5265
1000 bars	0.0031 ± 0.00287	-0.0015 ± 0.00195	0.0013 ± 0.00215	**0.0129** ± 0.00343	-0.0041 ± 0.00208	-0.0008 ± 0.00336	0.5173
1500 bars	-0.0109 ± 0.00825	-0.0032 ± 0.00204	0.0019 ± 0.0022	**0.0072** ± 0.00517	-0.0112 ± 0.00429	-0.0051 ± 0.00478	0.5113
2000 bars	0.0006 ± 0.00362	-0.0033 ± 0.00212	0.0028 ± 0.00204	**0.011** ± 0.00321	-0.006 ± 0.00245	0.0023 ± 0.00359	0.5113

Panel E: MDA feature importance for absolute skewness prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
25 bars	0.0013 ± 0.00074	**0.0374** ± 0.00494	0.0042 ± 0.00123	0.0106 ± 0.0015	-0.0009 ± 0.0009	0.0328 ± 0.00302	0.5447
50 bars	0.0012 ± 0.00087	**0.0288** ± 0.0045	0.002 ± 0.00072	0.0047 ± 0.00094	-0.0002 ± 0.00094	0.0276 ± 0.00246	0.537
250 bars	0.0028 ± 0.00264	0.0098 ± 0.00248	0.0001 ± 0.00117	0.0044 ± 0.0018	0.0002 ± 0.00213	**0.0179** ± 0.00315	0.5264
500 bars	-0.0005 ± 0.00293	0.0073 ± 0.00221	0.0006 ± 0.0013	0.0011 ± 0.00203	-0.0018 ± 0.00217	**0.0092** ± 0.00344	0.5166
1000 bars	-0.0033 ± 0.00303	**-0.0012** ± 0.00269	-0.0037 ± 0.00171	-0.0082 ± 0.00272	-0.0085 ± 0.00216	-0.008 ± 0.00293	0.5024
1500 bars	-0.0096 ± 0.00398	-0.0041 ± 0.00306	-0.0028 ± 0.00165	**-0.0021** ± 0.00251	-0.0039 ± 0.0024	-0.0084 ± 0.00421	0.4995
2000 bars	-0.0047 ± 0.00349	-0.0026 ± 0.00277	**-0.0019** ± 0.00166	-0.0052 ± 0.00286	-0.0027 ± 0.00255	-0.0025 ± 0.00406	0.504

Panel F: MDA feature importance for kurtosis prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
25 bars	0.0014 ± 0.00071	0.0062 ± 0.00121	0.005 ± 0.00075	0.0114 ± 0.00134	0.0001 ± 0.00061	**0.0968** ± 0.00597	0.5694
50 bars	0.0012 ± 0.00096	0.0047 ± 0.00137	0.0032 ± 0.00088	0.0094 ± 0.00104	-0.0007 ± 0.00094	**0.0844** ± 0.00482	0.5641
250 bars	0.0022 ± 0.0024	-0.0016 ± 0.00168	-0.0007 ± 0.0012	0.0077 ± 0.00198	-0.0023 ± 0.00184	**0.0461** ± 0.00421	0.5444
500 bars	-0.0039 ± 0.00315	-0.0035 ± 0.00185	-0.0018 ± 0.00156	0.0026 ± 0.00208	-0.0025 ± 0.00246	**0.0258** ± 0.00416	0.5227
1000 bars	-0.0065 ± 0.00308	-0.0025 ± 0.00265	**-0.0008** ± 0.00179	-0.0006 ± 0.00254	-0.0028 ± 0.00242	-0.002 ± 0.00441	0.5057
1500 bars	-0.0059 ± 0.00374	-0.0039 ± 0.0029	-0.0024 ± 0.00197	-0.0053 ± 0.00311	-0.0031 ± 0.0029	**0.0023** ± 0.00369	0.5046
2000 bars	-0.0089 ± 0.00317	-0.0043 ± 0.00292	-0.0026 ± 0.0022	-0.0068 ± 0.00334	-0.0022 ± 0.00349	**0.0009** ± 0.00411	0.5057

Table 4 MDA feature importance correlation between original random forest and adding max_depth=5

Variable	25 bars	250 bars	50 bars	500 bars	1000 bars	1500 bars	2000 bars
Kurtosis	0.998353	0.998687	0.997921	0.999611	0.999924	0.999904	0.999857
Bid-Ask spread	0.994484	0.99634	0.997186	0.993873	0.996883	0.997981	0.997974
Return variance	0.999724	0.99978	0.999759	0.999471	0.99923	0.999054	0.999315
Sequential correlation	0.999838	0.999787	0.999747	0.999869	0.999924	0.999784	0.999847
Skewness	0.999728	0.999669	0.999655	0.999896	0.99991	0.999963	0.999893
Jarque-Bera test	0.999695	0.999104	0.999393	0.999652	0.999931	0.999869	0.999815

Table 5 MDA feature importance correlation between original random forest and adding min_weight_fraction_leaf=0.01

Variable	25 bars	250 bars	50 bars	500 bars	1000 bars	1500 bars	2000 bars
Kurtosis	0.99845	0.999174	0.998438	0.99976	0.999984	0.999891	0.999861
Bid-Ask spread	0.996369	0.998045	0.998327	0.99646	0.998141	0.99838	0.998313
Return variance	0.999759	0.999895	0.999845	0.999754	0.999666	0.999218	0.999587
Sequential correlation	0.999918	0.999892	0.999823	0.999928	0.999867	0.999945	0.999906
Skewness	0.999677	0.999788	0.999632	0.999895	0.99996	0.999955	0.999968
Jarque-Bera test	0.999644	0.999318	0.999532	0.999794	0.999982	0.999899	0.999812

Table 6 Yearly feature importance, lookback window is fixed at 250 bars.

Panel A: MDA feature importance for bid-ask spread prediction.

Period	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
2012-2013	0.0013	0.0002	-0.0051	0.0014	0.0033	**0.0142**	0.4555
2013-2014	0.0034	-0.0012	-0.0067	-0.0008	0.0020	**0.0174**	0.4578
2014-2015	0.0035	-0.0003	-0.0040	0.0020	0.0009	**0.0142**	0.4480
2015-2016	0.0023	-0.0013	-0.0094	-0.0044	0.0038	**0.0128**	0.4549
2016-2017	0.0006	0.0008	-0.0038	-0.0005	0.0032	**0.0083**	0.4516

Panel B: MDA feature importance for realized volatility prediction.

Period	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
2012-2013	0.0079	0.0001	0.0034	**0.0170**	-0.0015	0.0128	0.5479
2013-2014	0.0117	-0.0027	0.0048	0.0141	-0.0046	**0.0198**	0.5507
2014-2015	0.0001	-0.0003	0.0105	**0.0215**	-0.0139	-0.0003	0.5337
2015-2016	0.0098	0.0012	0.0088	**0.0174**	-0.0092	0.0058	0.5431
2016-2017	-0.0032	0.0023	0.0040	**0.0143**	-0.0105	0.0171	0.5526

Panel C: MDA feature importance for Jarque-Bera test prediction.

Period	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
2012-2013	-0.0056	-0.0024	-0.0045	-0.0010	-0.0042	**0.0186**	0.5333
2013-2014	0.0040	0.0018	0.0025	0.0080	0.0038	**0.0372**	0.5556
2014-2015	-0.0130	-0.0031	0.0000	0.0033	-0.0111	**0.0131**	0.5319
2015-2016	-0.0049	-0.0039	-0.0045	-0.0008	-0.0162	**0.0269**	0.5382
2016-2017	-0.0155	0.0032	-0.0027	0.0006	-0.0125	**0.0057**	0.5326

Panel D: MDA feature importance for sequential correlation prediction.

Period	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
2012-2013	-0.0039	-0.0012	0.0057	**0.0313**	-0.0114	-0.0047	0.5435
2013-2014	-0.0069	-0.0018	0.0005	**0.0245**	-0.0091	-0.0056	0.5370
2014-2015	-0.0023	-0.0035	-0.0086	**0.0092**	-0.0077	-0.0158	0.5332
2015-2016	0.0003	-0.0033	0.0054	**0.0200**	-0.0067	-0.0144	0.5332
2016-2017	-0.0033	-0.0055	-0.0054	**0.0194**	-0.0112	-0.0144	0.5322

Panel E: MDA feature importance for absolute skewness prediction.

Period	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
2012-2013	-0.0015	-0.0035	-0.0030	-0.0021	0.0013	**0.0033**	0.5215

2013-2014	0.0052	0.0033	-0.0012	0.0008	-0.0043	**0.0090**	0.5309
2014-2015	-0.0049	**0.0033**	0.0013	0.0000	-0.0122	-0.0041	0.5137
2015-2016	-0.0028	0.0033	0.0000	-0.0017	-0.0066	**0.0105**	0.5208
2016-2017	-0.0083	**0.0027**	-0.0028	-0.0075	-0.0100	-0.0081	0.5165

Panel F: MDA feature importance for kurtosis prediction.

Period	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
2012-2013	-0.0032	-0.0012	-0.0023	0.0004	-0.0019	**0.0215**	0.5355
2013-2014	0.0028	-0.0022	-0.0016	0.0031	0.0030	**0.0358**	0.5568
2014-2015	-0.0066	-0.0007	0.0021	0.0055	-0.0070	**0.0269**	0.5378
2015-2016	-0.0062	-0.0012	0.0001	0.0027	-0.0138	**0.0325**	0.5423
2016-2017	-0.0144	0.0004	-0.0010	0.0005	-0.0177	**0.0070**	0.5343

Table 7 Performance comparison between Dollar-Volume bars and Time bars.

Variable	Average accuracy	
	DV bar	Time bar
Bid-ask spread	0.4539	**0.4699**
Jarque-Bera test	0.5237	**0.5269**
Kurtosis	**0.5309**	0.5304
Return variance	**0.5663**	0.5609
Sequential correlation	**0.5259**	0.5252
Skewness	**0.5187**	0.5142

Table 8 MDA feature importance correlation between logistic regression and random forest

Variable	25 bars	50 bars	250 bars	500 bars	1000 bars	1500 bars	2000 bars
Kurtosis	0.9818	0.9779	0.9602	0.9489	-0.1058	0.5929	0.6807
Bid-Ask spread	0.6478	0.5756	0.0999	0.1582	-0.0639	0.0177	0.1712
Return variance	0.9312	0.9223	0.8830	0.8264	0.9366	-0.3209	0.9492
Sequential correlation	0.9808	0.9929	0.9719	0.9268	0.9561	0.6887	0.9506
Skewness	0.9983	0.9809	0.8963	0.8058	0.6231	0.2737	0.7509
Jarque-Bera test	0.8611	0.9375	0.9589	0.9377	-0.0275	0.5359	0.7963

Table 9 MDA feature importance correlation between 50 bars forward window and 250 bars forward window

Variable	25 bars	50 bars	250 bars	500 bars	1000 bars	1500 bars	2000 bars
Kurtosis	0.9984	0.9987	0.9685	0.5405	0.3060	0.8690	0.5939
Bid-Ask spread	0.9951	0.9828	0.8229	0.9810	0.9612	0.8946	0.9459
Return variance	0.9911	0.9584	0.4131	0.9197	0.9801	0.9956	0.9864
Sequential correlation	0.9973	0.9979	0.8768	0.9415	0.9032	0.8292	0.9677
Skewness	0.9984	0.9947	0.6333	0.9140	-0.0871	0.3896	0.1941
Jarque-Bera test	0.9946	0.9945	0.8864	0.6871	0.2597	0.8482	0.7291

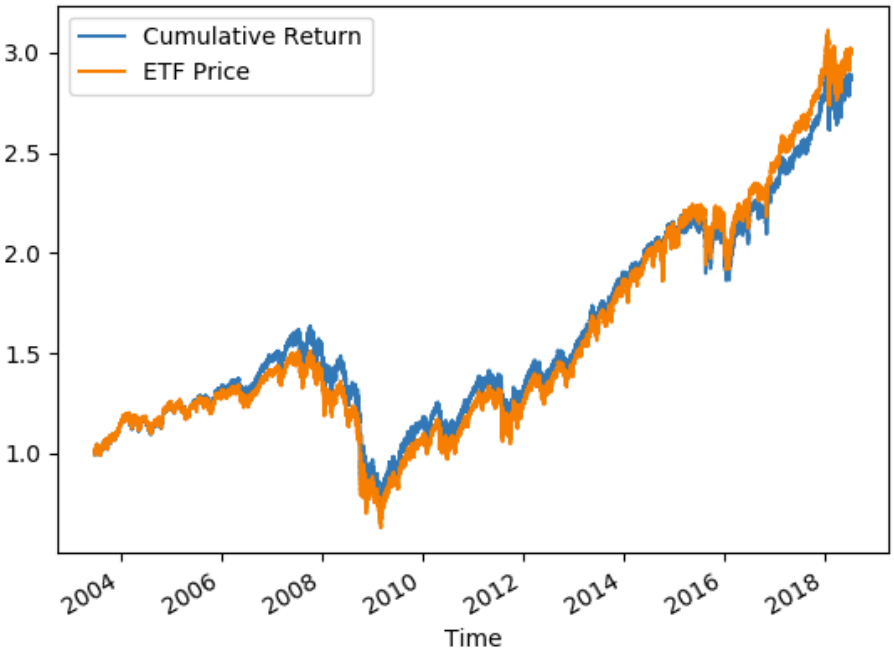


Figure 1: ES1 Index's cumulative return and ETF price.

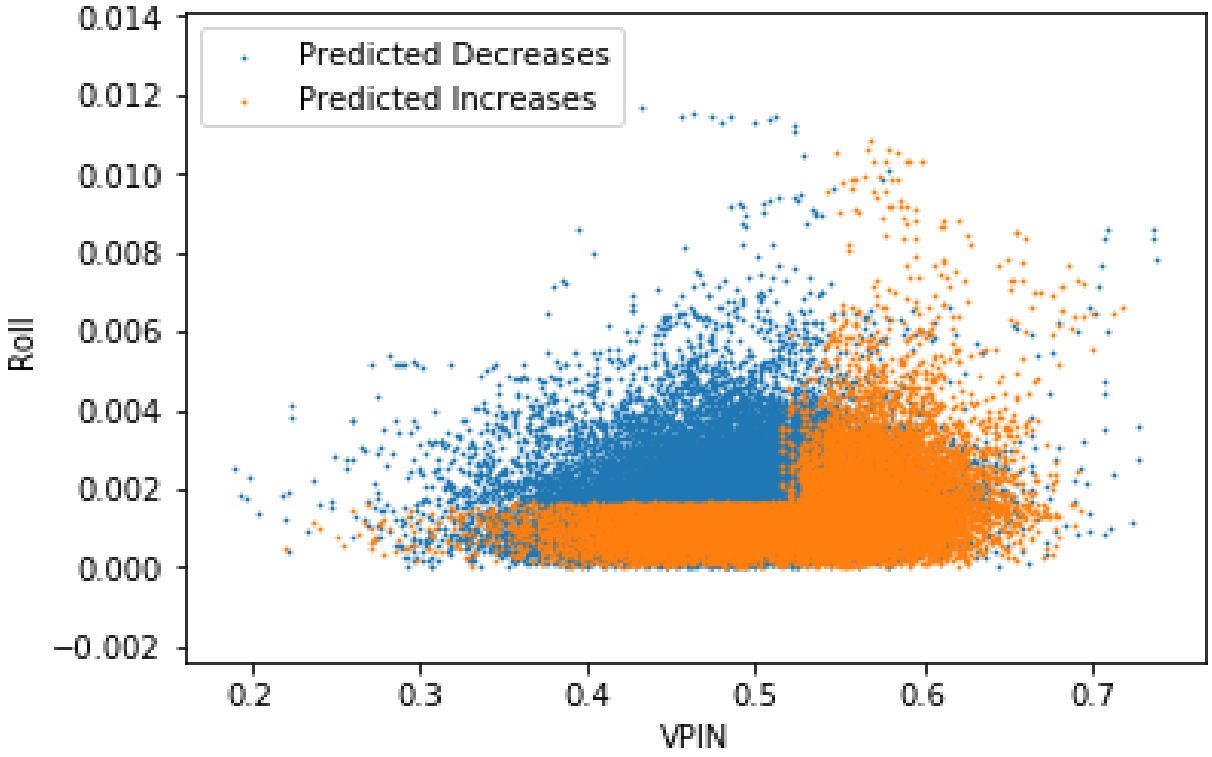


Figure 2: Scatter plot of predicted changes in spread as a function of the VPIN and Roll measures. The plot is for the ES1 Index with a lookback window of 25 bars.

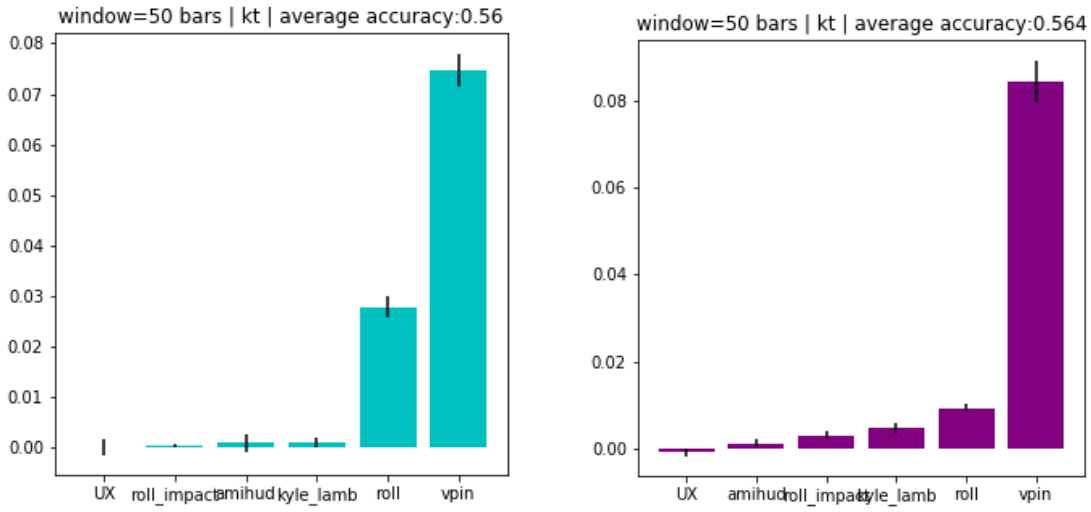


Figure 3: MDA feature importance for kurtosis prediction with window size = 50 bars. Left: time-bar. Right: dollar-volume-bar.

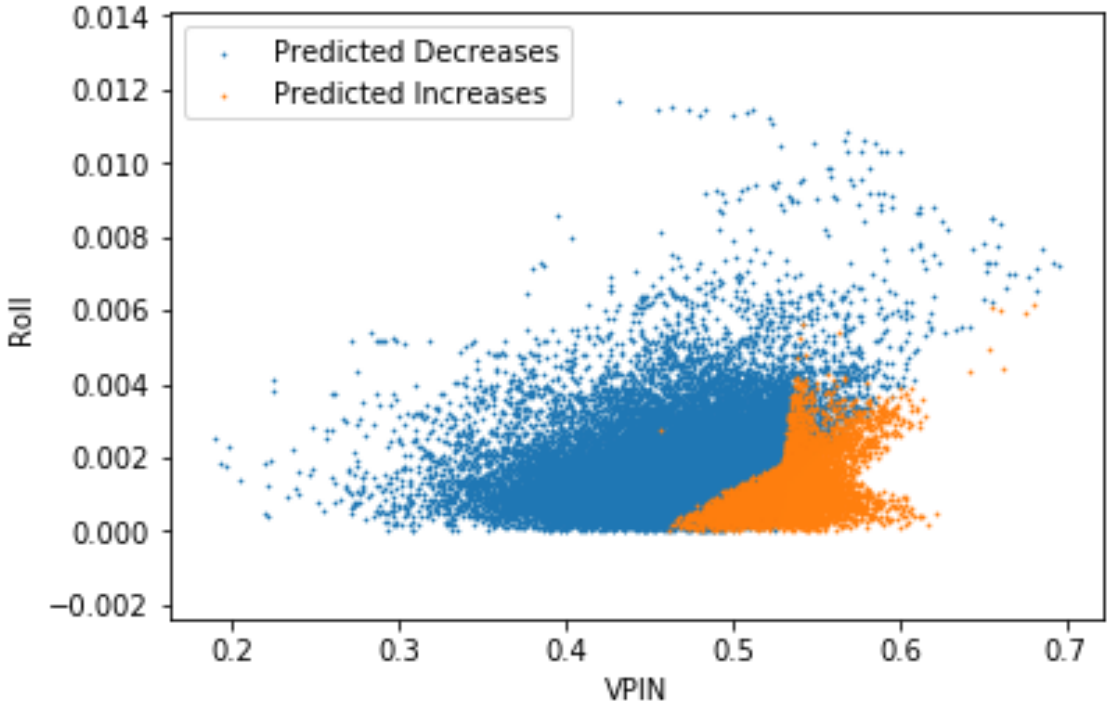


Figure 4: Scatter plot of predicted changes in spread as a function of the VPIN and Roll measures for the logistic regression. The plot is for the ES1 Index with a lookback window of 25 bars.

Appendix A.1 - Data

Table A.1 Data summary of 87 futures contracts.

This table lists and data and sources used in this paper. The data are futures contracts and each contract is identified by its ticker simple and general asset class. The data is divided into dollar volume bars. The sample period is from July 2, 2012 – October 2, 2017.

Instrument	Data Source	Size	Sample Period
AD1_Curncy	TickWrite	72553 Bars	2012-07-02 - 2017-10-02
BO1_Comdty	TickWrite	72306 Bars	2012-07-02 - 2017-10-02
BP1_Curncy	TickWrite	67939 Bars	2012-07-02 - 2017-10-02
BTS1_Comdty	TickWrite	30756 Bars	2015-10-01 - 2017-10-02
BZ1_Index	TickWrite	68054 Bars	2012-07-02 - 2017-10-02
CC1_Comdty	TickWrite	55568 Bars	2012-07-02 - 2017-10-02
CD1_Curncy	TickWrite	69808 Bars	2012-07-02 - 2017-10-02
CF1_Index	TickWrite	71056 Bars	2012-07-02 - 2017-10-02
CL1_Comdty	TickWrite	63875 Bars	2012-07-02 - 2017-10-02
CN1_Comdty	TickWrite	52872 Bars	2012-07-03 - 2017-10-02
CO1_Comdty	TickWrite	87235 Bars	2012-07-02 - 2017-10-02
CT1_Comdty	TickWrite	65074 Bars	2012-07-02 - 2017-10-02
C_1_Comdty	TickWrite	75564 Bars	2012-07-02 - 2017-10-02
DM1_Index	TickWrite	57993 Bars	2012-07-02 - 2017-10-02
DU1_Comdty	TickWrite	79664 Bars	2012-07-02 - 2017-10-02
DX1_Curncy	TickWrite	72221 Bars	2012-07-02 - 2017-10-02
EC1_Comdty	TickWrite	73491 Bars	2012-07-02 - 2017-10-02
EC1_Curncy	TickWrite	85801 Bars	2012-07-02 - 2017-10-02
ED1_Comdty	TickWrite	48863 Bars	2012-07-02 - 2017-10-02
EE1_Curncy	TickWrite	83571 Bars	2012-07-02 - 2017-10-02

EO1_Comdty	TickWrite	89594 Bars	2012-07-02 - 2017-10-02
EO1_Index	TickWrite	60815 Bars	2012-07-02 - 2017-10-02
ER1_Comdty	TickWrite	98910 Bars	2012-07-02 - 2017-06-28
ES1_Index	TickWrite	58347 Bars	2012-07-02 - 2017-10-02
FA1_Index	TickWrite	64184 Bars	2012-07-02 - 2017-10-02
FC1_Comdty	TickWrite	71374 Bars	2012-07-02 - 2017-10-02
FV1_Comdty	TickWrite	66428 Bars	2012-07-02 - 2017-10-02
GC1_Comdty	TickWrite	61138 Bars	2012-07-02 - 2017-10-02
GX1_Index	TickWrite	65780 Bars	2012-07-02 - 2017-10-02
G_1_Comdty	TickWrite	58371 Bars	2012-07-02 - 2017-10-02
HG1_Comdty	TickWrite	72404 Bars	2012-07-02 - 2017-10-02
HI1_Index	TickWrite	47631 Bars	2012-07-02 - 2017-09-29
HO1_Comdty	TickWrite	106393 Bars	2012-07-02 - 2017-10-02
IB1_Index	TickWrite	69305 Bars	2012-07-02 - 2017-10-02
IK1_Comdty	TickWrite	24956 Bars	2015-10-01 - 2017-10-02
IR1_Comdty	TickWrite	95442 Bars	2012-07-02 - 2017-10-02
JA1_Comdty	TickWrite	110637 Bars	2012-07-02 - 2017-10-02
JB1_Comdty	TickWrite	73579 Bars	2012-07-02 - 2017-10-02
JE1_Curncy	TickWrite	47231 Bars	2012-07-02 - 2017-10-02
JG1_Comdty	TickWrite	74719 Bars	2012-07-02 - 2017-10-02
JO1_Comdty	TickWrite	60186 Bars	2012-07-02 - 2017-10-02
JY1_Curncy	TickWrite	69056 Bars	2012-07-02 - 2017-10-02
KC1_Comdty	TickWrite	57345 Bars	2012-07-02 - 2017-10-02
LB1_Comdty	TickWrite	81342 Bars	2012-07-02 - 2017-10-02
LC1_Comdty	TickWrite	77021 Bars	2012-07-02 - 2017-10-02
LH1_Comdty	TickWrite	91351 Bars	2012-07-02 - 2017-10-02
L_1_Comdty	TickWrite	87387 Bars	2012-07-02 - 2017-06-28
MFS1_Index	TickWrite	49722 Bars	2012-07-02 - 2017-10-02

NG1_Comdty	TickWrite	79717 Bars	2012-07-02 - 2017-10-02
NI1_Index	TickWrite	73209 Bars	2012-07-02 - 2017-10-02
NK1_Index	TickWrite	63739 Bars	2012-07-02 - 2017-10-02
NQ1_Index	TickWrite	62864 Bars	2012-07-02 - 2017-10-02
NX1_Index	TickWrite	63741 Bars	2012-07-02 - 2017-10-02
OAT1_Comdty	TickWrite	25815 Bars	2015-10-01 - 2017-10-02
OE1_Comdty	TickWrite	62944 Bars	2012-07-02 - 2017-10-02
O_1_Comdty	TickWrite	111309 Bars	2012-07-02 - 2017-10-02
PA1_Comdty	TickWrite	74167 Bars	2012-07-02 - 2017-10-02
PE1_Curncy	TickWrite	62434 Bars	2012-07-02 - 2017-10-02
PT1_Index	TickWrite	57267 Bars	2012-07-03 - 2017-10-02
QS1_Comdty	TickWrite	104888 Bars	2012-07-02 - 2017-10-02
RR1_Comdty	TickWrite	82241 Bars	2012-07-02 - 2017-10-02
RTA1_Index	TickWrite	73603 Bars	2012-07-02 - 2017-10-02
RX1_Comdty	TickWrite	63887 Bars	2012-07-02 - 2017-10-02
SB1_Comdty	TickWrite	58299 Bars	2012-07-02 - 2017-10-02
SF1_Curncy	TickWrite	88162 Bars	2012-07-02 - 2017-10-02
SI1_Comdty	TickWrite	65028 Bars	2012-07-02 - 2017-10-02
SM1_Comdty	TickWrite	69252 Bars	2012-07-02 - 2017-10-02
SM1_Index	TickWrite	58143 Bars	2012-07-02 - 2017-10-02
SP1_Index	TickWrite	121578 Bars	2012-07-02 - 2017-10-02
ST1_Index	TickWrite	58630 Bars	2012-07-02 - 2017-10-02
S_1_Comdty	TickWrite	68785 Bars	2012-07-02 - 2017-10-02
TP1_Index	TickWrite	60620 Bars	2012-07-02 - 2017-10-02
TU1_Comdty	TickWrite	58332 Bars	2012-07-02 - 2017-10-02
TW1_Index	TickWrite	67037 Bars	2012-07-02 - 2017-10-02
TY1_Comdty	TickWrite	64273 Bars	2012-07-02 - 2017-10-02
UB1_Comdty	TickWrite	42292 Bars	2012-07-02 - 2017-10-02

US1_Comdty	TickWrite	71814 Bars	2012-07-02 - 2017-10-02
VG1_Index	TickWrite	61495 Bars	2012-07-02 - 2017-10-02
VH1_Index	TickWrite	80617 Bars	2012-07-02 - 2017-10-02
W_1_Comdty	TickWrite	77511 Bars	2012-07-02 - 2017-10-02
XB1_Comdty	TickWrite	98004 Bars	2012-07-02 - 2017-10-02
XG1_Comdty	TickWrite	134784 Bars	2012-07-02 - 2017-10-02
XM1_Comdty	TickWrite	52167 Bars	2012-07-02 - 2017-10-02
XP1_Index	TickWrite	63842 Bars	2012-07-02 - 2017-10-02
YM1_Comdty	TickWrite	74597 Bars	2012-07-02 - 2017-10-02
YS1_Comdty	TickWrite	172029 Bars	2012-07-02 - 2017-10-02
Z_1_Index	TickWrite	60405 Bars	2012-07-02 - 2017-10-02

Table A.2 A snippet of ES1 Index data

This table shows a selection of data for ES1 which is the ticker symbol for the CME E-mini S&P 500 front-month continuous contract.

Time	Open	Symbol	Close	Ticks	Volume	High	Low	ETF Price
7/1/2003 9:45	971.75	ESU03 Index	968.25	7691	69823	975	966.25	1
7/1/2003 10:05	968.25	ESU03 Index	962	5946	70143	969.75	962	0.993545
7/1/2003 10:19	962	ESU03 Index	963	5558	70304	964.25	960.25	0.994578
7/1/2003 10:39	963	ESU03 Index	963.5	4936	70188	965.75	962.5	0.995094
7/1/2003 11:17	963.5	ESU03 Index	964.5	5766	70171	965.25	961.75	0.996127
7/1/2003 12:34	964.5	ESU03 Index	964	7571	70050	967	963	0.995611
7/1/2003 13:27	964	ESU03 Index	970.75	6956	70024	970.75	962.75	1.002582
7/1/2003 14:20	970.75	ESU03 Index	972.25	8294	69783	972.25	968.5	1.004131
7/1/2003 14:42	972.25	ESU03 Index	976.75	6282	69424	977.25	971.75	1.008779

Table A.3 MDA feature importance for kurtosis prediction per instrument, lookback window = 250 bars

Instrument	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
GC1_Comdty	0.0246	0.0000	0.0044	0.0341	-0.0158	0.0124	0.5386
FC1_Comdty	-0.0428	0.0000	-0.0299	-0.0375	-0.0217	0.0834	0.5352
HG1_Comdty	0.0394	0.0000	0.0127	0.0344	0.0077	0.0442	0.5650
JE1_Curncy	0.0463	0.0000	-0.0078	0.0319	0.0072	0.0454	0.5678
EC1_Comdty	-0.0091	0.0000	0.0014	-0.0102	-0.0066	0.0424	0.5367
EC1_Curncy	-0.0093	-0.0001	0.0044	0.0206	-0.0155	-0.0091	0.5063
EE1_Curncy	0.0148	0.0000	-0.0042	-0.0196	-0.0138	-0.0090	0.5235
JA1_Comdty	0.0208	0.0000	-0.0086	-0.0021	0.0131	0.1045	0.5676
Z_1_Index	0.0133	0.0001	0.0201	0.0185	-0.0081	0.0365	0.5391
PA1_Comdty	-0.0564	0.0000	0.0025	0.0117	-0.0081	0.0100	0.5195
C_1_Comdty	0.0076	0.0000	-0.0004	0.0536	-0.0137	-0.0072	0.5127
NQ1_Index	-0.0209	0.0000	-0.0034	-0.0051	0.0015	0.0306	0.5324
JG1_Comdty	0.0373	0.0003	-0.0022	-0.0016	0.0229	0.0680	0.5553
QS1_Comdty	-0.0072	0.0000	0.0045	-0.0019	-0.0151	0.0461	0.5361
BO1_Comdty	0.0077	0.0000	0.0039	0.0284	0.0202	0.0553	0.5381
PT1_Index	-0.0171	0.0000	-0.0033	0.0029	0.0088	0.0436	0.5465
SF1_Curncy	-0.0166	0.0000	0.0002	0.0221	-0.0060	0.0407	0.5329
SB1_Comdty	-0.0015	-0.0073	0.0032	0.0022	-0.0336	0.0722	0.5372
NX1_Index	-0.0167	0.0000	0.0015	0.0079	0.0028	0.0089	0.5263
YS1_Comdty	-0.0024	0.0000	0.0082	0.0456	0.0384	0.0213	0.5397
S_1_Comdty	0.0004	0.0001	0.0053	0.0206	-0.0110	0.0575	0.5343
HO1_Comdty	0.0236	0.0000	0.0131	0.0146	0.0101	0.1053	0.5895
PE1_Curncy	0.0319	0.0000	-0.0051	0.0032	0.0012	0.0614	0.5470
IK1_Comdty	0.0075	-0.0001	-0.0125	0.0045	-0.0175	-0.0237	0.5155
LC1_Comdty	-0.0025	0.0000	-0.0106	-0.0201	-0.0114	0.0557	0.5329
ED1_Comdty	0.0048	-0.0031	-0.0089	0.0078	0.0023	0.0151	0.5309
LH1_Comdty	-0.0260	0.0000	-0.0182	-0.0162	-0.0296	0.0237	0.5168
FV1_Comdty	0.0253	0.0000	-0.0072	-0.0054	-0.0017	0.0450	0.5414
OE1_Comdty	0.0199	0.0108	-0.0008	0.0228	0.0075	0.0519	0.5600
SM1_Index	-0.0425	0.0000	-0.0269	-0.0244	0.0003	0.0223	0.5158
SI1_Comdty	0.0052	0.0000	-0.0008	0.0452	0.0044	0.0056	0.5378
JO1_Comdty	-0.0250	0.0000	-0.0157	-0.0330	-0.0022	-0.0102	0.4875
KC1_Comdty	0.0041	0.0000	0.0028	0.0068	0.0024	0.0579	0.5245
AD1_Curncy	0.0171	0.0000	0.0074	0.0166	0.0100	0.0532	0.5586
DX1_Curncy	0.0216	0.0000	0.0066	0.0289	-0.0044	0.0138	0.5332

DM1_Index	-0.0006	0.0000	-0.0277	-0.0186	0.0168	-0.0034	0.5166
RR1_Comdty	0.0153	0.0000	0.0108	0.0158	0.0121	0.0910	0.5596
ST1_Index	-0.0126	0.0000	0.0039	-0.0068	-0.0118	0.0614	0.5403
NG1_Comdty	-0.0154	0.0039	-0.0042	-0.0056	-0.0442	0.0353	0.5237
US1_Comdty	0.0056	-0.0059	-0.0028	-0.0021	0.0087	0.0023	0.5229
EO1_Index	-0.0200	0.0000	-0.0122	-0.0138	-0.0080	0.0220	0.5130
XB1_Comdty	-0.0010	0.0000	0.0006	-0.0048	0.0117	0.1128	0.5797
ER1_Comdty	0.0148	0.0000	-0.0032	0.0290	0.0041	0.1174	0.5955
GX1_Index	-0.0084	0.0000	-0.0152	-0.0006	-0.0192	0.0656	0.5395
RTA1_Index	-0.0151	0.0001	0.0073	0.0087	0.0185	0.0662	0.5583
BZ1_Index	0.0215	0.0000	-0.0160	-0.0111	-0.0026	0.0605	0.5517
L_1_Comdty	0.0086	0.0000	0.0065	0.0141	-0.0003	0.1527	0.6042
SM1_Comdty	0.0189	0.0031	0.0007	-0.0091	0.0043	0.0088	0.5233
VH1_Index	-0.0077	0.0000	0.0011	0.0204	-0.0102	0.0889	0.5591
RX1_Comdty	0.0555	0.0014	-0.0117	0.0183	-0.0175	0.0506	0.5684
IR1_Comdty	0.0169	0.0000	-0.0072	0.0103	0.0280	0.1125	0.5930
G_1_Comdty	-0.0083	0.0026	0.0122	0.0166	-0.0221	0.0319	0.5365
TP1_Index	-0.0252	-0.0088	0.0064	0.0522	-0.0139	0.0305	0.5545
FA1_Index	0.0326	0.0000	0.0008	0.0132	0.0291	0.0874	0.5762
W_1_Comdty	0.0162	-0.0226	0.0059	0.0106	0.0015	0.0164	0.5213
XP1_Index	0.0280	0.0000	0.0091	0.0065	0.0310	0.1105	0.5787
JY1_Curncy	0.0141	0.0001	0.0030	0.0130	0.0053	0.0185	0.5441
BP1_Curncy	0.0092	0.0000	0.0266	0.0203	-0.0187	0.0463	0.5395
VG1_Index	-0.0038	-0.0128	0.0051	0.0022	-0.0114	0.0605	0.5509
XM1_Comdty	0.0104	0.0001	-0.0173	-0.0350	-0.0367	0.0093	0.5008
IB1_Index	-0.0045	-0.0207	0.0127	-0.0006	-0.0282	0.0113	0.5215
XG1_Comdty	0.0081	0.0000	0.0008	0.0152	0.0103	0.0275	0.5371
O_1_Comdty	-0.0003	0.0000	-0.0099	-0.0047	0.0144	0.1078	0.5725
NK1_Index	0.0529	0.0000	0.0128	0.0268	0.0071	0.0839	0.5791
SP1_Index	0.0413	0.0000	-0.0174	-0.0184	-0.0378	0.0319	0.5281
TY1_Comdty	0.0004	-0.0022	0.0030	0.0233	-0.0168	-0.0076	0.5253
HI1_Index	-0.0082	-0.0108	-0.0144	-0.0159	0.0039	0.0343	0.5346
JB1_Comdty	0.0033	0.0000	-0.0164	-0.0061	-0.0104	0.0526	0.5373
TU1_Comdty	0.0288	0.0688	-0.0015	0.0094	-0.0014	0.0326	0.5758
OAT1_Comdty	-0.0456	0.0055	0.0560	0.0529	-0.0298	0.0374	0.5241
CN1_Comdty	0.0021	0.0001	0.0034	0.0235	0.0359	0.0599	0.5628
CD1_Curncy	0.0299	0.0000	-0.0020	0.0097	0.0200	0.0387	0.5639
DU1_Comdty	-0.0175	0.0003	-0.0037	0.0007	-0.0024	0.0108	0.5159
MFS1_Index	-0.0091	-0.0027	-0.0188	-0.0080	-0.0137	0.0289	0.5215
CF1_Index	0.0168	0.0000	0.0140	0.0129	0.0083	0.0713	0.5384

BTS1_Comdty	-0.0589	0.0000	-0.0108	0.0298	-0.0024	-0.0275	0.5188
EO1_Comdty	-0.0239	0.0000	0.0032	0.0105	-0.0156	0.0396	0.5390
CC1_Comdty	-0.0023	0.0000	-0.0017	0.0102	0.0121	0.0396	0.5312
YM1_Comdty	-0.0057	0.0153	0.0102	0.0290	-0.0083	-0.0041	0.5367
ES1_Index	-0.0234	-0.0136	-0.0098	-0.0039	0.0115	0.0278	0.5291
LB1_Comdty	0.0010	0.0000	-0.0113	0.0058	0.0067	0.1094	0.5727
CT1_Comdty	0.0208	0.0000	-0.0078	-0.0084	0.0043	0.0468	0.5419
UB1_Comdty	0.0260	0.0000	-0.0051	-0.0024	0.0179	0.0938	0.5726
CL1_Comdty	0.0212	0.0045	0.0353	0.0129	-0.0079	0.0500	0.5864
CO1_Comdty	0.0056	0.0000	0.0113	-0.0010	-0.0028	0.0155	0.5283

Table A.4 MDA feature importance for logistic regression

Panel A: MDA feature importance for bid-ask spread prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
25 bars	0.001 ± 0.0004	0.0033 ± 0.0024	**0.0467** ± 0.0056	0.113 ± 0.0053	0.0021 ± 0.0011	0.0404 ± 0.0023	0.4737
50 bars	0.0019 ± 0.0005	-0.0 ± 0.0	**0.0467** ± 0.0065	0.1044 ± 0.006	0.0005 ± 0.0002	0.0308 ± 0.0018	0.4705
250 bars	0.0056 ± 0.0013	-0.0 ± 0.0	0.0339 ± 0.0035	0.0703 ± 0.0035	0.006 ± 0.0013	**0.0342** ± 0.0023	0.4754
500 bars	0.004 ± 0.0009	-0.0 ± 0.0	**0.0602** ± 0.0063	0.0677 ± 0.0062	0.0083 ± 0.0017	0.0561 ± 0.0026	0.4792
1000 bars	0.0082 ± 0.0018	0.0 ± 0.0	**0.0625** ± 0.0067	0.0604 ± 0.0068	0.0139 ± 0.0028	0.0462 ± 0.002	0.4662
1500 bars	0.0155 ± 0.0026	-0.0001 ± 0.0001	**0.0528** ± 0.0065	0.049 ± 0.0067	0.0189 ± 0.0038	0.0394 ± 0.0021	0.4551
2000 bars	0.0187 ± 0.0028	-0.0 ± 0.0	**0.0412** ± 0.0059	0.0369 ± 0.0063	0.0207 ± 0.004	0.0349 ± 0.0025	0.4465

Panel B: MDA feature importance for realized volatility prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
25 bars	0.003 ± 0.0006	0.0006 ± 0.0006	0.0167 ± 0.0042	**0.1206** ± 0.0071	0.0 ± 0.0003	0.0611 ± 0.0064	0.6269
50 bars	0.0055 ± 0.0011	0.0 ± 0.0	0.0213 ± 0.0049	**0.108** ± 0.0063	0.0004 ± 0.0004	0.0486 ± 0.0057	0.6083

250 bars	0.0279 ± 0.0038	0.0 ± 0.0	0.0075 ± 0.004	**0.0937** ± 0.0065	-0.0012 ± 0.0012	0.016 ± 0.0038	0.5759
500 bars	0.03 ± 0.0042	-0.0 ± 0.0	0.005 ± 0.0023	0.0445 ± 0.0054	0.0064 ± 0.0023	**0.061** ± 0.0052	0.5746
1000 bars	0.0165 ± 0.0052	-0.0001 ± 0.0	0.0091 ± 0.0032	0.0222 ± 0.0048	0.0109 ± 0.0031	**0.0867** ± 0.0079	0.5922
1500 bars	-0.2528 ± 0.2774	-0.0001 ± 0.0001	0.0109 ± 0.0031	**0.0165** ± 0.0045	-0.044 ± 0.0662	-0.1378 ± 0.2459	0.6102
2000 bars	0.0179 ± 0.0034	-0.0 ± 0.0	0.0103 ± 0.0032	0.0135 ± 0.004	0.0298 ± 0.0052	**0.0984** ± 0.0065	0.6221

Panel C: MDA feature importance for Jarque-Bera test prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
25 bars	0.0004 ± 0.0004	0.0003 ± 0.0003	0.0103 ± 0.0042	0.0366 ± 0.0048	0.0007 ± 0.0004	**0.0441** ± 0.0081	0.5388
50 bars	0.0007 ± 0.0006	-0.0 ± 0.0	0.0091 ± 0.0019	0.038 ± 0.0035	0.0004 ± 0.0005	**0.0683** ± 0.0072	0.5604
250 bars	0.0159 ± 0.0024	-0.0 ± 0.0	0.007 ± 0.0015	0.043 ± 0.0053	0.002 ± 0.001	**0.0789** ± 0.0057	0.5817
500 bars	0.0115 ± 0.0031	0.0001 ± 0.0001	0.0065 ± 0.0031	0.0228 ± 0.0039	0.0 ± 0.0011	**0.0534** ± 0.006	0.5516
1000 bars	0.0008 ± 0.003	-0.0 ± 0.0	-0.0053 ± 0.0026	**0.0026** ± 0.0041	0.0004 ± 0.0026	0.0015 ± 0.0039	0.5119
1500 bars	-0.0079 ± 0.0035	-0.0001 ± 0.0	-0.003 ± 0.002	-0.016 ± 0.0053	-0.0039 ± 0.0036	**0.0019** ± 0.0042	0.5109
2000 bars	-0.0096 ± 0.0032	0.0 ± 0.0	-0.0082 ± 0.0043	-0.0128 ± 0.0044	-0.0005 ± 0.0034	**0.0117** ± 0.0047	0.5195

Panel D: MDA feature importance for sequential correlation prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
25 bars	0.0042 ± 0.0009	-0.0002 ± 0.0002	0.0127 ± 0.0032	**0.108** ± 0.0083	0.0038 ± 0.0022	0.0091 ± 0.0022	0.5683
50 bars	0.0066 ± 0.0014	0.0 ± 0.0	0.0069 ± 0.0017	**0.0921** ± 0.0087	0.001 ± 0.0005	0.01 ± 0.0023	0.5586
250 bars	0.0285 ± 0.0034	0.0 ± 0.0	0.0019 ± 0.0024	**0.0688** ± 0.0106	-0.0006 ± 0.0012	0.0044 ± 0.0027	0.5641
500 bars	0.0206 ± 0.0037	0.0 ± 0.0	-0.0005 ± 0.0022	**0.0404** ± 0.0077	-0.0035 ± 0.0014	0.0078 ± 0.0033	0.5456
1000 bars	0.0107 ± 0.0027	-0.0 ± 0.0	-0.0004 ± 0.0026	**0.0281** ± 0.0067	-0.005 ± 0.0025	0.008 ± 0.0043	0.5333
1500 bars	0.0048 ± 0.0026	-0.0001 ± 0.0001	-0.0012 ± 0.0028	**0.0178** ± 0.0048	-0.0063 ± 0.0022	0.011 ± 0.0043	0.5267
2000 bars	0.0034 ± 0.0032	0.0 ± 0.0	0.0004 ± 0.0022	**0.0162** ± 0.0056	-0.0085 ± 0.0029	0.0015 ± 0.0046	0.5178

Panel E: MDA feature importance for absolute skewness prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
25 bars	0.0022 ± 0.0007	-0.0005 ± 0.0005	0.0097 ± 0.0017	0.0283 ± 0.0038	0.0 ± 0.0009	**0.0604** ± 0.0031	0.5491
50 bars	0.0028 ± 0.0011	0.0 ± 0.0	0.0122 ± 0.0027	0.0289 ± 0.0029	0.0002 ± 0.0004	**0.0608** ± 0.0036	0.5496
250 bars	0.014 ± 0.0029	-0.0 ± 0.0	0.0042 ± 0.0017	0.0261 ± 0.004	0.0015 ± 0.0013	**0.0432** ± 0.0044	0.5513
500 bars	0.0043 ± 0.0023	0.0 ± 0.0	0.0012 ± 0.0023	0.0105 ± 0.0034	-0.0016 ± 0.0017	**0.0255** ± 0.0057	0.531
1000 bars	-0.0007 ± 0.0027	-**0.0** ± 0.0	-0.0045 ± 0.0027	-0.0028 ± 0.0031	-0.002 ± 0.0022	-0.0034 ± 0.0031	0.5059
1500 bars	-0.008 ± 0.0026	-**0.0001** ± 0.0001	-0.0056 ± 0.0023	-0.0097 ± 0.004	-0.0049 ± 0.0031	-0.0034 ± 0.0033	0.5047
2000 bars	-0.0052 ± 0.0022	**0.0001** ± 0.0001	-0.0029 ± 0.0026	-0.0026 ± 0.0034	-0.0019 ± 0.0031	-0.0013 ± 0.0033	0.509

Panel F: MDA feature importance for kurtosis prediction.

Window size	Amihud	Kyle Lambda	Roll Impact	Roll Measure	VIX	VPIN	Accuracy
25 bars	0.0024 ± 0.0004	-0.0009 ± 0.0009	0.0202 ± 0.0051	0.0437 ± 0.0055	0.0015 ± 0.0004	**0.1431** ± 0.0055	0.6027
50 bars	0.0005 ± 0.0004	0.0 ± 0.0	0.0101 ± 0.0022	0.0429 ± 0.0037	0.0011 ± 0.0006	**0.1313** ± 0.005	0.5992
250 bars	0.0161 ± 0.0025	-0.0 ± 0.0	0.0067 ± 0.0017	0.0429 ± 0.0053	0.0024 ± 0.0012	**0.0899** ± 0.0054	0.5874
500 bars	0.0102 ± 0.0029	0.0001 ± 0.0001	0.004 ± 0.0032	0.021 ± 0.0039	-0.0012 ± 0.0014	**0.0569** ± 0.006	0.5524
1000 bars	**0.0002** ± 0.0032	0.0 ± 0.0	-0.006 ± 0.0026	0.0001 ± 0.0043	-0.0009 ± 0.0025	-0.0001 ± 0.004	0.5104
1500 bars	-0.0077 ± 0.0035	-0.0 ± 0.0	-0.0018 ± 0.002	-0.0152 ± 0.0052	-0.0041 ± 0.0036	**0.0015** ± 0.0043	0.5096
2000 bars	-0.009 ± 0.0031	-0.0 ± 0.0	-0.0084 ± 0.0044	-0.013 ± 0.0044	0.0004 ± 0.0034	**0.0125** ± 0.0047	0.5199

Appendix A.2 --- Creating the ETF from the raw data series

The calculation details are as follows.

- o_t is the raw open price at bar $t = 1, \dots, T$.
- p_t is the raw open price at bar $t = 1, \dots, T$.
- h_t is the fractional number of shares invested in the instrument at $t = 1, \dots, T$.
- K_t is ETF price, which is the value of the initial \$1 investment in the instrument at $t = 1, \dots, T$. By definition $K_0 = 1$.
- B is the subset of bars when the contract rolls. $t \in B$ means the contract rolls at time t.

Then

$$h_t = \begin{cases} \dfrac{K_t}{o_{t+1}} & \text{if } t \in B \\ h_{t-1} & \text{otherwise} \end{cases}$$

$$\delta_t = \begin{cases} p_t - o_t & \text{if } (t-1) \in B \\ \Delta p_t & \text{otherwise} \end{cases}$$

$$K_t = K_{t-1} + h_{t-1}\delta_t \ .$$