# Forecasting high-frequency stock market returns using embedded limit order book data

Marius Sterling
Niels Wesselhöfft

International Research Training Group 1792
Ladislaus von Bortkiewicz Chair of Statistics
Humboldt–Universität zu Berlin

http://irtg1792.hu-berlin.de
http://lvb.wiwi.hu-berlin.de

# Motivation

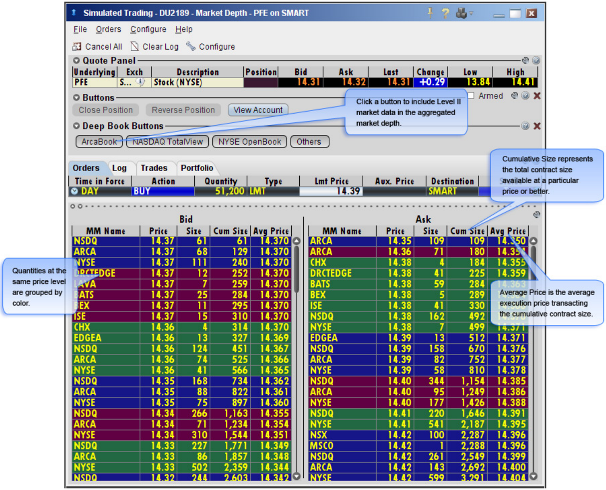



Figure: Limit Order Book (LOB) Level 2

# Data Description

- Time span and aggregation
  - Time span from 01/07/2015 to 31/12/2015 (first iteration)
  - Tick data aggregated to second (last tick)
- Source
  - LOB tick data for NASDAQ stocks including Amazon from Lobster
  - Trading from 9:30 a.m. to 04:00 p.m. representing 23,400 seconds per day, 3m for half a year

# Raw Data

- Limit prices and volumes (Bid | Ask) over
  - time $t = 1, \ldots, T$,
  - depth $d = 1, \ldots, D$ with $D = 200$
- Limit ask and bid price $P^a_{t,d}$, $P^b_{t,d}$
- Volume $V^a_{t,d}$, $V^b_{t,d}$
- Log price $p^a_{t,d} = \log P^a_{t,d}$

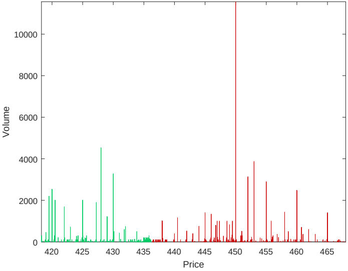



Figure: Limit Order Book in $t$

# Cumulated Volumes

- Limit price vector and volume vector

$$P_t^a = \left[P_{t,1}^a, \ldots, P_{t,D}^a\right]^\top$$
$$V_t^a = \left[V_{t,1}^a, \ldots, V_{t,D}^a\right]^\top$$

- Cumulated Limit prices and volumes (Bid | Ask)

$$\overline{V}_{t,d}^b = \sum_{i=1}^{d} V_{t,i}^b$$

$$\overline{V}_{t,d}^a = \sum_{i=1}^{d} V_{t,i}^a$$

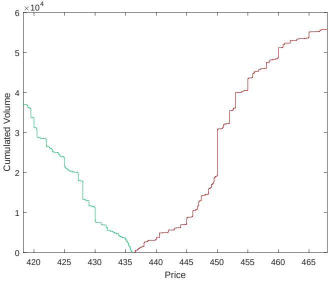



Figure: Limit Order Book with cumulated volumes in $t$

# LOB over time

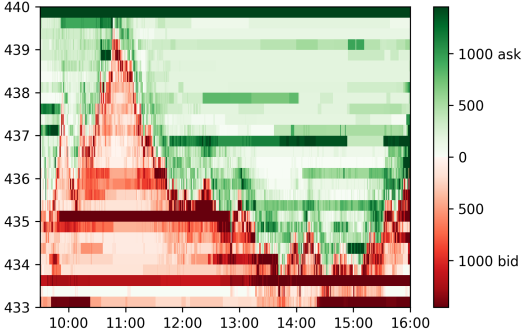
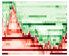



Figure: Limit Order Book (LOB) over one day



# Embedding

- Embeddings for LOB data
  - Price and volume data
  - Return and volume data
  - Factor models
    - Semiparametric Factor Model (Hautsch, Härdle, Mihoci, 2012)
  - Measure for buying/selling pressure

# Measure for buying/selling pressure

- Absolute difference from mid price

$$\delta^b_{t,d} = |P^b_{t,d} - P_t| \quad (1)$$

$$\delta^a_{t,d} = |P^a_{t,d} - P_t| \quad (2)$$

- Mid price $P_t \in \left[P^b_{t,1}, P^a_{t,1}\right]$

$$P_t = \sqrt{P^b_{t,1} P^a_{t,1}} \quad (3)$$

$$P_t = \frac{1}{2}\left(P^b_{t,1} + P^a_{t,1}\right) \quad (4)$$

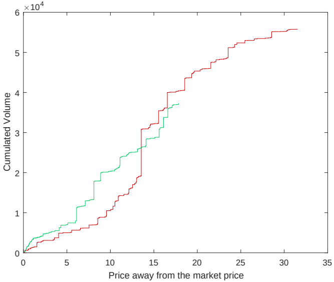



Figure: Cumulated volume over difference from mid price $P_t$ in $t$

# Theta $\theta_t$

- Theta $\theta_t \in \mathbb{R}$

$$\theta_t = f(\overline{V}_{t,d}, \delta_{t,d}) \qquad (5)$$

- Theta - difference

$$\overline{\theta}_{t,d} = \overline{V}^a_{t,d} - \overline{V}^b_{t,d} \qquad (6)$$

- $d \in \{1, \ldots, D\}$ is the specific depth chosen

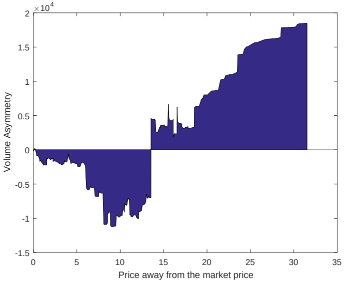



Figure: Cumulated volume difference $\overline{\theta}_{t,d}$

# Theta $\theta_t$

- Theta - ratio

$$\widetilde{\theta}_{t,d} = \begin{cases} \widetilde{\theta}^{+}_{t,d} = \overline{V}^{b}_{t,d} / \overline{V}^{a}_{t,d}, \ \overline{V}^{b}_{t,d} > \overline{V}^{a}_{t,d} \\ \widetilde{\theta}^{-}_{t,d} = -\overline{V}^{a}_{t,d} / \overline{V}^{b}_{t,d}, \ \overline{V}^{b}_{t,d} \leq \overline{V}^{a}_{t,d} \end{cases} \tag{7}$$

- Interpretation
  - $\widetilde{\theta}^{+}_{t,d}$ as multiple of bid over ask cum volume (buying pressure)
  - $\widetilde{\theta}^{-}_{t,d}$ as multiple of ask over bid cum volume (selling pressure)
- Depth weights
  - Uniform
  - Exponential

# Return forecasting

- Return forecast up to 600 seconds

$$r_{t+i} = f(X_t, X_{t-1}, \ldots) + \varepsilon_{t+i}, \tag{8}$$

$$\varepsilon_{t+i} \sim \mathrm{F}(), \ \ i = 1, \ldots, 600 \tag{9}$$

- Dependent variable: Log return

$$r_t = \log P_t - \log P_{t-1} = p_t - p_{t-1} \tag{10}$$

- Independent variable(s) $X_t, X_{t-1}, \ldots$

$$X_t = \begin{cases} 1) \ [P^b_{t,1}, V^b_{t,1}, P^a_{t,1}, V^a_{t,1}, \ldots, P^b_{t,D}, V^b_{t,D}, P^a_{t,D}, V^a_{t,D}] \\ 2) \ [r^b_{t,1}, V^b_{t,1}, r^a_{t,1}, V^a_{t,1}, \ldots, r^b_{t,D}, V^b_{t,D}, r^a_{t,D}, V^a_{t,D}] \\ 3) \ \theta_t = f(\overline{V}_{t,d}, \delta_{t,d}) \end{cases} \tag{11}$$

# Modeling Approaches

| **Embedding**<br>**Model** | 1 Prices and volumes | 2 Log returns and volumes | 3 Asymmetry estimator $\theta$ |
|---|---|---|---|
| 1 Linear Regression | 11 | 12 | 13 |
| 2 RNN (e.g. LSTM) | 21 | 22 | 23 |
| 3 TCN | 31 | 32 | 33 |

Table: Modeling matrix

# 13 Linear Regression $+ \theta_t$

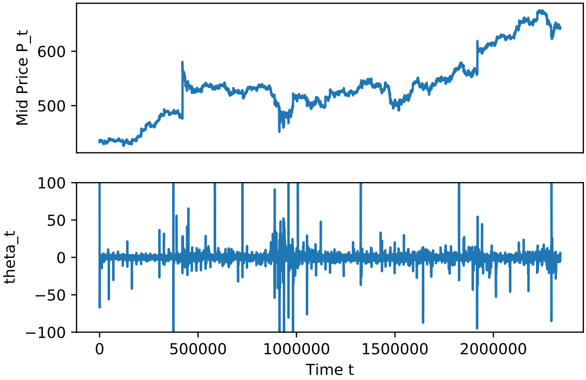



Figure: (upper) Amazon prices series (lower) $\theta_t$ over time

# 13 Linear Regression $+ \theta_t$

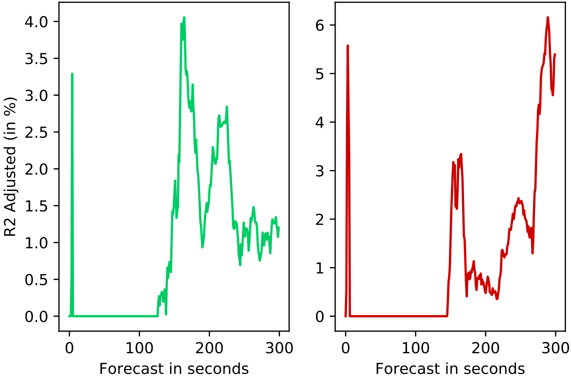
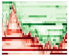



Figure: (left) $R^2$ for $X_t = [\theta_t^+]$ (right) $R^2$ for $X_t = [\theta_t^-]$



# 23 Linear Regression $+\theta_t + \theta_t^2$

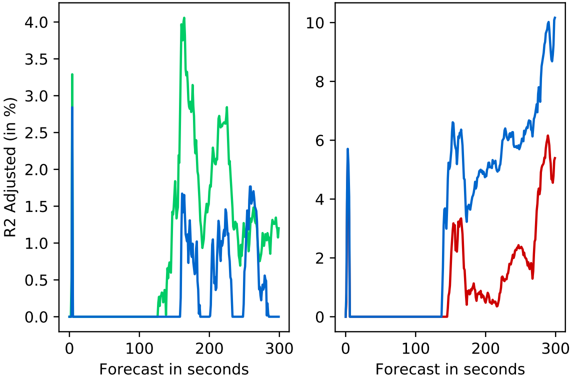
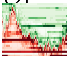


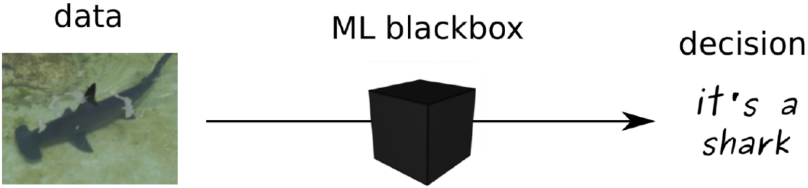

Figure: (left) $R^2$ for $X_t = [\theta_t^+, (\theta_t^+)^2]$ (right) $R^2$ for $X_t = [\theta_t^-, (\theta_t^-)^2]$



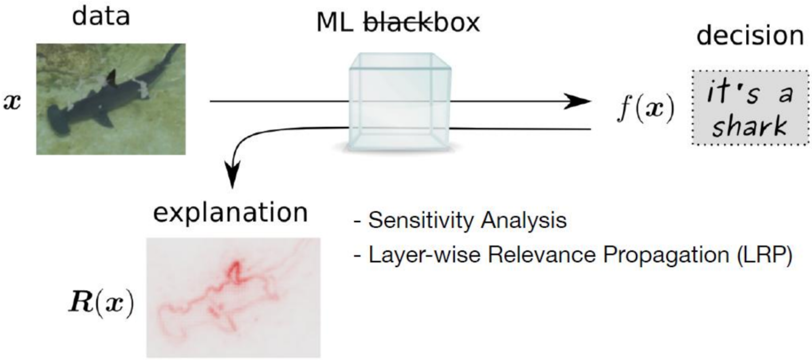
# TCN inspired model architecture

- Layers
  - Feature generating layer
  - Inception layers
  - Dimension reduction layers
- Training parameters
  - Depth: 10
  - Lag: 64 seconds
  - Prediction horizon: 15 seconds
  - Loss: MSE
  - Optimizer: Adam
  - Batchsize: 128

# NNs as a blackbox



# LRP




- Sensitivity Analysis
- Layer-wise Relevance Propagation (LRP)

# Input relevance for 15 second forecast

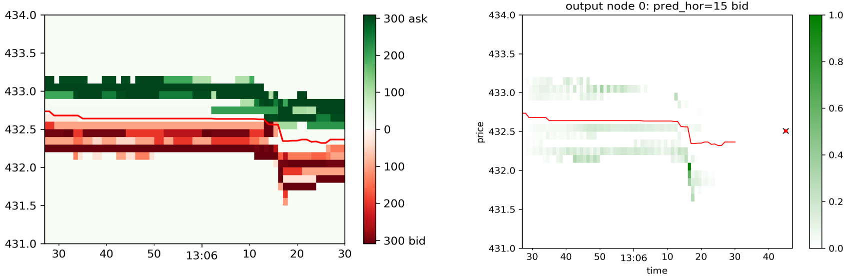



Figure: (left) LOB input and (right) scaled forecast relevance (LRP) for input of 15 sec ahead bid price prediction, — mid price, • price, × predicted price

# Input relevance for 15 second forecast

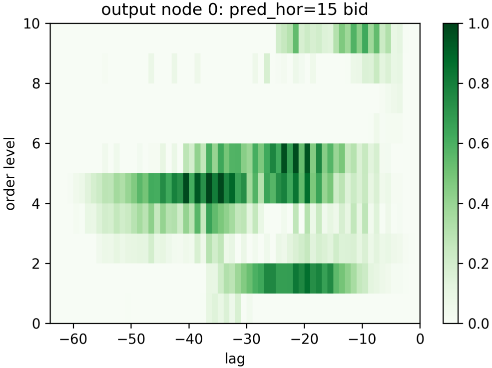



Figure: Dept and Lag relevance

# Research goals

- Supervised methods
  - Stock markets returns forecasts for high frequencies
- Embeddings
  - Will unstructured information still lead to superior NNs?
  - Compare parametric depth structure (uniform, exponential) to the forecast relevance of NNs (LRP)
- Out-of-sample Test
  - Performance holds under transactions costs
  - Out-of-sample and out-of-stock