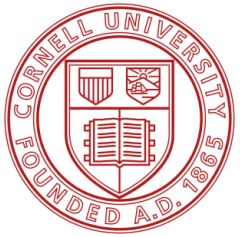


Advances in Financial Machine Learning: Lecture 8/10

Prof. Marcos López de Prado

Advances in Financial Machine Learning

ORIE 5256

What are we going to learn today? (1/2)

- Structural Breaks
  - CUSUM tests
  - Explosiveness tests
    - Right-tail unit-root tests
    - Sub/super-martingale tests
- Entropy Features
  - Shannon entropy
  - The plug-in estimator
  - Lempel-Ziv estimators
  - Encoding schemes
  - Entropy of a Gaussian process
  - Entropy and the generalized mean

What are we going to learn today? (2/2)

- Microstructural Features
  - First generation: Price sequences
    - The tick rule
    - The roll model
    - The high-low volatility estimator
    - The Corwin-Schulz bid-ask spread model
  - Second generation: Strategic trade models
    - Kyle's lambda
    - Amihud's lambda
    - Hasbrouck's lambda
  - Third generation: Sequential trade models
    - Probability of information-based trading
    - Volume-synchronized probability of informed trading

SECTION I
Structural Breaks

CUSUM test: Brown-Durbin-Evans

- It estimates standardized recursive least squares forecasting errors as

$$\widehat{\omega}_t = \frac{y_t - \widehat{\beta}'_{t-1} x_t}{\sqrt{f_t}}$$

$$f_t = \hat{\sigma}_\varepsilon^2 [1 + x_t{}' (X_t' X_t)^{-1} x_t]$$

- It compares the observed cumulative sum of forecasting errors (S_t) against its theoretical distribution.

$$S_t = \sum_{j=k+1}^{t} \frac{\widehat{\omega}_j}{\widehat{\sigma}_\omega}$$

$$\hat{\sigma}_\omega^2 = \frac{1}{T-k} \sum_{t=k}^{T} (\widehat{\omega}_t - \mathrm{E}[\widehat{\omega}_t])^2$$

- Under the null hypothesis $H_0: \beta_t = \beta$, then $S_t \sim N[0, t-k-1]$.
- **Caveat**: Results are sensitive to starting point, which is chosen arbitrarily.

CUSUM test: Chu-Stinchcombe-White

- It assumes $\mathrm{E}_{t-1}[\Delta y_t] = 0$, and works directly with levels y_t (log-prices).
- We compute the standardized departure of log-price y_t relative to the log-price at y_n, $t > n$, as

$$S_{n,t} = (y_t - y_n)\left(\hat{\sigma}_t\sqrt{t-n}\right)^{-1}$$

$$\hat{\sigma}_t^2 = (t-1)^{-1}\sum_{i=2}^{t}(\Delta y_i)^2$$

- Under the null hypothesis $H_0: \beta_t = 0$, then $S_{n,t} \sim N[0,1]$.
- The time-dependent critical value for the *one-sided test* is ($b_\alpha \approx 4.6$ for $\alpha = .05$)

$$c_\alpha[n,t] = \sqrt{b_\alpha + \log[t-n]}$$

- **Caveat**: Results are sensitive to the reference level, y_n, which is chosen arbitrarily.

EXPLOSIVENESS: Chow-Type Dickey-Fuller (1/2)

- Consider the first order autoregressive process with white noise ε_t

$$y_t = \rho y_{t-1} + \varepsilon_t$$

- The null hypothesis is that y_t follows a random walk, $H_0: \rho = 1$, and the alternative hypothesis is that y_t starts as a random walk but changes at time $\tau^* T$, where $\tau^* \in (0,1)$, into an explosive process:

$$H_1: y_t = \begin{cases} y_{t-1} + \varepsilon_t \text{ for } t = 1, \ldots, \tau^* T \\ \rho y_{t-1} + \varepsilon_t \text{ for } t = \tau^* T + 1, \ldots, T, \text{with } \rho > 1 \end{cases}$$

- At time T we can test for a switch (from random walk to explosive process) having taken place at time $\tau^* T$ (break date). In order to test this hypothesis, we fit the following specification,

$$\Delta y_t = \delta y_{t-1} D_t[\tau^*] + \varepsilon_t$$

where $D_t[\tau^*]$ is a dummy variable that takes zero value if $t < \tau^* T$, and takes the value one if $t \geq \tau^* T$.

EXPLOSIVENESS: Chow-Type Dickey-Fuller (2/2)

- Then, the null hypothesis $H_0: \delta = 0$ is tested against the (one-sided) alternative $H_1: \delta > 1$:

$$DFC_{\tau^*} = \frac{\hat{\delta}}{\hat{\sigma}_\delta}$$

- The main drawback of this method is that τ^* is unknown.
- To address this issue, Andrews [1993] proposed a new test where all possible τ^* are tried, within some interval $\tau^* \in [\tau_0, 1 - \tau_0]$.
- The test statistic for an unknown τ^* is the maximum of all $T(1 - 2\tau_0)$ values of DFC_{τ^*}

$$SDFC = \sup_{\tau^* \in [\tau_0, 1-\tau_0]} \{DFC_{\tau^*}\}$$

- Another drawback of Chow's approach is that it assumes that there is only one break date $\tau^* T$, and that the bubble runs up to the end of the sample (there is no switch back to a random walk). For situations where three or more regimes (random walk → bubble → random walk . . .) exist, this is problematic.

EXPLOSIVENESS: SADF (1/2)

- In the words of Phillips, Wu and Yu [2011]: *"[S]tandard unit root and cointegration tests are inappropriate tools for detecting bubble behavior because they cannot effectively distinguish between a stationary process and a periodically collapsing bubble model. Patterns of periodically collapsing bubbles in the data look more like data generated from a unit root or stationary autoregression than a potentially explosive process."*
- To address this flaw, these authors propose fitting the regression specification

$$\Delta y_t = \alpha + \beta y_{t-1} + \sum_{l=1}^{L} \gamma_l \, \Delta y_{t-l} + \varepsilon_t$$

where we test for $H_0: \beta \leq 0$, $H_1: \beta > 0$. Inspired by Andrews [1993], Phillips and Yu [2011] and Phillips, Wu and Yu [2011] proposed the Supremum Augmented Dickey-Fuller test (SADF).

EXPLOSIVENESS: SADF (2/2)

- SADF fits the above regression at each end point t with backwards expanding start points, then computes

$$SADF_t = \sup_{t_0 \in [1, t-\tau]} \{ADF_{t_0,t}\} = \sup_{t_0 \in [1, t-\tau]} \left\{ \frac{\hat{\beta}_{t_0,t}}{\hat{\sigma}_{\beta_{t_0,t}}} \right\}$$

  where $\hat{\beta}_{t_0,t}$ is estimated on a sample that starts at t_0 and ends at t, τ is the minimum sample length used in the analysis, t_0 is the left bound of the backwards expanding window, and $t = \tau, \ldots, T$. For the estimation of $SADF_t$, the right side of the window is fixed at t. The standard ADF tests is a special case of $SADF_t$, where $\tau = t$.

- There are two critical differences between $SADF_t$ and SDFC: First, $SADF_t$ is computed at each $t \in [\tau, T]$, whereas SDFC is computed only at T. Second, instead of introducing a dummy variable, SADF recursively expands the beginning of the sample ($t_0 \in [1, t - \tau]$). By trying all combinations of a nested double loop on (t_0, t), SADF does not assume a known number of regime switches or break dates.

EXPLOSIVENESS: Sub/Super-Martingale (1/3)

- Consider a process that is either a sub- or super-martingale. Given some observations $\{y_t\}$, we would like to test for the existence of an explosive time trend, $H_0: \beta = 0$, $H_1: \beta \neq 0$, under alternative specifications.
- Polynomial trend (SM-Poly1):

$$y_t = \alpha + \gamma t + \beta t^2 + \varepsilon_t$$

- Polynomial trend (SM-Poly2):

$$\log[y_t] = \alpha + \gamma t + \beta t^2 + \varepsilon_t$$

- Exponential trend (SM-Exp):

$$y_t = \alpha e^{\beta t} + \varepsilon_t \Longrightarrow \log[y_t] = \log[\alpha] + \beta t + \xi_t$$

- Power trend (SM-Power):

$$y_t = \alpha t^{\beta} + \varepsilon_t \Longrightarrow \log[y_t] = \log[\alpha] + \beta \log[t] + \xi_t$$

EXPLOSIVENESS: Sub/Super-Martingale (2/3)

- Similar to SADF, we fit any of these specifications to each end point $t = \tau, \ldots, T$, with backwards expanding start points, then compute

$$SMT_t = \sup_{t_0 \in [1, t-\tau]} \left\{ \frac{\left| \hat{\beta}_{t_0,t} \right|}{\hat{\sigma}_{\beta_{t_0,t}}} \right\}$$

- The reason for the absolute value is that we are equally interested in explosive growth and collapse. In the simple regression case (Greene [2008], p. 48), the variance of β is $\hat{\sigma}_\beta^2 = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_{xx}^2 (t - t_0)}$, hence $\lim_{t \to \infty} \hat{\sigma}_{\beta_{t_0,t}} = 0$. The same result is generalizable to the multivariate linear regression case (Greene [2008], pp. 51–52).
- **Problem**: The $\hat{\sigma}_\beta^2$ of a weak long-run bubble may be smaller than the $\hat{\sigma}_\beta^2$ of a strong short-run bubble, hence biasing the method towards long-run bubbles.

EXPLOSIVENESS: Sub/Super-Martingale (3/3)

- **Solution**: We can penalize large sample lengths by determining the coefficient $\varphi \in [0,1]$ that yields best explosiveness signals

$$SMT_t = \sup_{t_0 \in [1, t-\tau]} \left\{ \frac{\left|\hat{\beta}_{t_0,t}\right|}{\hat{\sigma}_{\beta_{t_0,t}} (t - t_0)^{\varphi}} \right\}$$

- For instance,
  - when $\varphi = 0.5$, we compensate for the lower $\hat{\sigma}_{\beta_{t_0,t}}$ associated with longer sample lengths, in the simple regression case.
  - For $\varphi \to 0$, SMT_t will exhibit longer trends, as that compensation wanes and long-run bubbles mask short-run bubbles.
  - For $\varphi \to 1$, SMT_t becomes noisier, because more short-run bubbles are selected over long-run bubbles.
- Consequently, this is a natural way to adjust the explosiveness signal, so that it filters opportunities targeting a particular holding period.
- The features used by the ML algorithm may include SMT_t estimated from a wide range of φ values.

SECTION II
Entropy Features

Entropy

- Let X be a discrete random variable that takes a value x from the set S_X with probability $p[x]$. The entropy of X is defined as

$$H[X] = -\sum_{x \in S_X} p[x] \log\big[p[x]\big]$$

- A few observations:
  - The value $\frac{1}{p[x]}$ measures how surprising an observation is, because surprising observations are characterized by their low probability.
  - Entropy is the expected value of those surprises, where the $\log[.]$ function prevents that $p[x]$ cancels $\frac{1}{p[x]}$ and endows entropy with desirable mathematical properties.
  - Accordingly, entropy can be interpreted as the amount of uncertainty associated with X. Entropy is zero when all probability is concentrated in a single element of S_X. Entropy reaches a maximum at $\log[\|S_X\|]$ when X is distributed uniformly, $p[x] = \frac{1}{\|S_X\|}, \forall x \in S_X$.

Time Series Entropy: The Plug-In Estimator

- Given a data sequence x_1^n, comprising the string of values starting in position 1 and ending in position *n*, we can form a dictionary of all words of length $w < n$ in that sequence, A^w.
- Consider an arbitrary word $y_1^w \in A^w$ of length w. We denote $\hat{p}_w[y_1^w]$ the empirical probability of the word y_1^w in x_1^n, which means that $\hat{p}_w[y_1^w]$ is the frequency with which y_1^w appears in x_1^n.
- Assuming that the data is generated by a stationary and ergodic process, then the law of large numbers guarantees that, for a fixed *w* and large *n*, the empirical distribution $\hat{p}_w$ will be close to the true distribution p_w.
- Under these circumstances, a natural estimator for the entropy rate (i.e., average entropy per bit) is

$$\widehat{H}_{n,w} = -\frac{1}{w} \sum_{y_1^w \in A^w} \hat{p}_w[y_1^w] \log_2 \hat{p}_w[y_1^w]$$

Time Series Entropy: Lempel-Ziv Estimators

- Plug-in estimators require large samples.
- Kontoyiannis [1998] attempts to make a more efficient use of the information available in a message.
- Let us define L_i^n as 1 plus the length of the longest match found in the *n* bits prior to *i*,

$$L_i^n = 1 + \max\left\{l \middle| x_i^{i+l} = x_j^{j+l} \text{ for some } i-n \le j \le i-1, l \in [0,n]\right\}$$

- The general intuition is, as we increase the available history, we expect that messages with high entropy will produce relatively shorter non-redundant substrings. In contrast, messages with low entropy will produce relatively longer non-redundant substrings as we parse through the message.
- The sliding-window LZ estimator $\widehat{H}_{n,k} = \widehat{H}_{n,k}\left[x_{-n+1}^{n+k-1}\right]$ is defined by

$$\widehat{H}_{n,k} = \left[\frac{1}{k}\sum_{i=1}^{k}\frac{L_i^n}{\log_2[n]}\right]^{-1}$$

Time Series Entropy: Encoding Schemes (1/2)

- Entropy rate estimation requires the discretization of a continuous variable, so that each value can be assigned a code from a finite alphabet.
- **Binary Encoding**:
  - A stream of returns r_t can be encoded according to the sign, 1 for $r_t > 0$, 0 for $r_t < 0$, removing cases where $r_t = 0$.
  - Binary encoding arises naturally in the case of returns series sampled from price bars (i.e., bars that contain prices fluctuating between two symmetric horizontal barriers, centered around the start price), because $|r_t|$ is approximately constant.
- **Quantile Encoding**:
  - Unless price bars are used, it is likely that more than two codes will be needed.
  - One approach consists in assigning a code to each r_t according to the quantile it belongs to.
  - The quantile boundaries are determined using an in-sample period (training set).
  - There will be the same number of observations assigned to each letter for the overall in-sample, and close to the same number of observations per letter out-of-sample.
  - This uniform (in-sample) or close to uniform (out-of-sample) distribution of codes tends to increase entropy readings on average.

Time Series Entropy: Encoding Schemes (2/2)

- **Sigma Encoding**:
  - As an alternative approach, rather than fixing the number of codes, we could let the price stream determine the actual dictionary.
  - Suppose we fix a discretization step, σ. Then, we assign the value 0 to $r_t \in [\min\{r\}, \min\{r\} + \sigma)$, 1 to $r_t \in [\min\{r\} + \sigma, \min\{r\} + 2\sigma)$ and so on until every observation has been encoded with a total of $ceil\left[\frac{\max\{r\} - \min\{r\}}{\sigma}\right]$ codes, where $ceil[.]$ is the ceiling function.
  - Unlike quantile encoding, now each code covers the same fraction of r_t's range.
  - Because codes are not uniformly distributed, entropy readings will tend to be smaller than in quantile encoding on average; however, the appearance of a "rare" code will cause spikes in entropy readings.

Entropy and the Generalized Mean (1/4)

- Here is an interesting way of thinking about entropy.
- Consider a set of real numbers $x = \{x_i\}_{i=1,\dots,n}$ and weights $p = \{p_i\}_{i=1,\dots,n}$, such that $0 \leq p_i \leq 1, \forall i$ and $\sum_{i=1}^{n} p_i = 1$.
- The generalized weighted mean of *x* with weights *p* on a power $q \neq 0$ is defined as

$$M_q[x,p] = \left(\sum_{i=1}^{n} p_i x_i^q\right)^{1/q}$$

- For $q < 0$, we must require that $x_i > 0, \forall i$.

Entropy and the Generalized Mean (2/4)

- The reason this is a generalized mean is that other means can be obtained as special cases:
  - Minimum: $\lim_{q\to-\infty} M_q[x,p] = \min_i\{x_i\}$
  - Harmonic mean: $M_{-1}[x,p] = \left(\sum_{i=1}^n p_i x_i^{-1}\right)^{-1}$
  - Geometric mean: $\lim_{q\to 0} M_q[x,p] = e^{\sum_{i=1}^n p_i \log[x_i]} = \prod_{i=1}^n x_i{}^{p_i}$
  - Arithmetic mean: $M_1\left[x,\{n^{-1}\}_{i=1,\ldots,n}\right] = n^{-1}\sum_{i=1}^n x_i$
  - Weighted mean: $M_1[x,p] = \sum_{i=1}^n p_i x_i$
  - Quadratic mean: $M_2[x,p] = \left(\sum_{i=1}^n p_i x_i^2\right)^{1/2}$
  - Maximum: $\lim_{q\to+\infty} M_q[x,p] = \max_i\{x_i\}$
- In the context of information theory, an interesting special case is $x = \{p_i\}_{i=1,\ldots,n}$, hence

$$M_q[p,p] = \left(\sum_{i=1}^n p_i p_i^q\right)^{1/q}$$

Entropy and the Generalized Mean (3/4)

- Let us define the quantity $N_q[p] = \frac{1}{M_{q-1}[p,p]}$, for some $q \neq 1$.
- Again, for $q < 1$ in $N_q[p]$, we must have $p_i > 0, \forall i$.
- If $p_i = \frac{1}{k}$ for $k \in [1, n]$ different indices and $p_i = 0$ elsewhere, then the weight is spread evenly across *k* different items, and $N_q[p] = k$ for $q > 1$.
- In other words, $N_q[p]$ gives us the *effective number* or *diversity* of items in *p*, according to some weighting scheme set by *q*.
- Using Jensen's inequality, we can prove that $\frac{\partial M_q[p,p]}{\partial q} \geq 0$, hence $\frac{\partial N_q[p]}{\partial q} \leq 0$. Smaller values of *q* assign a more uniform weight to elements of the partition, giving relatively more weight to less common elements, and $\lim_{q \to 0} N_q[p]$ is simply the total number of nonzero p_i.

Entropy and the Generalized Mean (4/4)

- Shannon's entropy is $H[p] = \sum_{i=1}^{n} -p_i \log[p_i] = -\log\left[\lim_{q\to 0} M_q[p]\right] = \log\left[\lim_{q\to 1} N_q[p]\right]$.
- This shows that entropy can be interpreted as the logarithm of the *effective number* of items in a list *p*, where $q \to 1$.
- Intuitively, entropy measures information as the level of *diversity* contained in a random variable. This intuition is formalized through the notion of generalized mean.
- The implication is that Shannon's entropy is a special case of a diversity measure (hence its connection with volatility).
- We can now define and compute alternative measures of diversity, other than entropy, where $q \neq 1$.

SECTION III
Microstructural Features

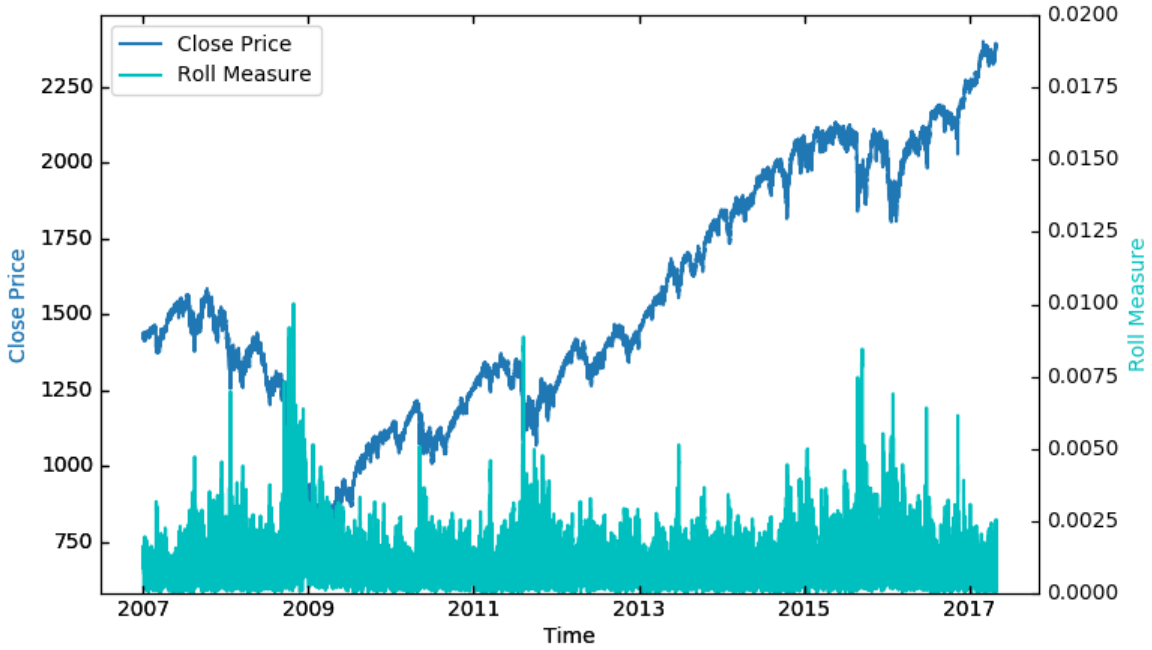
A brief History of Market Microstructure Models

- First generation: price sequences
  - The Roll model [1984]
  - High-Low Volatility Estimator: Beckers [1983], Parkinson [1980]
  - Corwin and Schultz [2012]
- Second generation: strategic trade models
  - Kyle's lambda [1985]
  - Amihud's lambda [2002]
  - Hasbrouck's lambda [2009]
- Third generation: sequential trade models
  - Probability of information-based trading (PIN): Easley et al. [1996]
  - Volume-synchronized probability of informed trading (VPIN): Easley et al. [2011]

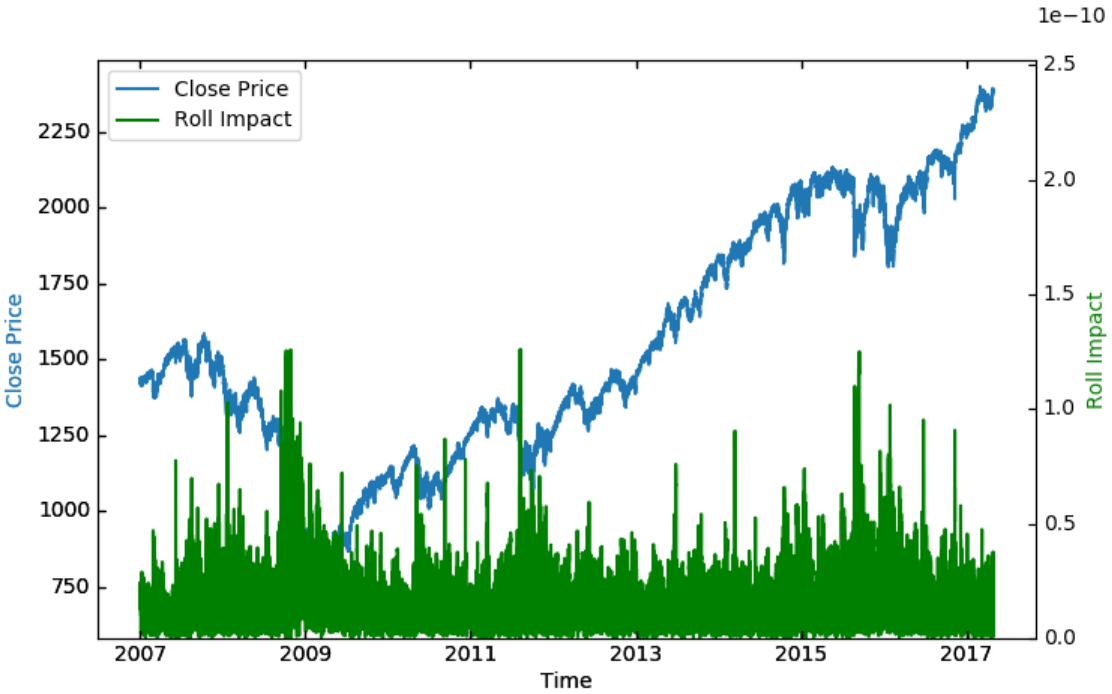
Roll Measure

$$2\sqrt{|\mathrm{cov}[\Delta p_t\ , \Delta p_{t-1}]|}$$

- Δp_t is the change in close price between two bars at time t.
- The covariance is evaluated on a rolling basis for different window size



Roll Impact

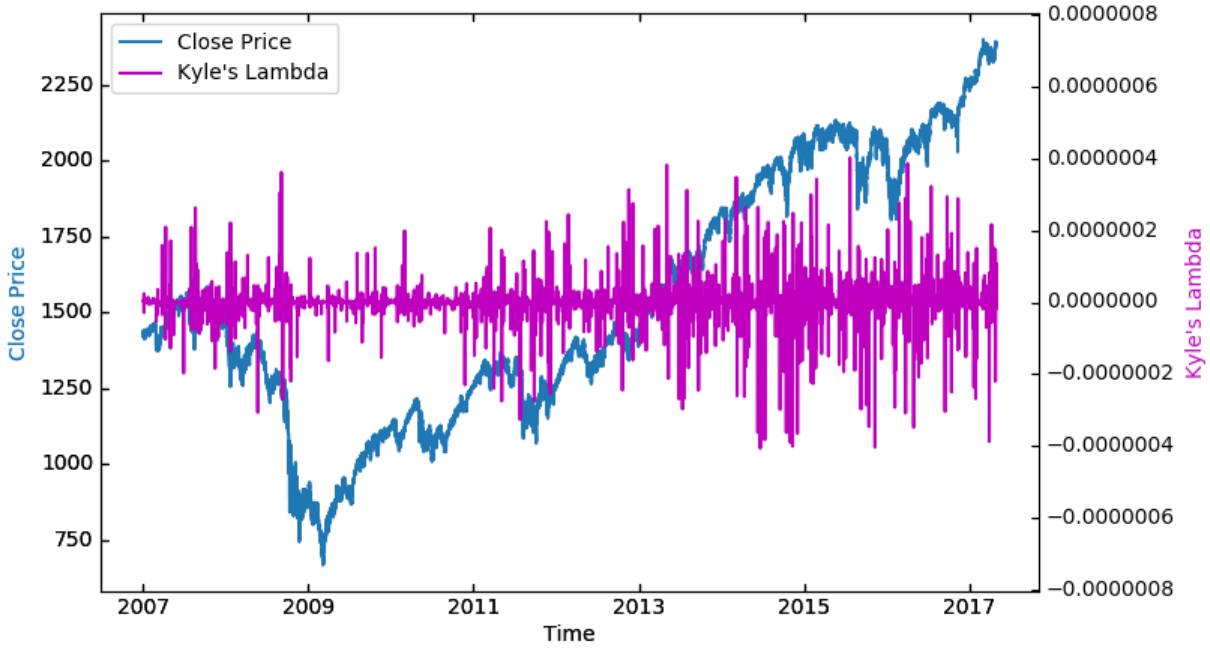


$$\frac{2\sqrt{|\mathrm{cov}[\Delta p_t, \Delta p_{t-1}]|}}{Dollar\ Volume_t}$$

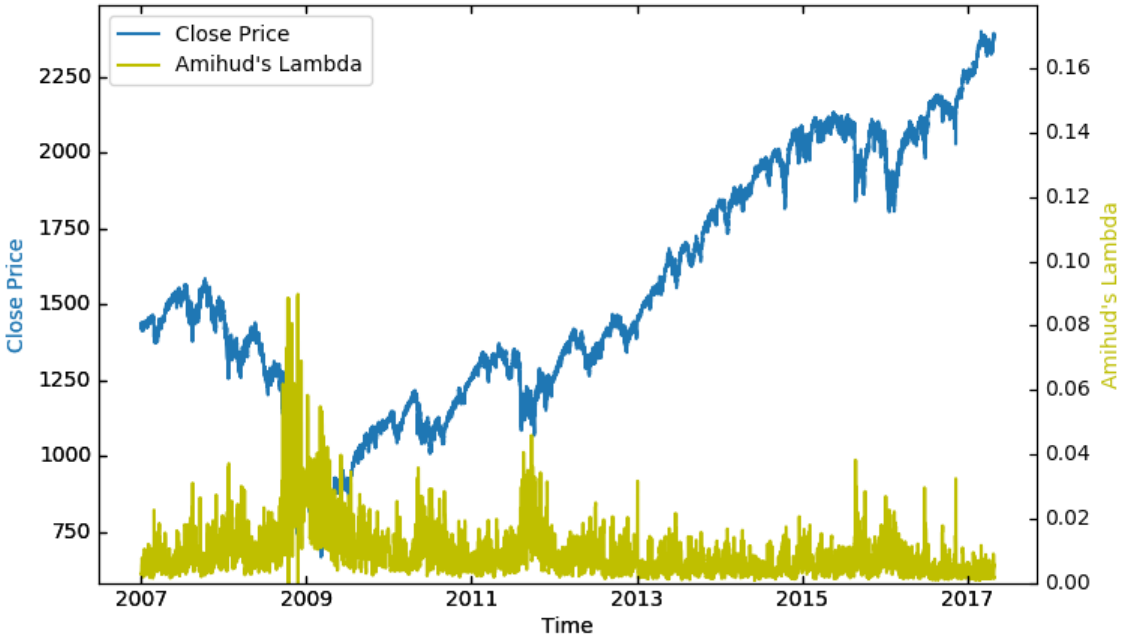
Kyle's Lambda

$$\Delta p_t = \lambda b_t V_t$$

- λ is derived from the regression with a rolling window
- V_t is the volume and $b_t = \text{sign}[p_t - p_{t-1}]$ which is computed on bar level.



Amihud's Lambda

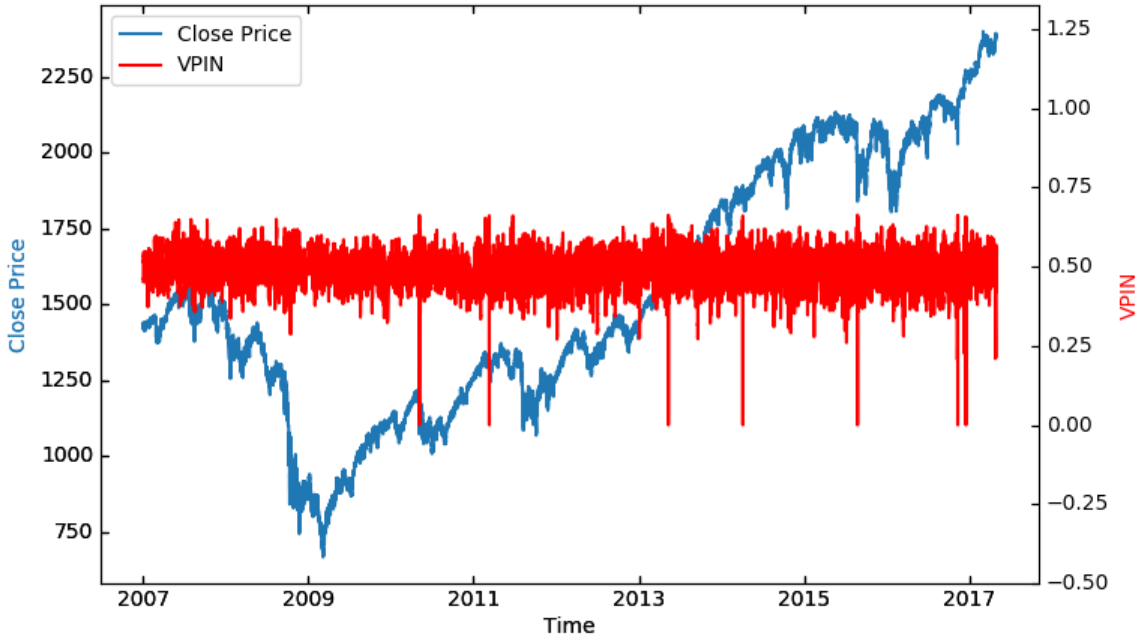


$$Moving\ Average\left[\frac{|Return_t|}{Dollar\ Volume_t}, window\right]$$

VPIN

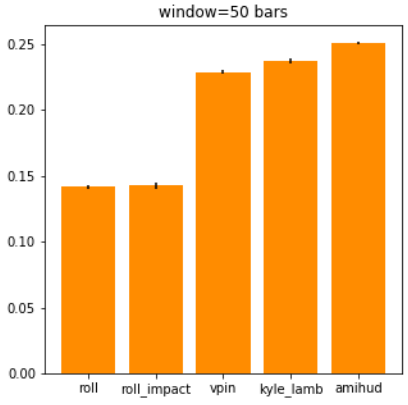
$$\frac{\sum |V_t^S - V_t^B|}{nV}$$

- $V_t^B = V_t Z\left[\frac{\Delta p_t}{\sigma_{\Delta p_t}}\right]$, $V_t^S = V_t - V_t^B$
- n is the number of bars used

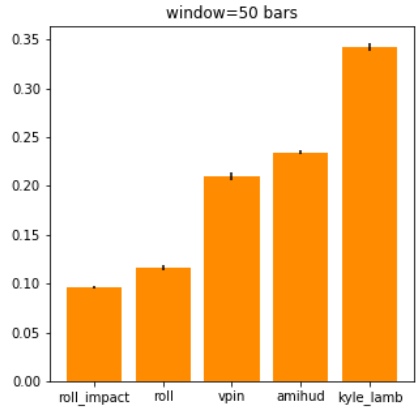


Kyle & Amihud are best In-Sample (1/2)

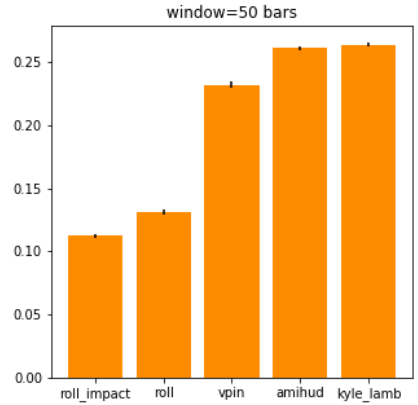
- MDI results have strong similarity across all labels. It is observed that MDI is biased towards features with higher variance (Altmann et al. [2010]). See below for MDI results with 50 bars window.



Corwin-Schultz



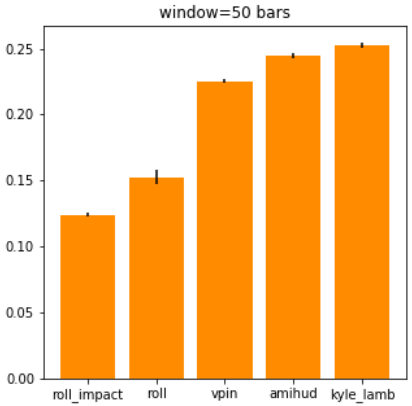
Realized volatility



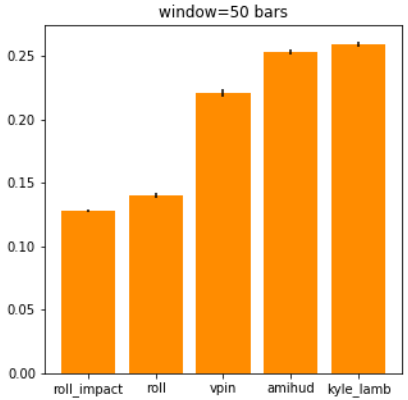
JB statistics

Kyle & Amihud are best In-Sample (2/2)

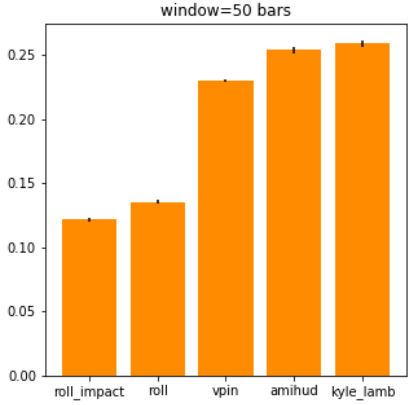
- MDI results have strong similarity across all labels. It is observed that MDI is biased towards features with higher variance (Altmann et al. [2010]). See below for MDI results with 50 bars window.



Sequential correlation



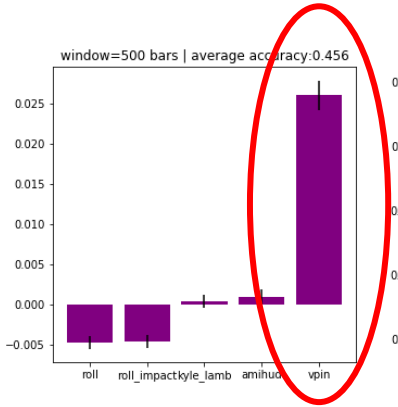
Return skewness



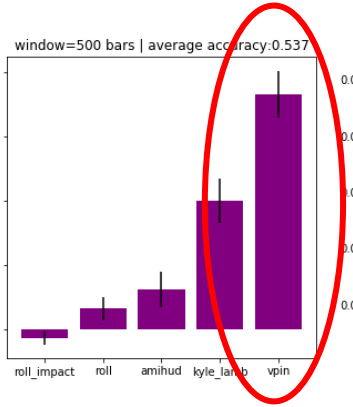
Return kurtosis

VPIN is best Out-Of-Sample (1/2)

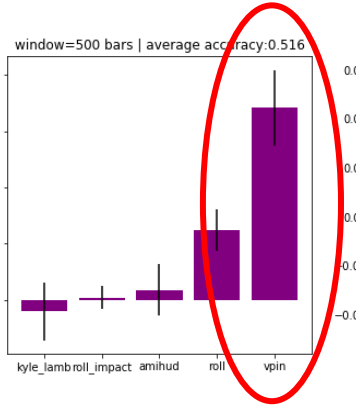
- When the backward window is large, only VPIN can contribute positively to out-of-sample prediction across all labels except sequential correlation. Below are MDA result with 500 bar window



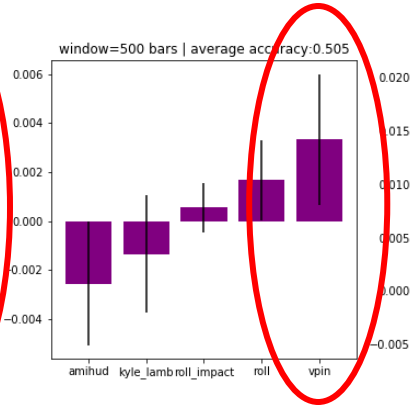
Corwin-Schultz



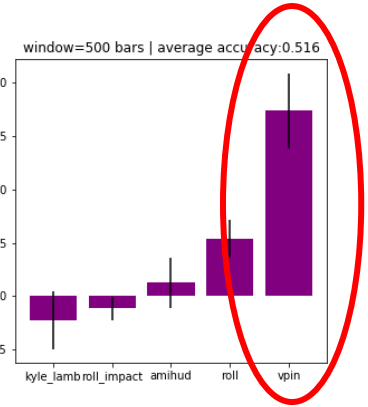
Realized volatility



JB statistics



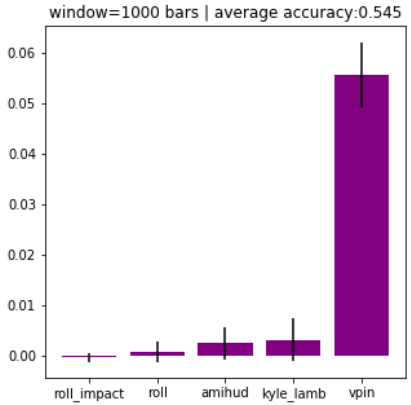
Return skewness



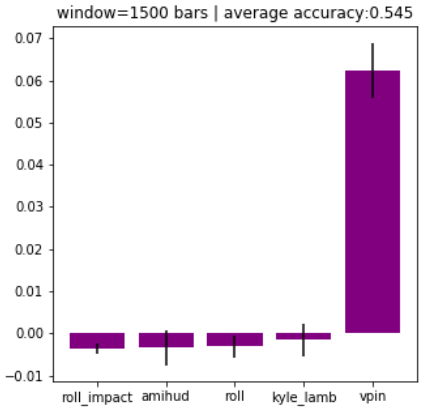
Return kurtosis

VPIN is best Out-Of-Sample (2/2)

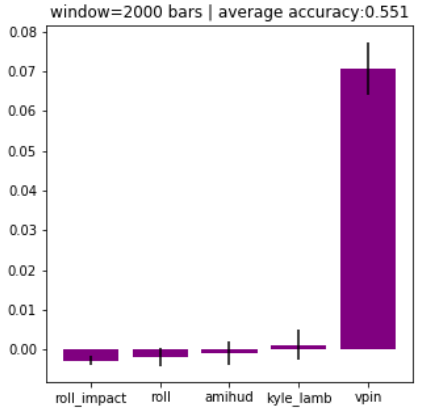
- VPIN's MDA importance at predicting realized volatility remains significant with large window size, even when other variables become irrelevant



1000 bars

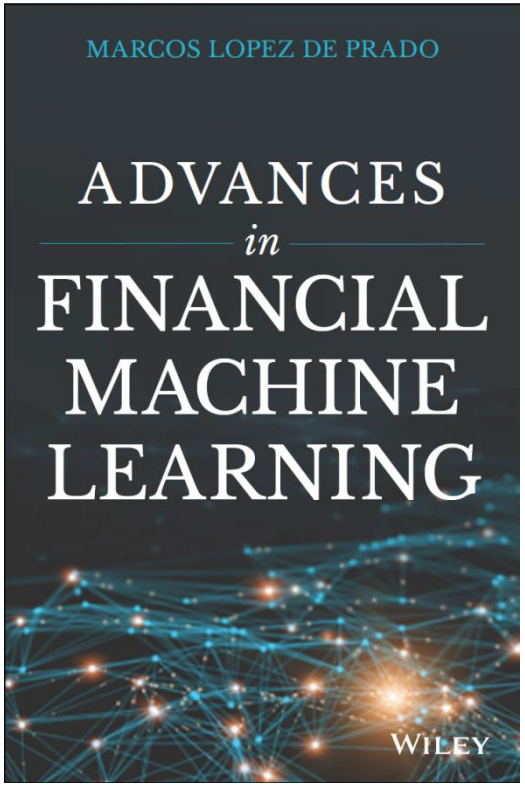


1500 bars



2000 bars

For Additional Details



The first wave of quantitative innovation in finance was led by Markowitz optimization. Machine Learning is the second wave and it will touch every aspect of finance. López de Prado's Advances in Financial Machine Learning is essential for readers who want to be ahead of the technology rather than being replaced by it.
— Prof. **Campbell Harvey**, Duke University. Former President of the American Finance Association.

Financial problems require very distinct machine learning solutions. Dr. López de Prado's book is the first one to characterize what makes standard machine learning tools fail when applied to the field of finance, and the first one to provide practical solutions to unique challenges faced by asset managers. Everyone who wants to understand the future of finance should read this book.
— Prof. **Frank Fabozzi**, EDHEC Business School. Editor of The Journal of Portfolio Management.

Disclaimer

- The views expressed in this document are the authors' and do not necessarily reflect those of the organizations he is affiliated with.
- No investment decision or particular course of action is recommended by this presentation.
-

www.QuantResearch.org

Market Microstructure in the Age of Machine Learning

Marcos López de Prado

Lawrence Berkeley National Laboratory
Computational Research Division

A brief History of Market Microstructure Models

- First generation: price sequences
  - The Roll model [1984]
  - High-Low Volatility Estimator: Beckers [1983], Parkinson [1980]
  - Corwin and Schultz [2012]
- Second generation: strategic trade models
  - Kyle's lambda [1985]
  - Amihud's lambda [2002]
  - Hasbrouck's lambda [2009]
- Third generation: sequential trade models
  - Probability of information-based trading (PIN): Easley et al. [1996]
  - Volume-synchronized probability of informed trading (VPIN): Easley et al. [2011]

Advantages of the AI Age

- Financial Big Data (tick/book level)
- New and advanced statistical techniques (Machine Learning)
- Unprecedented computational power (Supercomputers)

Microstructural relationships often are non-linear and hard to parameterize. When used properly, these technologies can uncover relationships unknown to traditional approaches.

Goals of this Presentation

- Investigate all three-generation market microstructure variables jointly on the most recent 10 year market data
- Apply ML based feature importance analysis and test the usefulness of microstructure variables at predicting various market movements
- Study how the feature importance pattern changes with different labels and different time scales

Section I
Data & Variables

Market Data

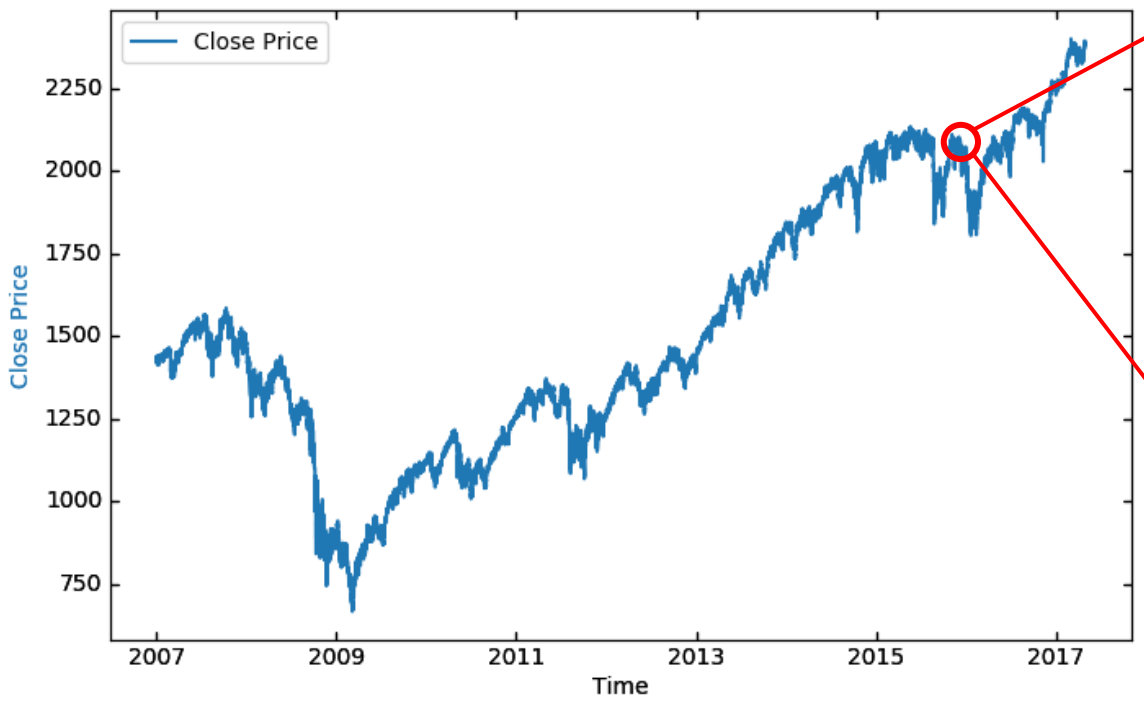
- 87 liquid futures traded globally (index, currency, commodity and fixed-income) with 10 years history
- Tick level trade data, aggregated into bars (price-volume bars). Each bar is formed with a timestamp t when

$$\sum_{\tau=t-1}^{t} p_\tau V_\tau \geq L$$

- The threshold L is chosen to have roughly 50 bars per day in the year 2017. Each bar contains Close, Open, High, Low and Volume

Market Data

A snapshot of ES1 Index data



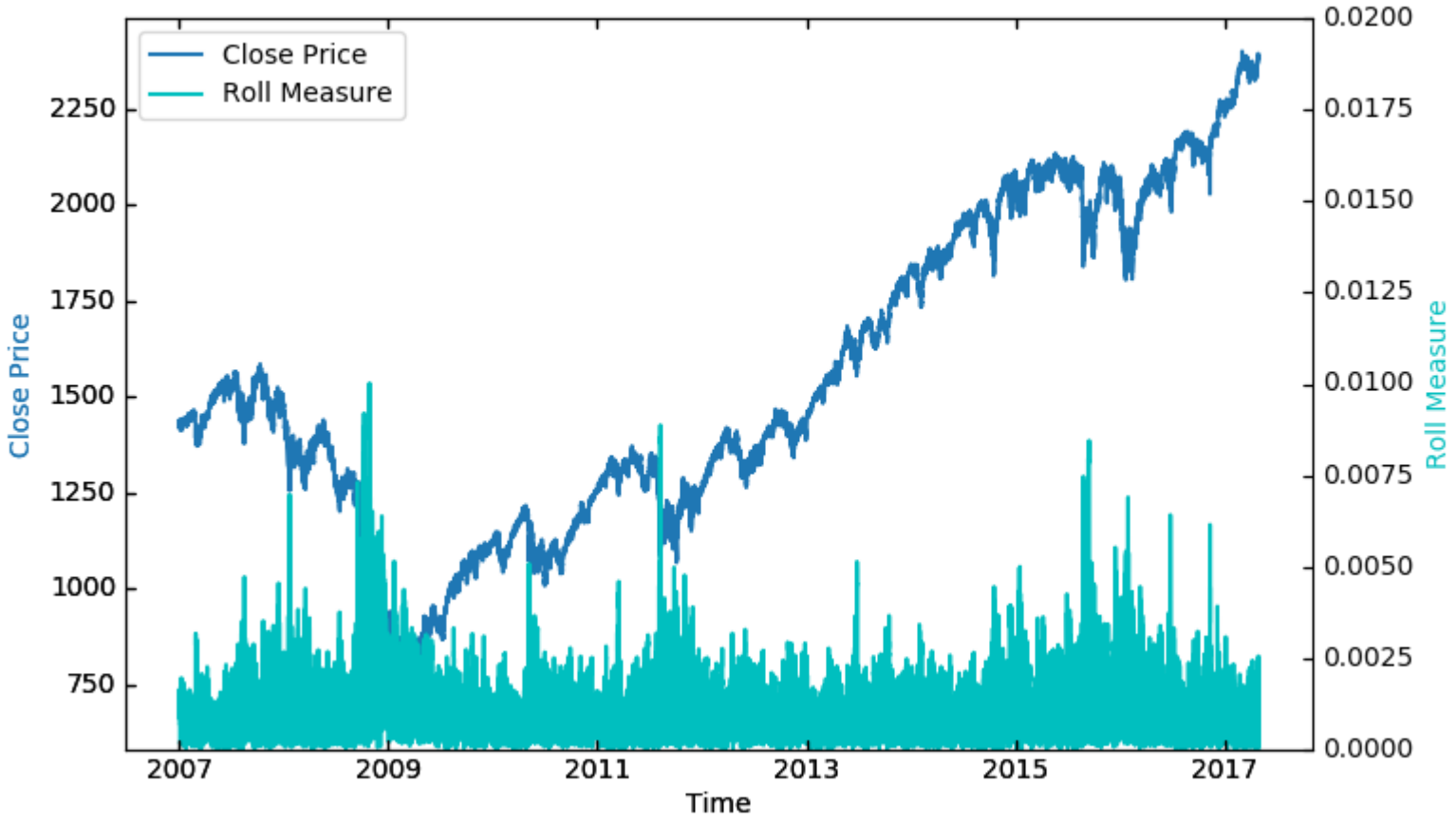
Timestamp	Open	Close	High	Low	Volume
5/23/16 3:14 AM	2052	2045	2053.75	2045	33111
5/23/16 4:10 AM	2045	2052.5	2053.75	2043.5	33033
5/23/16 5:31 AM	2052.5	2051.25	2056	2050.25	32916
5/23/16 7:07 AM	2051.25	2046.25	2052.25	2045	33020
5/23/16 8:43 AM	2046	2049.5	2051.75	2046	33026
5/23/16 9:30 AM	2049.5	2048	2049.75	2047.25	33061
5/23/16 9:35 AM	2048	2049.5	2051.25	2047.5	32996
5/23/16 9:40 AM	2049.5	2050.75	2051	2046.5	33006
5/23/16 9:48 AM	2050.75	2049.5	2051.25	2048	32984
5/23/16 9:54 AM	2049.5	2047	2050	2046.25	33021
5/23/16 10:00 AM	2047	2048.75	2049.75	2046.5	33016
5/23/16 10:06 AM	2048.75	2051.75	2052.25	2048.75	32964
5/23/16 10:14 AM	2051.75	2050	2051.75	2049.25	32972
5/23/16 10:22 AM	2050	2051.5	2052.75	2048.75	33006
5/23/16 10:32 AM	2051.5	2049.5	2052.25	2049.5	32966
5/23/16 10:44 AM	2049.5	2048.75	2051.5	2048.25	33057

Microstructure variables

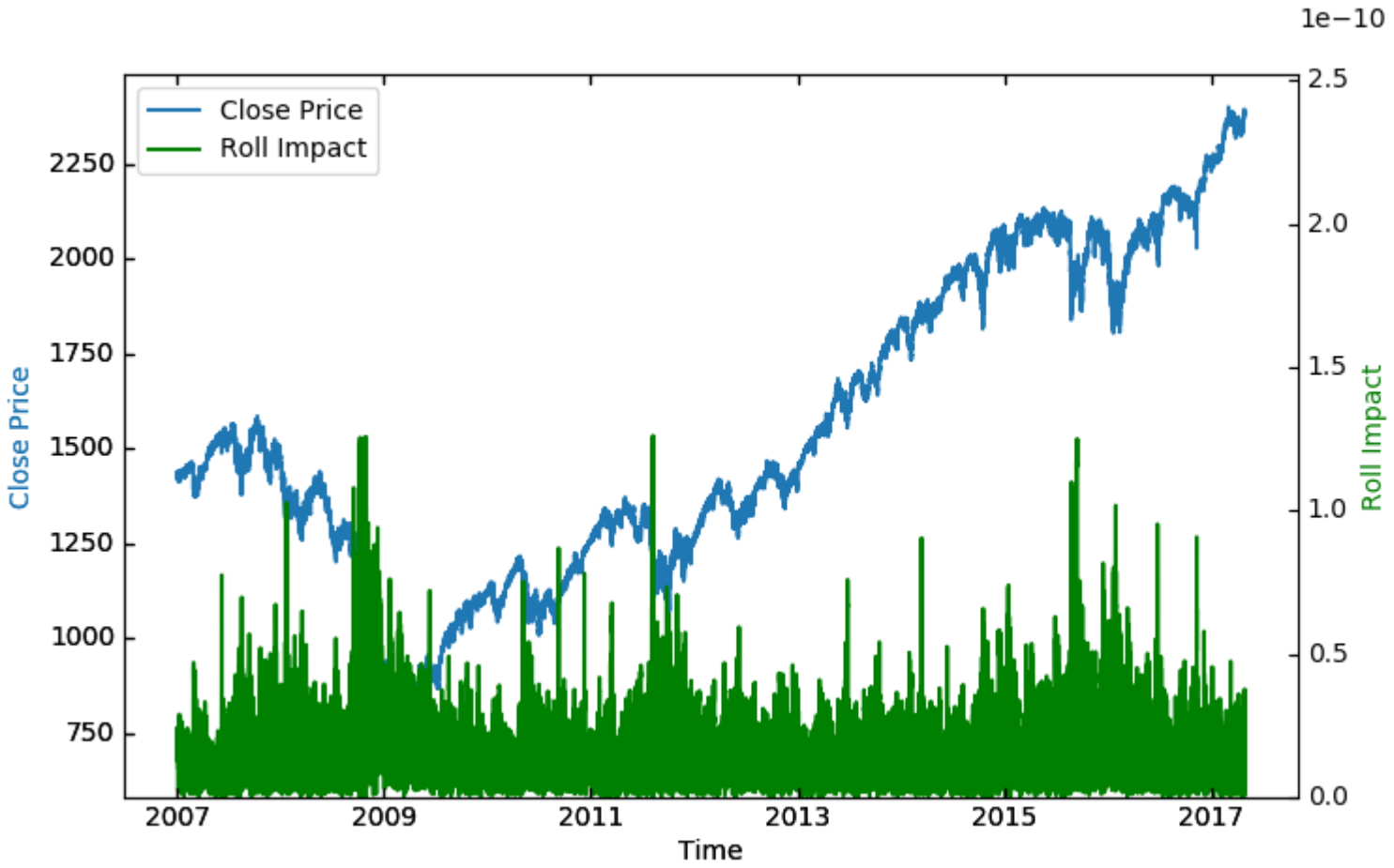
Roll measure

$$2\sqrt{|\mathrm{cov}[\Delta p_t\ , \Delta p_{t-1}]|}$$

- Δp_t is the change in close price between two bars at time t.
- The covariance is evaluated on a rolling basis for different window size (all plots are done with window = 50 bars)



Microstructure variables



Roll impact

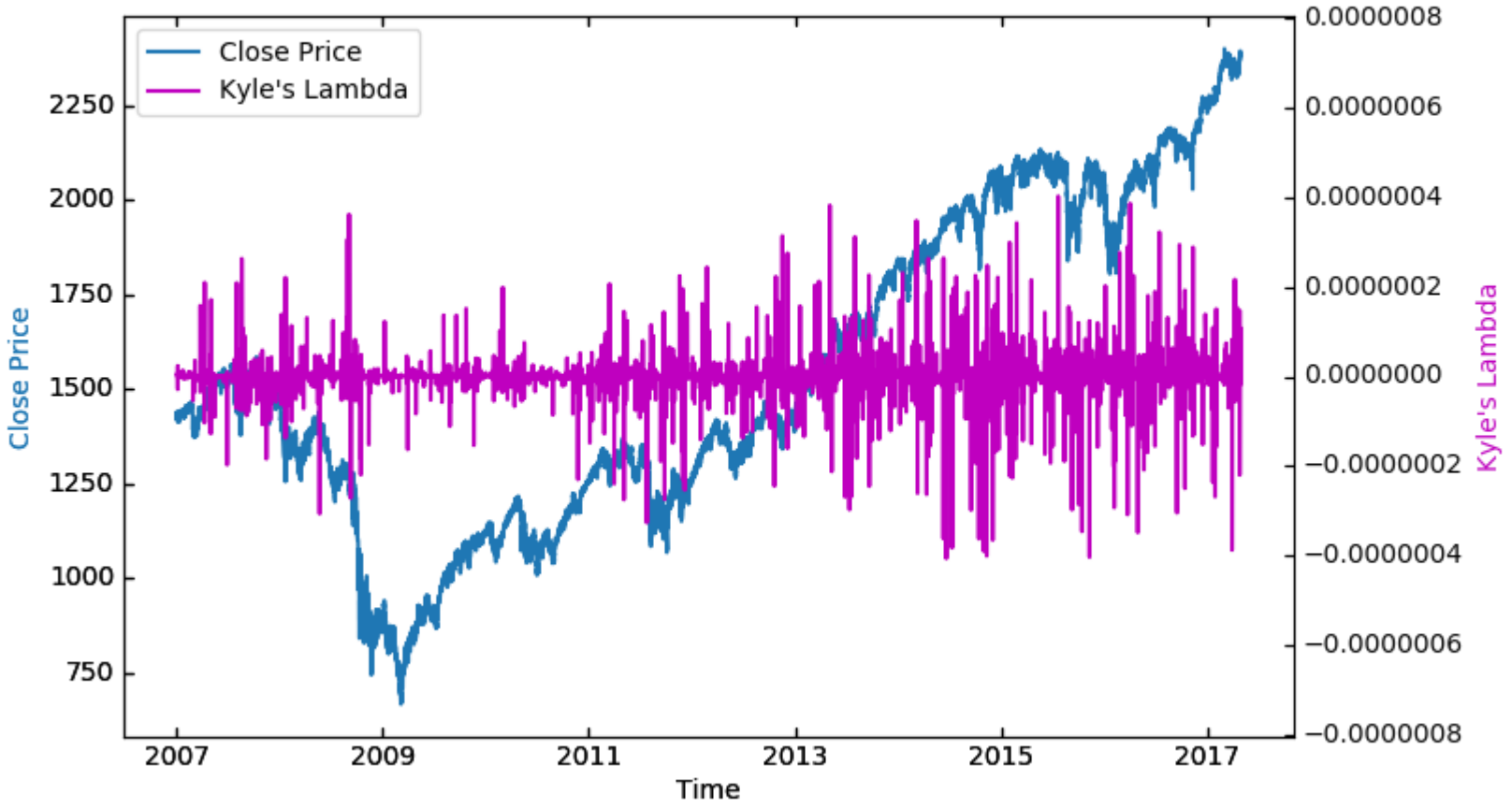
$$\frac{2\sqrt{|\mathrm{cov}[\Delta p_t, \Delta p_{t-1}]|}}{Dollar\ Volume_t}$$

Microstructure variables

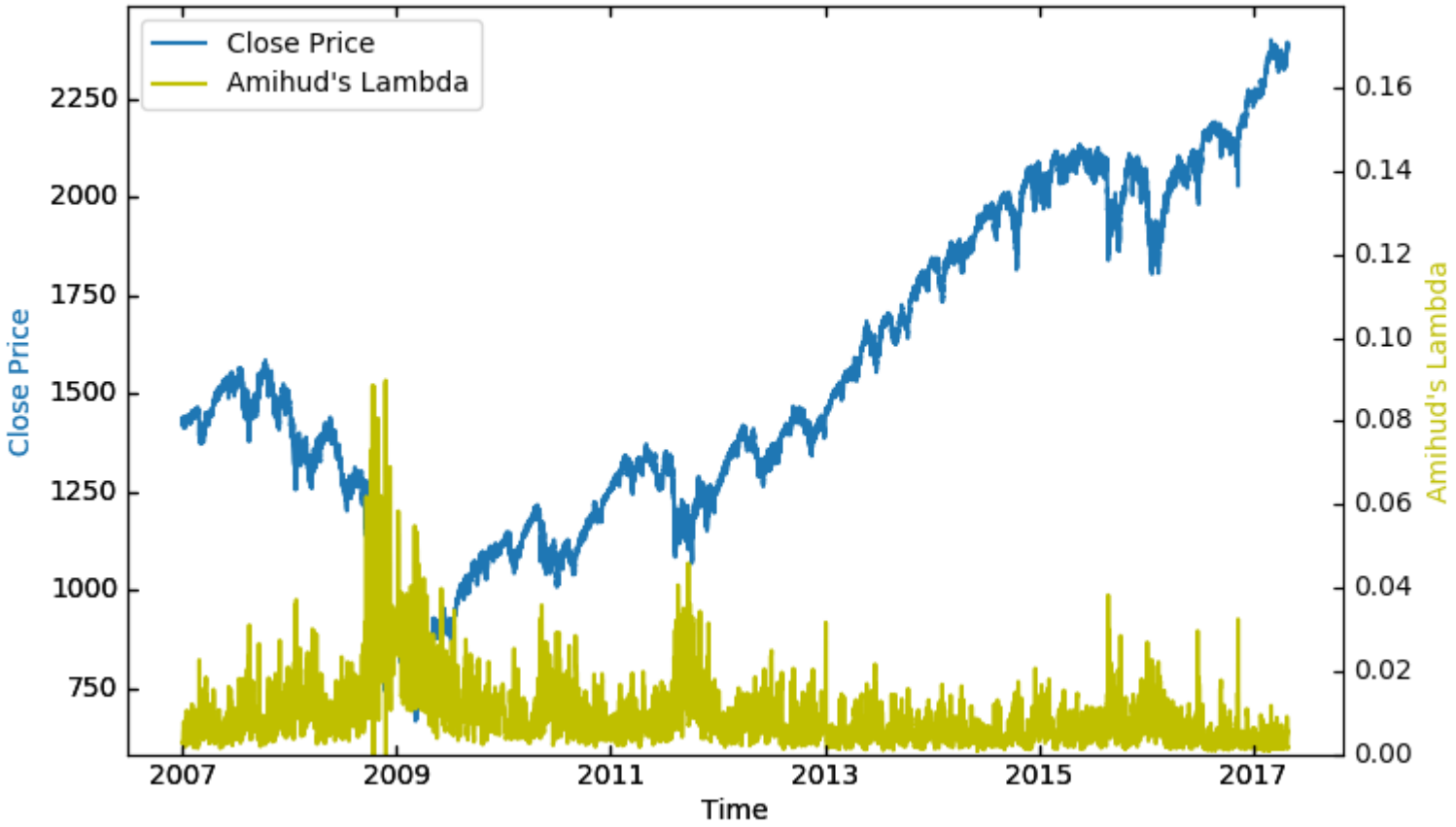
Kyle's λ

$$\Delta p_t = \lambda b_t V_t$$

- λ is derived from the regression with a rolling window
- V_t is the volume and $b_t = \text{sign}[p_t - p_{t-1}]$ which is computed on bar level.



Microstructure variables



Amihud measure

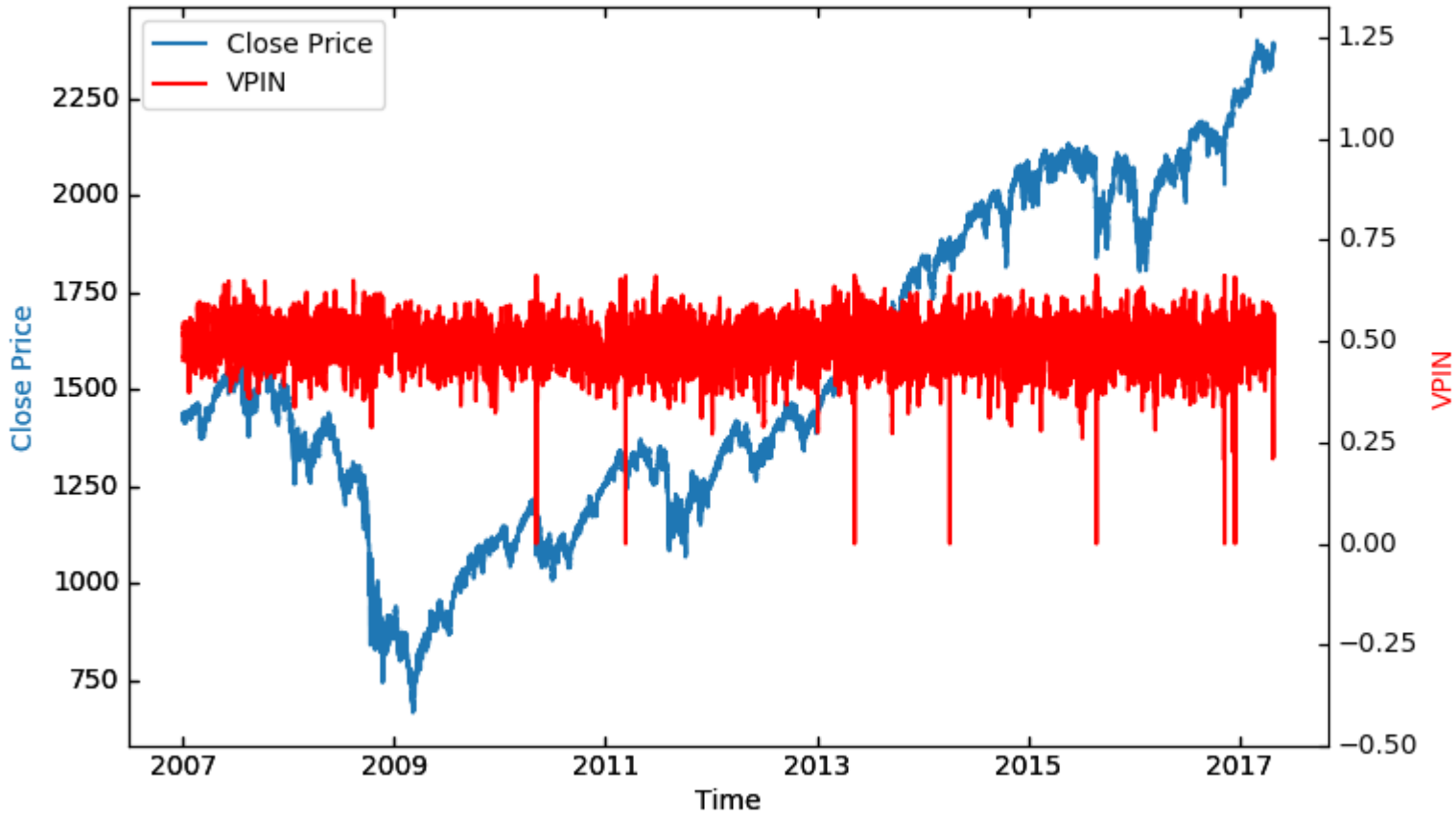
$$Moving\ Average\left[\frac{|Return_t|}{Dollar\ Volume_t}, window\right]$$

Microstructure variables

VPIN

$$\frac{\sum |V_t^S - V_t^B|}{nV}$$

- $V_t^B = V_t Z\left[\frac{\Delta p_t}{\sigma_{\Delta p_t}}\right], \quad V_t^S = V_t - V_t^B$
- n is the number of bars used



Section II
Algorithms & Research Tools

Selection of ML algorithm: Random Forest

- Random Forest is an ensemble method. The bootstrapping process brings in randomness that can reduce potential overfitting, a common issue in finance problems.

- As a tree-based algorithm, Random Forest has a tree-based feature importance method (Mean Decreased Impurity). We can compare it with a more generic ML feature importance method (Mean Decreased Accuracy).

ML-based feature importance analysis

- Mean-Decreased Impurity (MDI). MDI is built for all tree-based ML algorithms including random forest. All tree-based algorithms consist of multiple data splits on selected features, and each split is obtained by minimizing the impurity. MDI measures how much sample size weighted total impurity each feature reduces during training, and rank the feature with largest decreased impurity the highest.

- MDI feature importance is **in-sample** as it is extracted from training data only. It is a statement on explanatory importance rather than predictive importance.

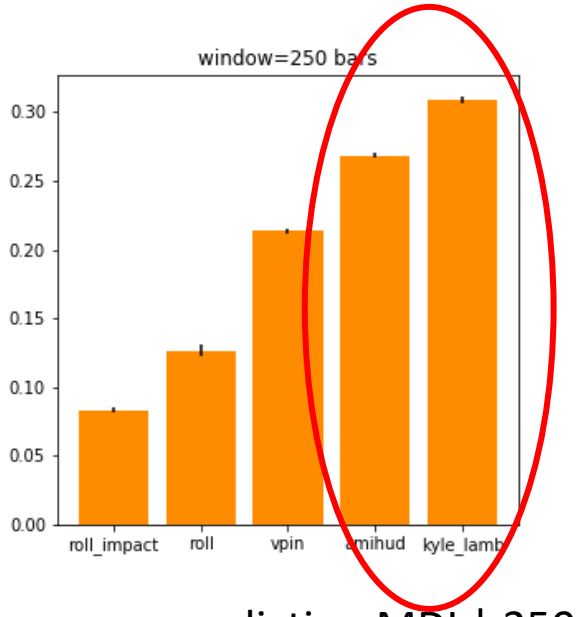
ML-based feature importance analysis

- Mean-Decreased Accuracy (MDA). MDA is a generic feature importance method that applies to all ML algorithms. The idea is to randomly permute the values of each feature and measure how much the permutation decreases the model's out-of-sample accuracy. The more the accuracy decreases indicates the feature is more important.

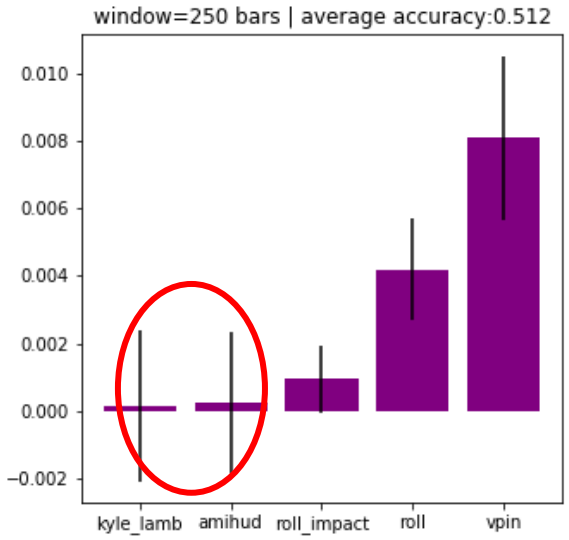
- In contrast to MDI, MDA tests the actual importance for the **out-of-sample** performance.

Analysis

- Feature importance rankings from MDI and MDA are generally different. Important features from MDA might contribute little or even negatively to out-of-sample prediction performance.



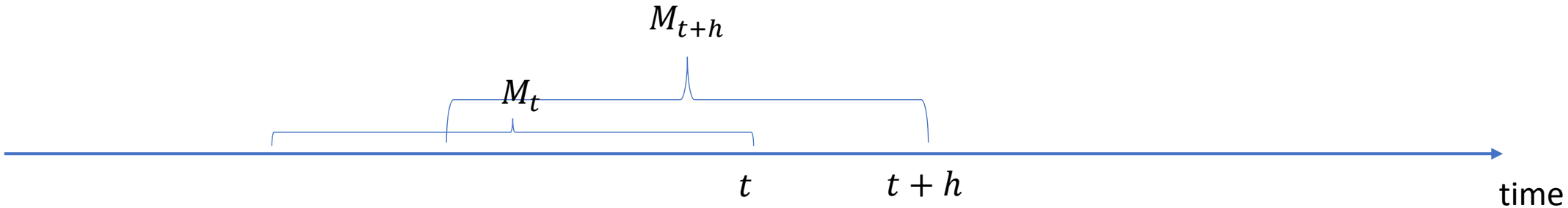
return skewness prediction MDI | 250 bars



return skewness prediction MDA | 250 bars

Prediction labels

Binary Label generation: at time t, we generate a binary label by computing a measure M_t that is constructed with a backward window and comparing it to its value at $t + h$, with $h = 250$ bars (around 5 trading days) for all labels hereafter in this study.



Label at t:
sign[M_{t+h}- M_t]

Prediction labels

Label 1. Sign of change in bid-ask spread estimator (Corwin-Schultz)

$$\text{sign}[S_{t+h} - S_t]$$

$$S_t = \frac{2(e^{\alpha_t}-1)}{1+e^{\alpha_t}}, \quad \alpha_t = \frac{\sqrt{2\beta_t}-\sqrt{\beta_t}}{3-2\sqrt{2}} - \sqrt{\frac{\gamma_t}{3-2\sqrt{2}}}, \quad \beta_t = E\left[\sum_{j=0}^{1}\left[log\left(\frac{H_{t-j}}{L_{t-j}}\right)\right]^2\right], \quad \gamma_t = \left[log\left(\frac{H_{t-1,t}}{L_{t-1,t}}\right)\right]^2$$

Label 2. Sign of change in realized volatility

$$\text{sign}[\sigma_{t+h} - \sigma_t]$$

σ_t is the standard deviation of realized return

Prediction labels

Label 3. Sign of change in Jarque-Bera statistics of realized returns

$$\text{sign}(JB[r_{t+h}] - JB[r_t]) ,$$

$JB[r] = \frac{n}{6}\left(S^2 + \frac{1}{4}(C-3)^2\right)$, S is the skewness and C is the kurtosis of realized return r in the past window

Label 4. Sign of change in serial correlation of realized returns

$$\text{sign}[sc_{t+h} - sc_t]$$

$sc_t = \text{corr}[r_t, r_{t-1}]$, the correlation between returns of two consecutive bars

Prediction labels

Label 5. Sign of change in absolute skewness of realized returns

$$\text{sign}[skew_{t+h} - skew_t]$$

Label 6. Sign of change in kurtosis of realized returns

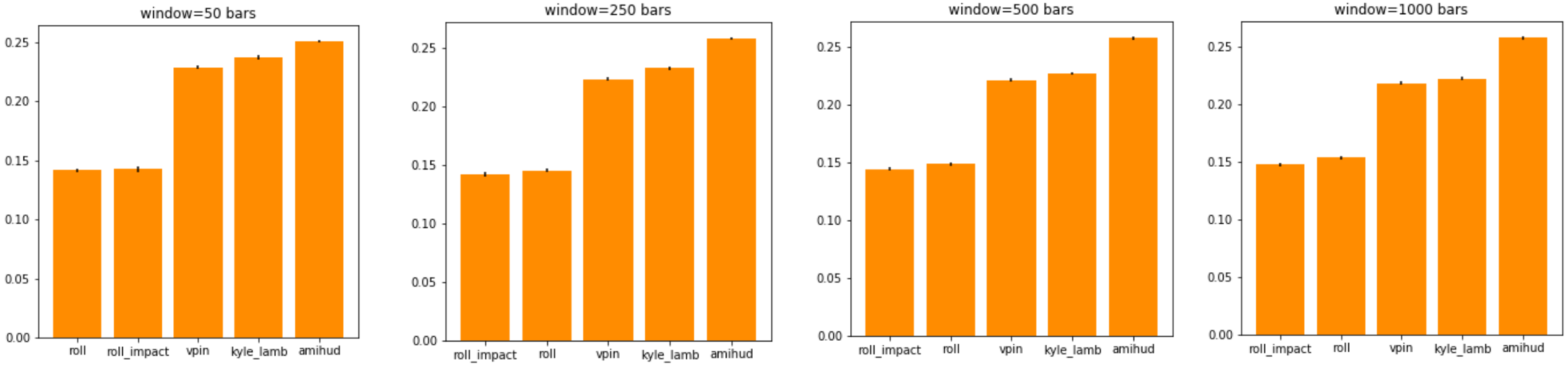
$$\text{sign}[Kurt_{t+h} - Kurt_t]$$

Section III
Results

Microstructure variable feature importance

1. Sign of change in Corwin-Schultz estimator

MDI result

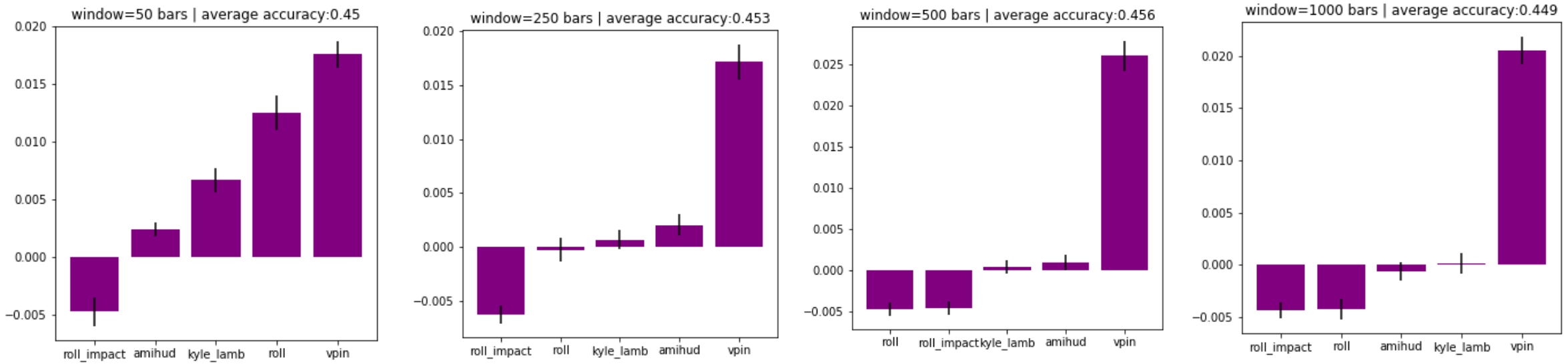


Increasing backward window size for feature generation

Microstructure variable feature importance

1. Sign of change in Corwin-Schultz estimator

MDA result

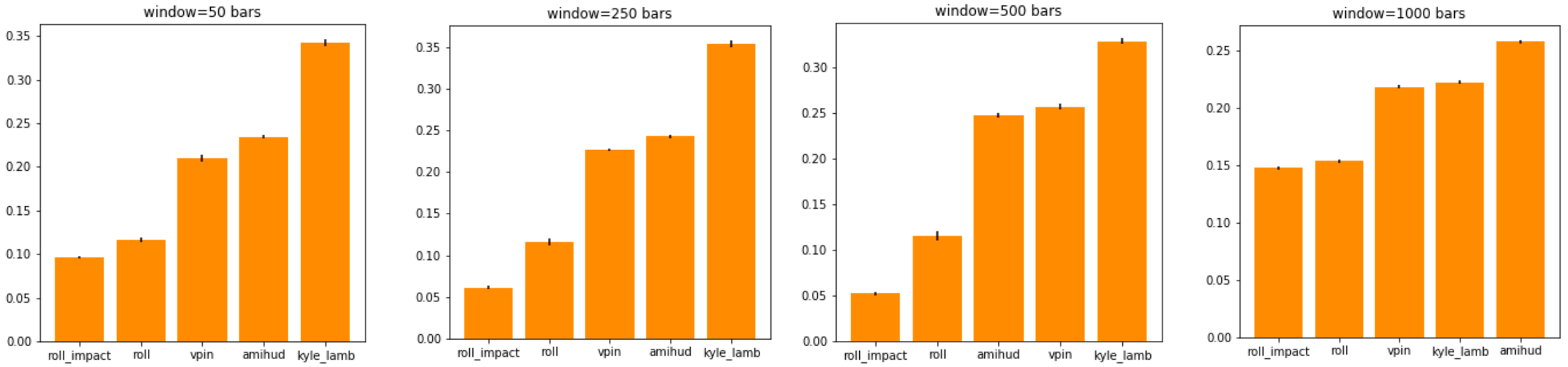


Increasing backward window size for feature generation

Microstructure variable feature importance

2. Sign of change in realized volatility

MDI result

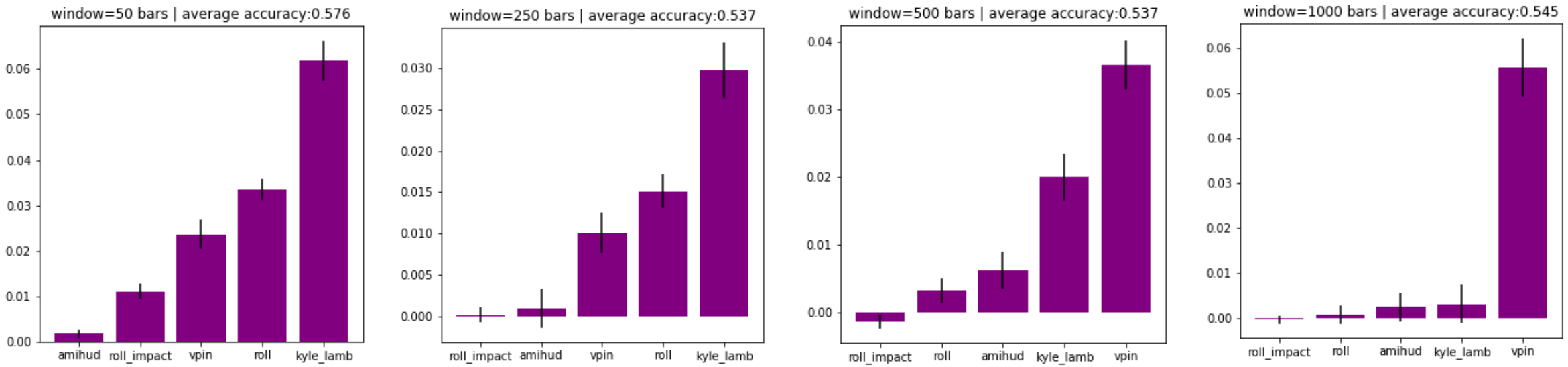


Increasing backward window size for feature generation

Microstructure variable feature importance

2. Sign of change in realized volatility

MDA result

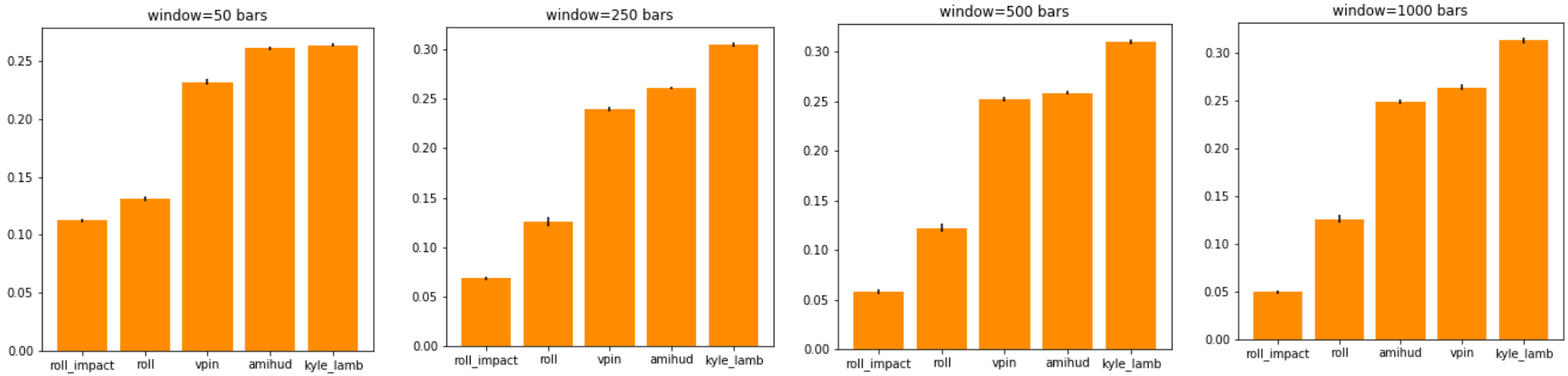


Increasing backward window size for feature generation

Microstructure variable feature importance

3. Sign of change in Jarque-Bera statistics of realized returns

MDI result

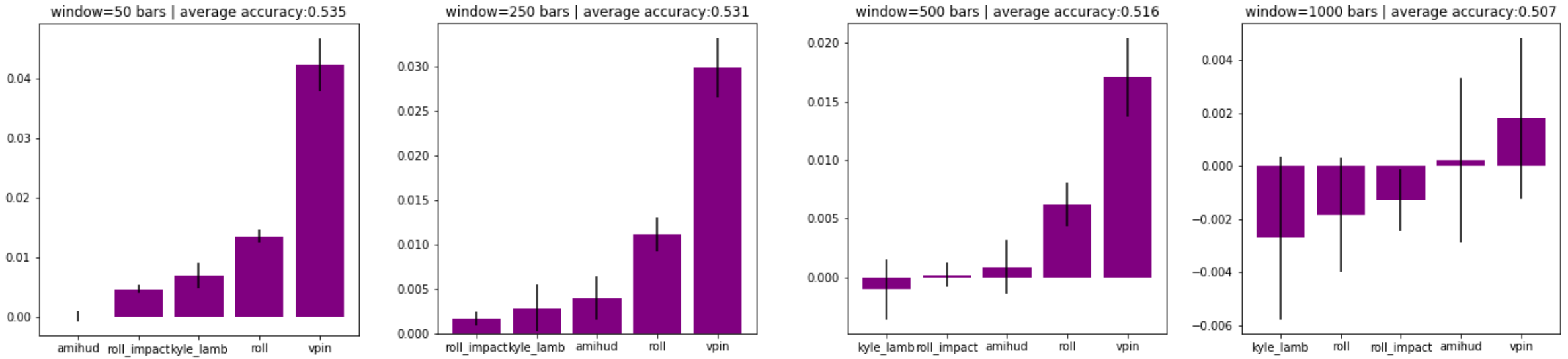


Increasing backward window size for feature generation

Microstructure variable feature importance

3. Sign of change in Jarque-Bera statistics of realized returns

MDA result

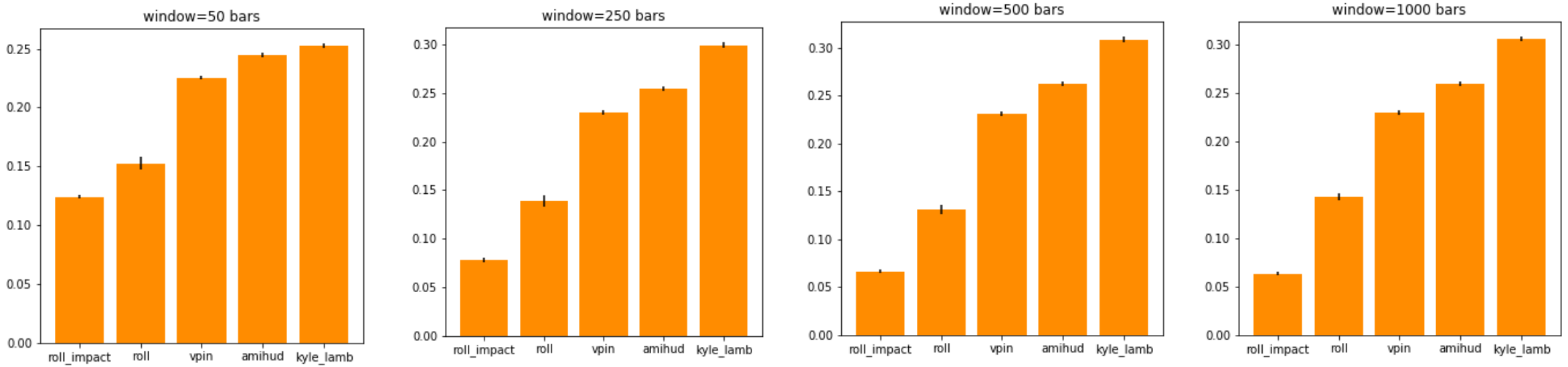


Increasing backward window size for feature generation

Microstructure variable feature importance

4. Sign of change in serial correlation of realized returns

MDI result

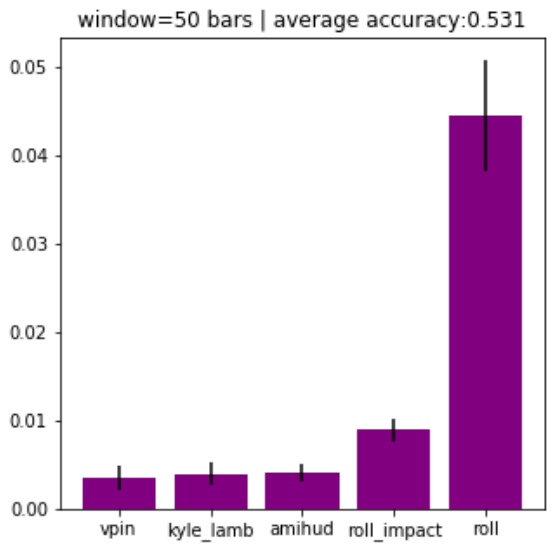


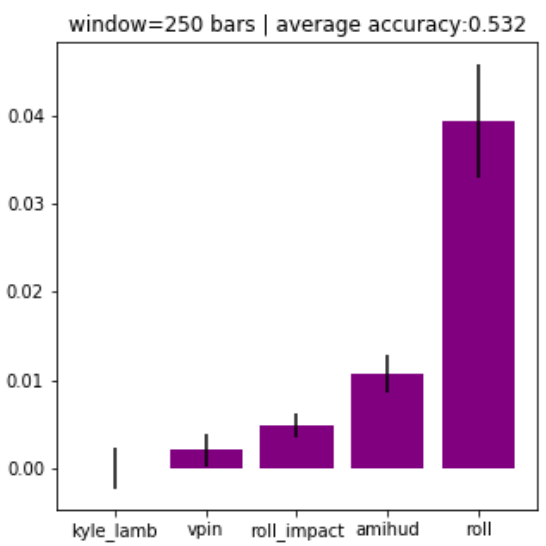
Increasing backward window size for feature generation

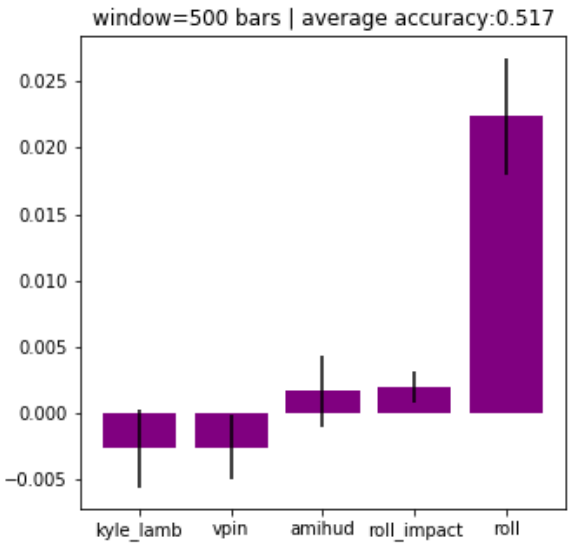
Microstructure variable feature importance

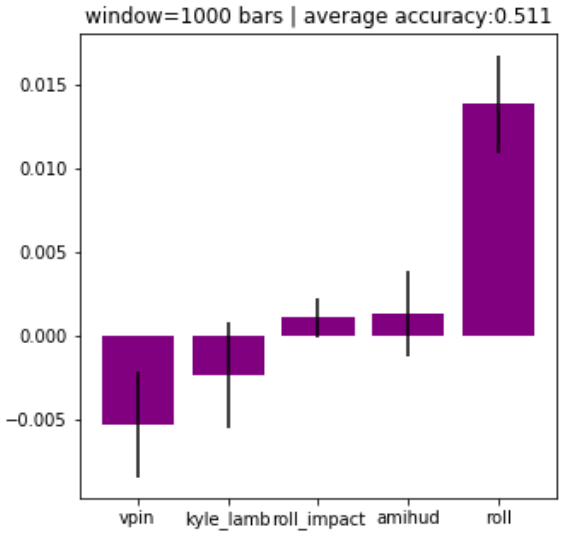
4. Sign of change in serial correlation of realized returns

MDA result







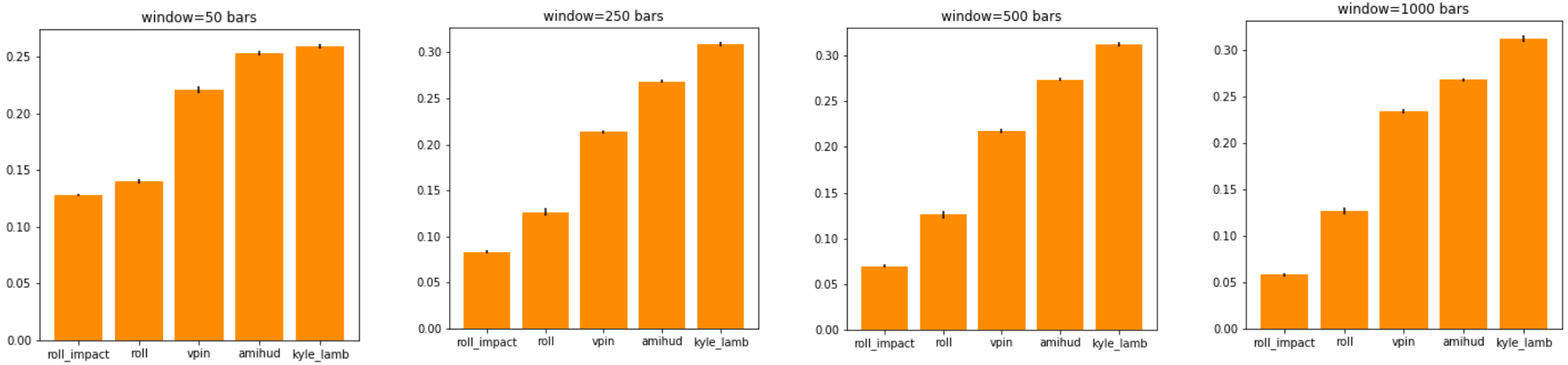


Increasing backward window size for feature generation

Microstructure variable feature importance

5. Sign of change in absolute skewness of realized returns

MDI result

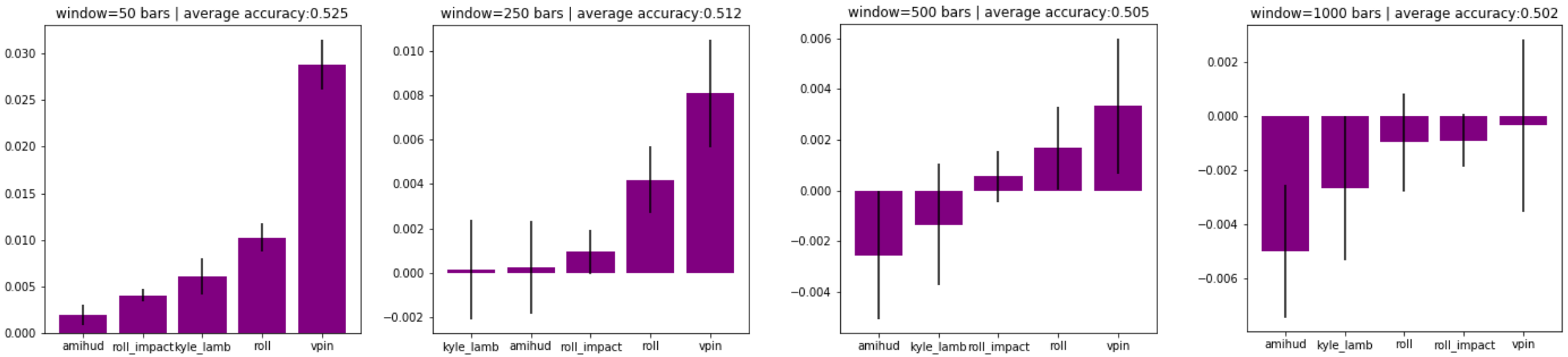


Increasing backward window size for feature generation

Microstructure variable feature importance

5. Sign of change in absolute skewness of realized returns

MDA result

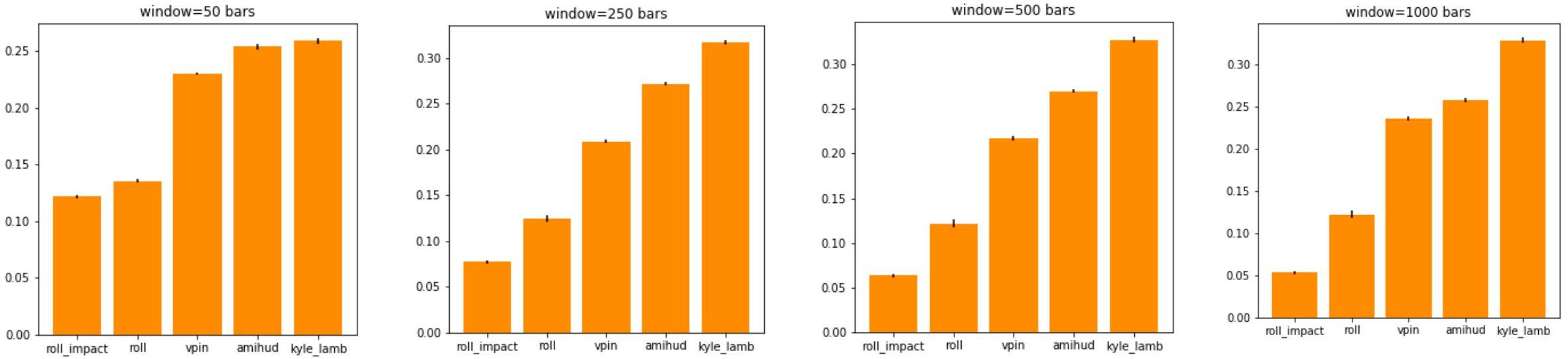


Increasing backward window size for feature generation

Microstructure variable feature importance

6. Sign of change in kurtosis of realized returns

MDI result

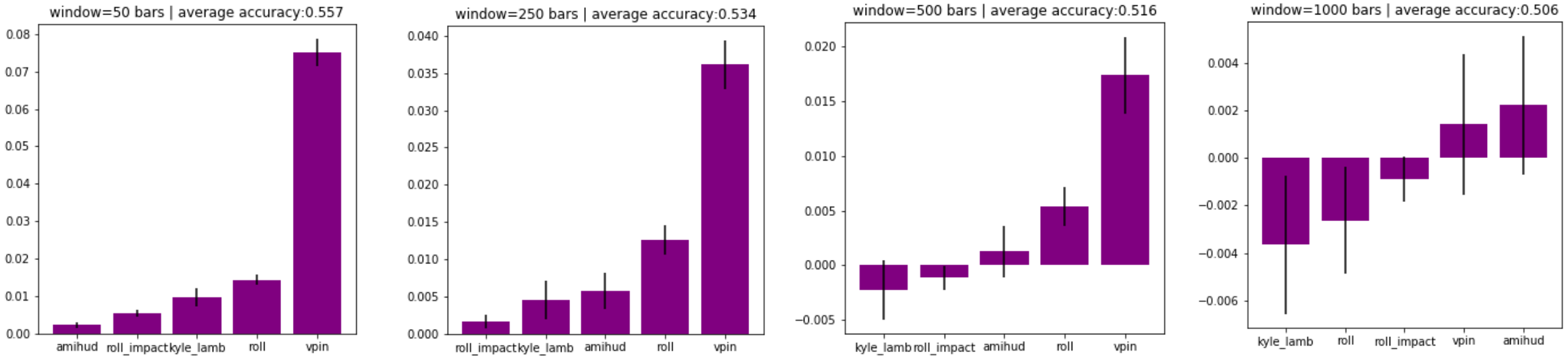


Increasing backward window size for feature generation

Microstructure variable feature importance

6. Sign of change in kurtosis of realized returns

MDA result

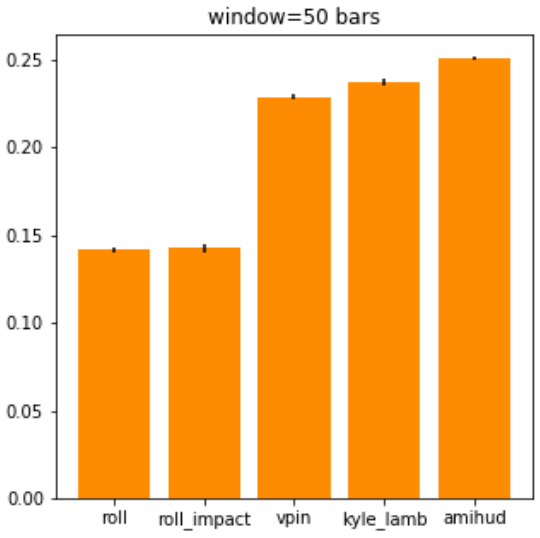


Increasing backward window size for feature generation

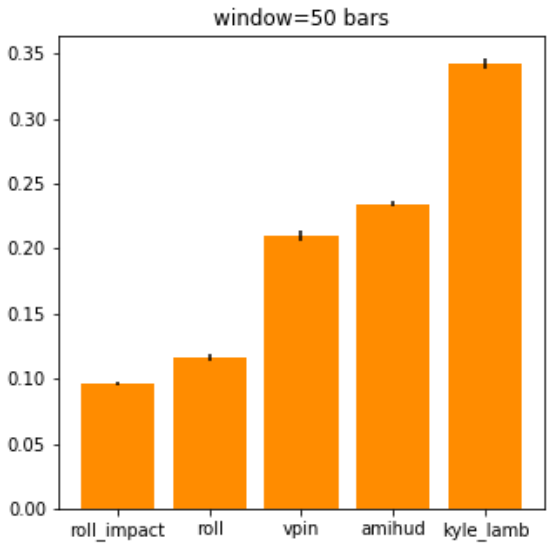
Section IV
Analysis

Kyle & Amihud are best In-Sample (1/2)

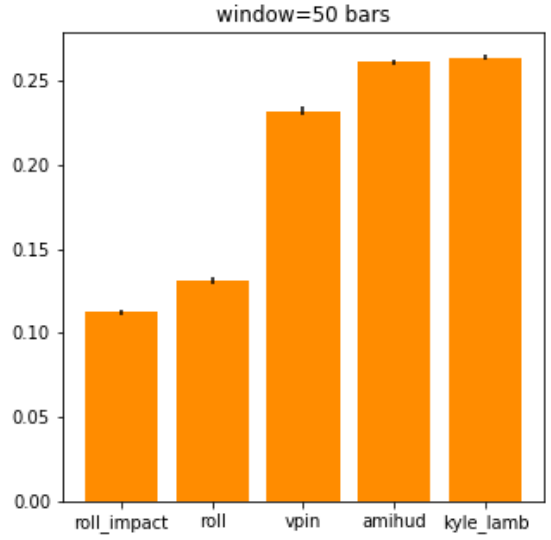
- MDI results have strong similarity across all labels. It is observed that MDI is biased towards features with higher variance (Altmann et al. [2010]). See below for MDI results with 50 bars window.



Corwin-Schultz



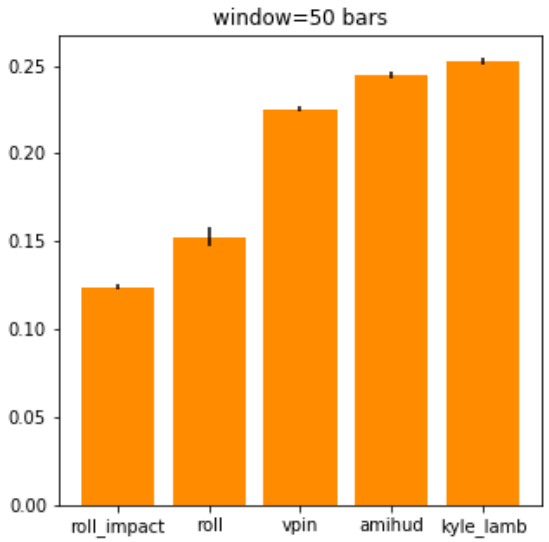
Realized volatility



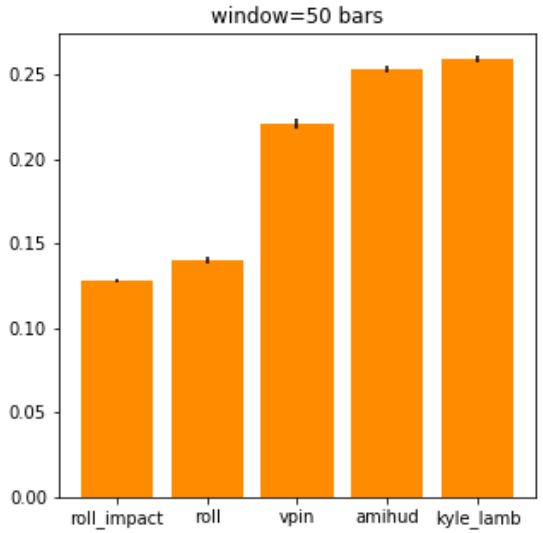
JB statistics

Kyle & Amihud are best In-Sample (2/2)

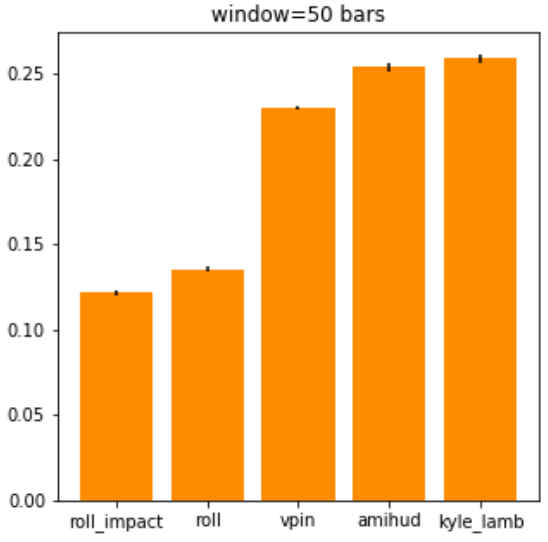
- MDI results have strong similarity across all labels. It is observed that MDI is biased towards features with higher variance (Altmann et al. [2010]). See below for MDI results with 50 bars window.



Sequential correlation



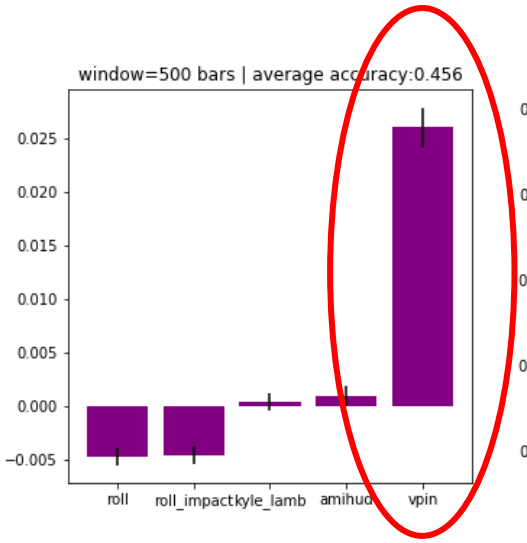
Return skewness



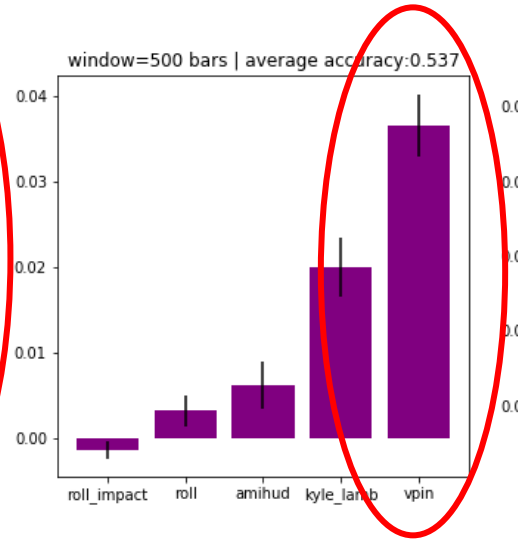
Return kurtosis

VPIN is best Out-Of-Sample (1/2)

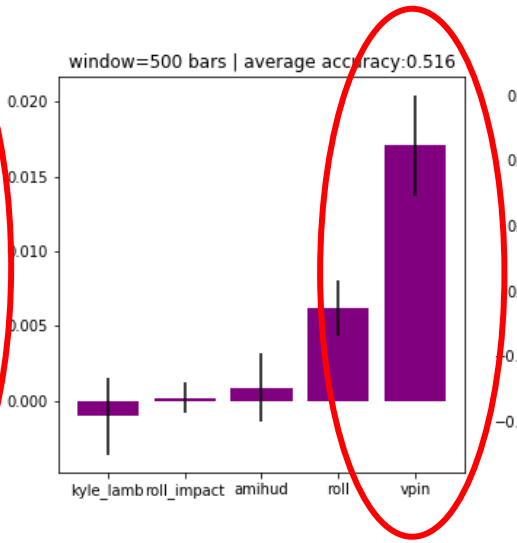
- When the backward window is large, only VPIN can contribute positively to out-of-sample prediction across all labels except sequential correlation. Below are MDA result with 500 bar window



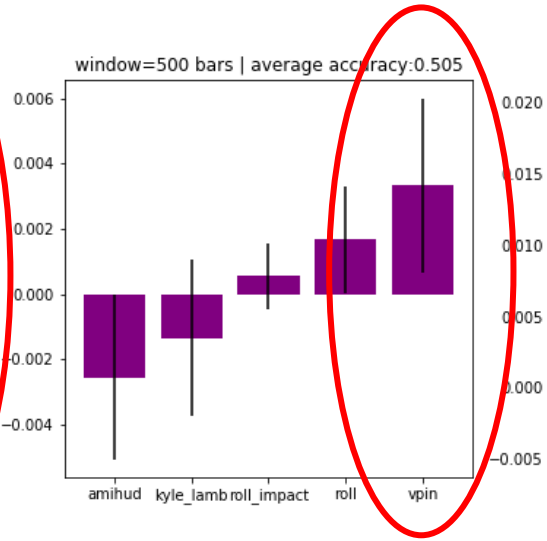
Corwin-Schultz



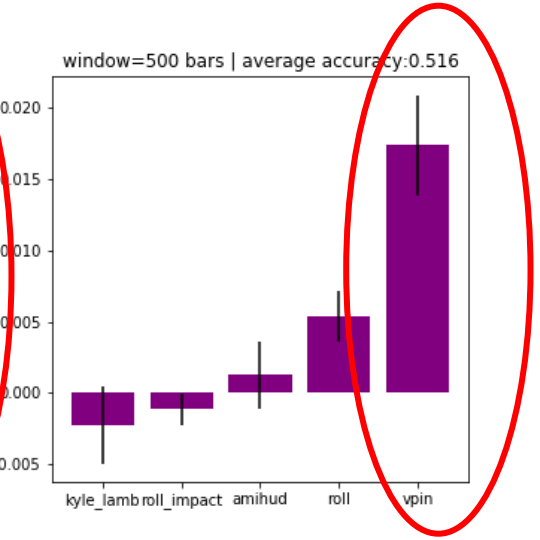
Realized volatility



JB statistics



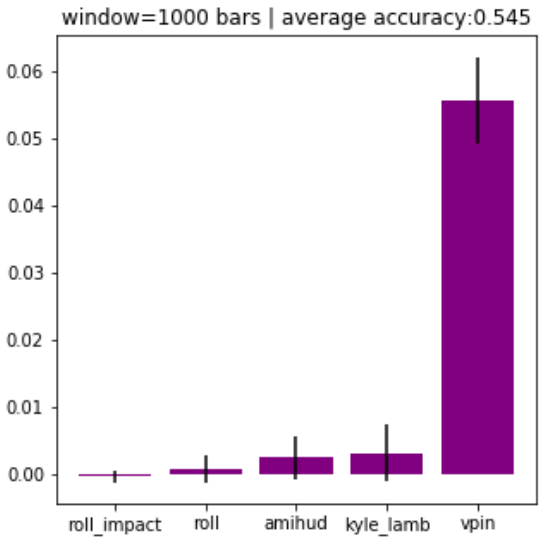
Return skewness



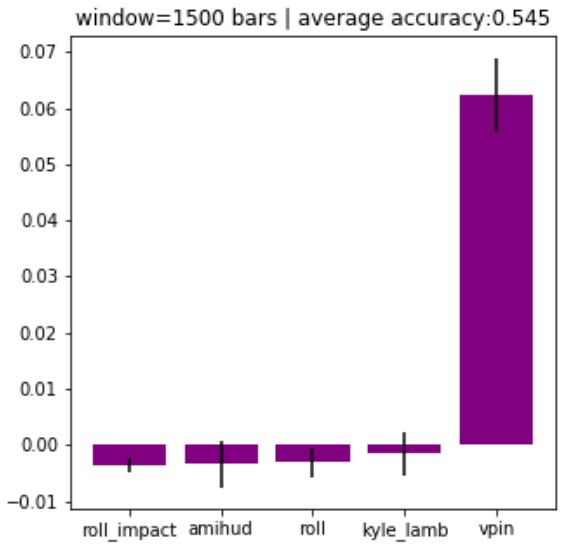
Return kurtosis

VPIN is best Out-Of-Sample (2/2)

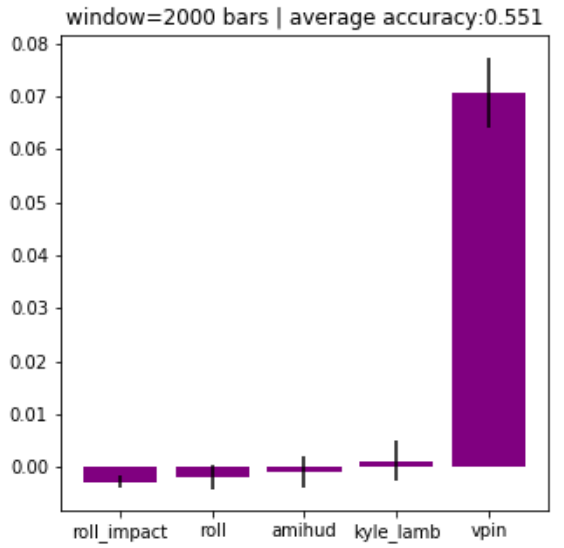
- VPIN's MDA importance at predicting realized volatility remains significant with large window size, even when other variables become irrelevant



1000 bars



1500 bars



2000 bars

Section V
Conclusions

Conclusions (1/3)

- For the prediction of the estimated bid-ask spread, the feature importance remains almost the same across all window sizes, for both MDI and MDA, indicating universality.

- For the prediction of the realized volatility, while Amihud and Roll measures remain stable, VPIN's importance increases while Kyle' lambda decreases as the window size expands, indicating the growth of predictability of VPIN with larger look back window size.

- For the prediction of the JB test the result is similar to realized volatility. VPIN's importance increases while Amihud measure decreases as the window size expands, while the others remain almost the same, indicating the growth of predictability of VPIN with larger look back window size.

Conclusions (2/3)

- For the prediction of the sequential correlation, MDA results demonstrate the Roll measure is much more predictive than all other variables, corresponding to the fact that it is built on past sequential correlation of returns.

- For the prediction of various moments of realized return (volatility, skewness and kurtosis), MDA feature importance shows that VPIN gives the largest contribution consistently, indicating universality.

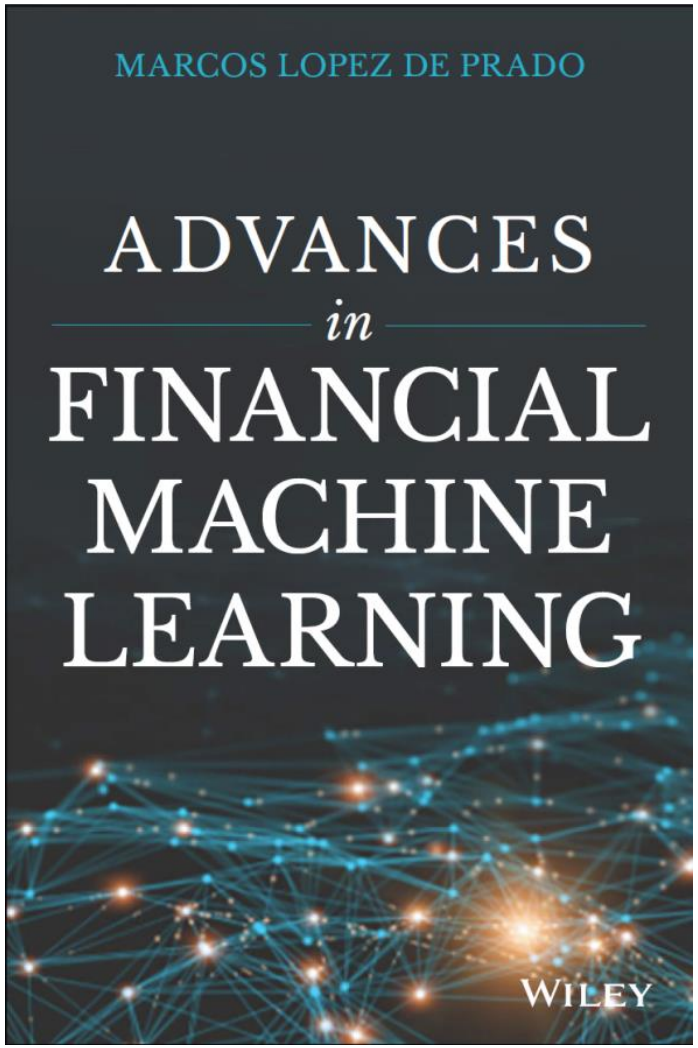
Conclusions (3/3)

- Technologies at the Age of AI provide new methods and perspectives for market microstructure study.
- ML offers new prediction framework with classification on categorical output, compared to traditional regression-based models.
- The 5 prototypical market microstructure variables (Roll Measure, Roll Impact, Kyle's Lambda, Amihud's Lambda and VPIN) show different importances in-sample and out-of-sample. This demonstrates explanatory power does not fully relate to predictability.
- For all prediction labels tested, VPIN is shown to have consistently high importance.

Future directions

- Incorporate more market indicators as features (e.g. VIX)
- Experiment with different bar formations (Volume/Dollar Imbalance Bars)
- Vary forecast horizon and test other prediction labels

For Additional Details



How does one make sense of todays' financial markets in which complex algorithms route orders, financial data is voluminous, and trading speeds are measured in nanoseconds? For academics and practitioners alike, this book fills an important gap in our understanding of investment management in the machine age.
— Prof. **Maureen O'Hara**, Cornell University. Former President of the American Finance Association.

The first wave of quantitative innovation in finance was led by Markowitz optimization. Machine Learning is the second wave and it will touch every aspect of finance. López de Prado's Advances in Financial Machine Learning is essential for readers who want to be ahead of the technology rather than being replaced by it.
— Prof. **Campbell Harvey**, Duke University. Former President of the American Finance Association.

Disclaimer

- The views expressed in this document are the authors' and do not necessarily reflect those of the organizations he is affiliated with.
- No investment decision or particular course of action is recommended by this presentation.
-