Group 1

Algorithmic Trading

Generative Models for LOB Data

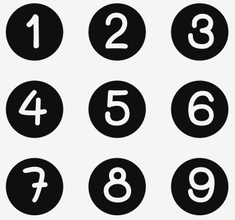
Content

Why Simulate Data?

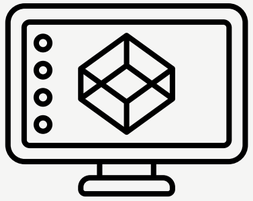
- Not Enough Data - Real LOB data is hard to access due to cost/privacy
- Reactivity - Historical Data won’t react to new asks
- Build Scenarios
- Useful for testing trading algorithms, market resilience, and modeling rare events

Generative Models

Generative models aim to learn the distribution of data and sample from it



Inspiration from text generation (e.g., GPT) but adapted to financial order data



Can include simple probabilistic models or more complex sequence models

Our Approach

- Use LOBSTER dataset to learn distributions of order flow and book dynamics.
- Extract mid-price, spread, imbalance.
- Model order arrival, side, size, and limit order prices separately.
- Simulate a full day of orders and compare statistics with real data.

Binning

Discretizing time for money

Time is split into 0.001-second bins.

Each bin records whether a market/limit order arrives.

Enables binary classification models and autocorrelation analysis.

Order Arrival Model

- Modeled market and limit orders separately.

- Simulated order types using fixed probabilities estimated from the data.
  - Assumes independence from features like spread or imbalance.
  - Simple Bernoulli sampling with constant probability.

- Used logistic regression to predict the probability of an order being MO or LO based on Spread and Imbalance.
  - Produces context-dependent probabilities, i.e., the probability changes depending on order characteristics.

- Also tried RNN’s (Logistic Regression gave the most realistic result)

Model Side and Size

How did we model buy/sell direction and size?

Filtered the dataset into market and limit orders, creating separate groups for analyzing their directional behavior.

Computed the probability of buy-side orders (Direction == 1) within each time interval (TimeBin) to capture how order direction varies over time.

Generated time-dependent Bernoulli models that can be used to simulate whether a new order is buy or sell based on its type and timestamp.

Model Size

Lognormal Distribution

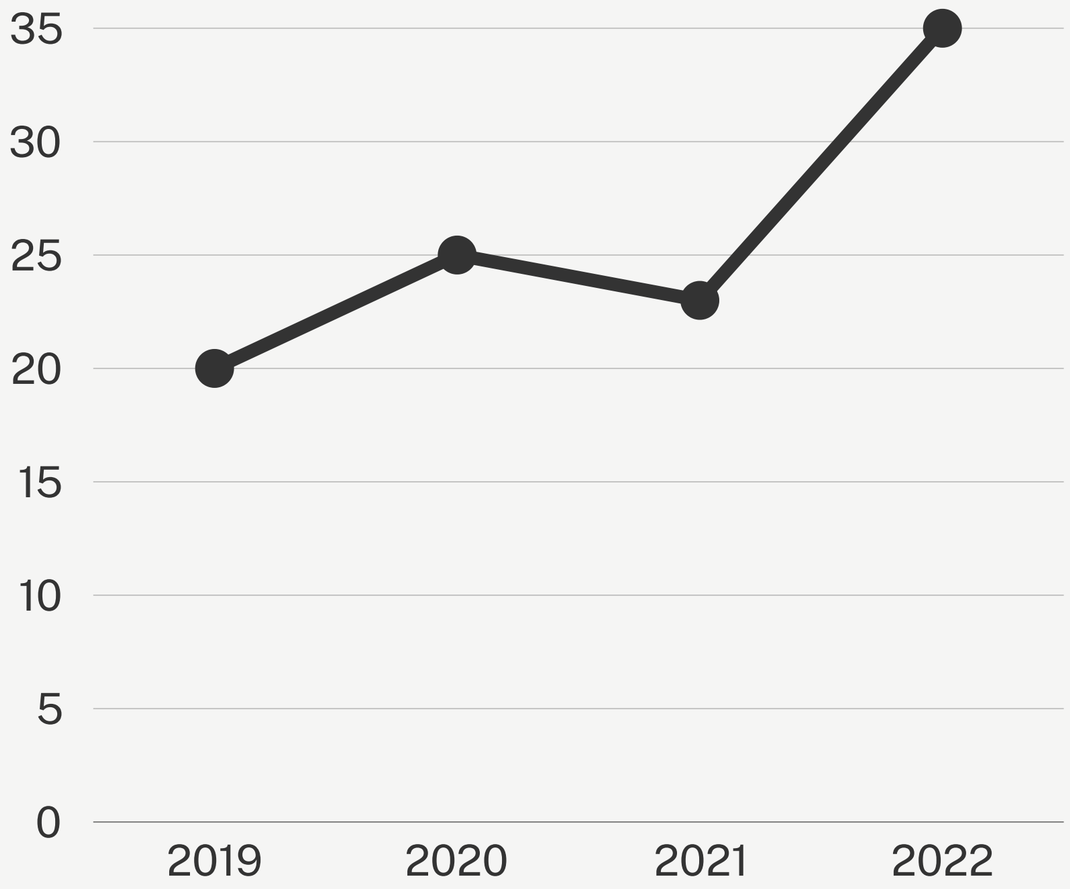
- Fitted power law distributions to order sizes for each group (buy/sell × market/limit) to model their heavy-tailed behavior.
- Defined a sampling function that retrieves the appropriate distribution parameters based on order type and direction.
- Generated and rounded sampled sizes to the nearest 100 units, enforcing a minimum round-lot size of 100.

Order Types

- Limit Buy
- Limit Sell
- Market Buy
- Market Sell

Mid-Price Modeling

- Fitted a lognormal distribution to the absolute distance between limit order prices and the mid-price to model how far from the mid-price LOs are typically placed.
- Defined a function to simulate limit order prices, adjusting the mid-price up or down based on direction and a sampled distance.
- Modeled mid-price movement as a random walk, fitting a lognormal distribution to multiplicative returns and simulating price paths using those returns.



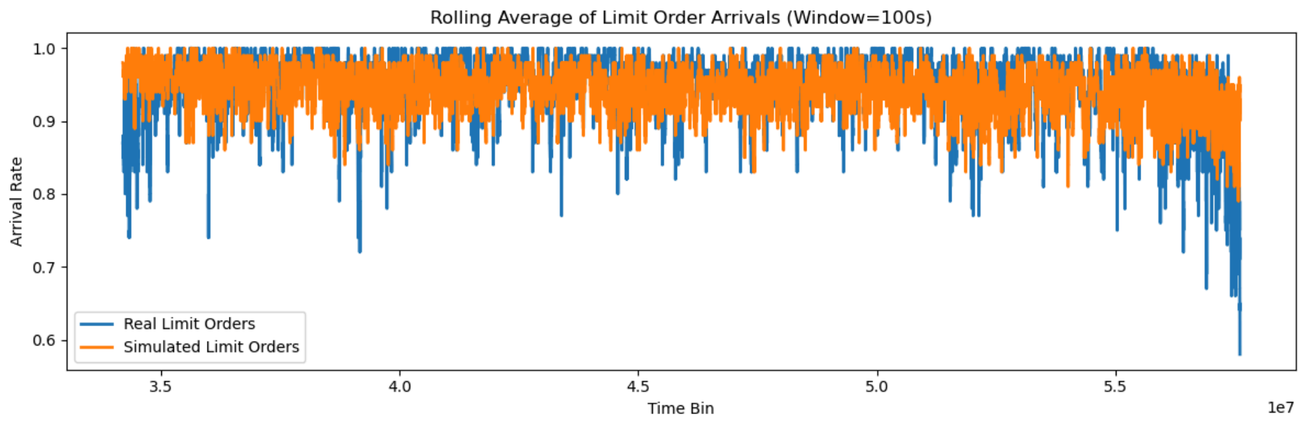
Synthetic Order Generation

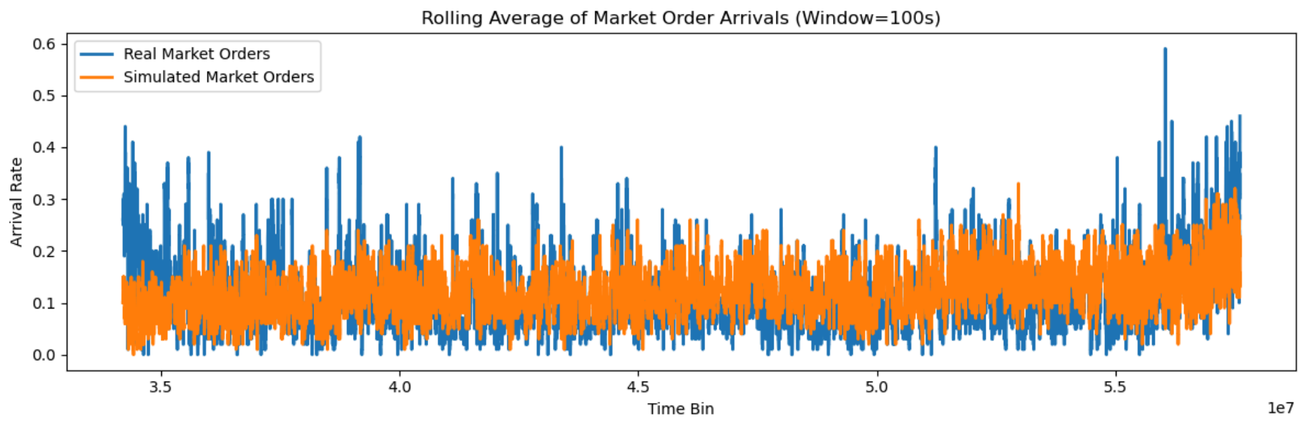
- For each time bin: sample arrival, type, side, size, and (if LO) price.
- Used statistical models estimated from LOBSTER data.
- Ensured structure mimics original message file format.

	Time	Type	OrderID	Size	Price	Direction
0	34200004	Limit	1	2500	585.840000	1
1	34200025	Limit	2	100	585.490000	-1
2	34200050	Limit	3	100	585.690000	1
3	34200201	Limit	4	400	585.670000	1
4	34200205	Limit	5	1800	585.670000	1
...	...	...	...	...	...	...
168981	57599355	Limit	168982	1400	580.730000	1
168982	57599383	Limit	168983	100	580.770000	1
168983	57599444	Market	168984	2200	580.658579	1
168984	57599444	Limit	168985	100	580.630000	-1
168985	57599913	Limit	168986	100	580.550000	-1

168986 rows × 6 columns

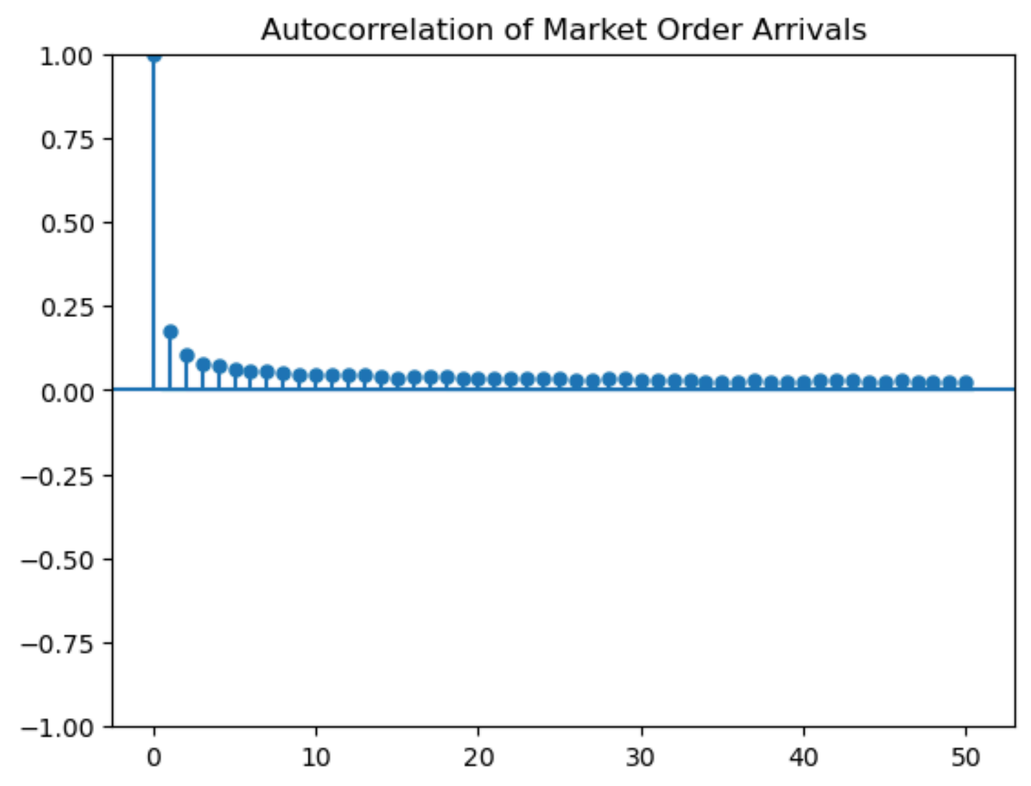
Results - Order Arrivals





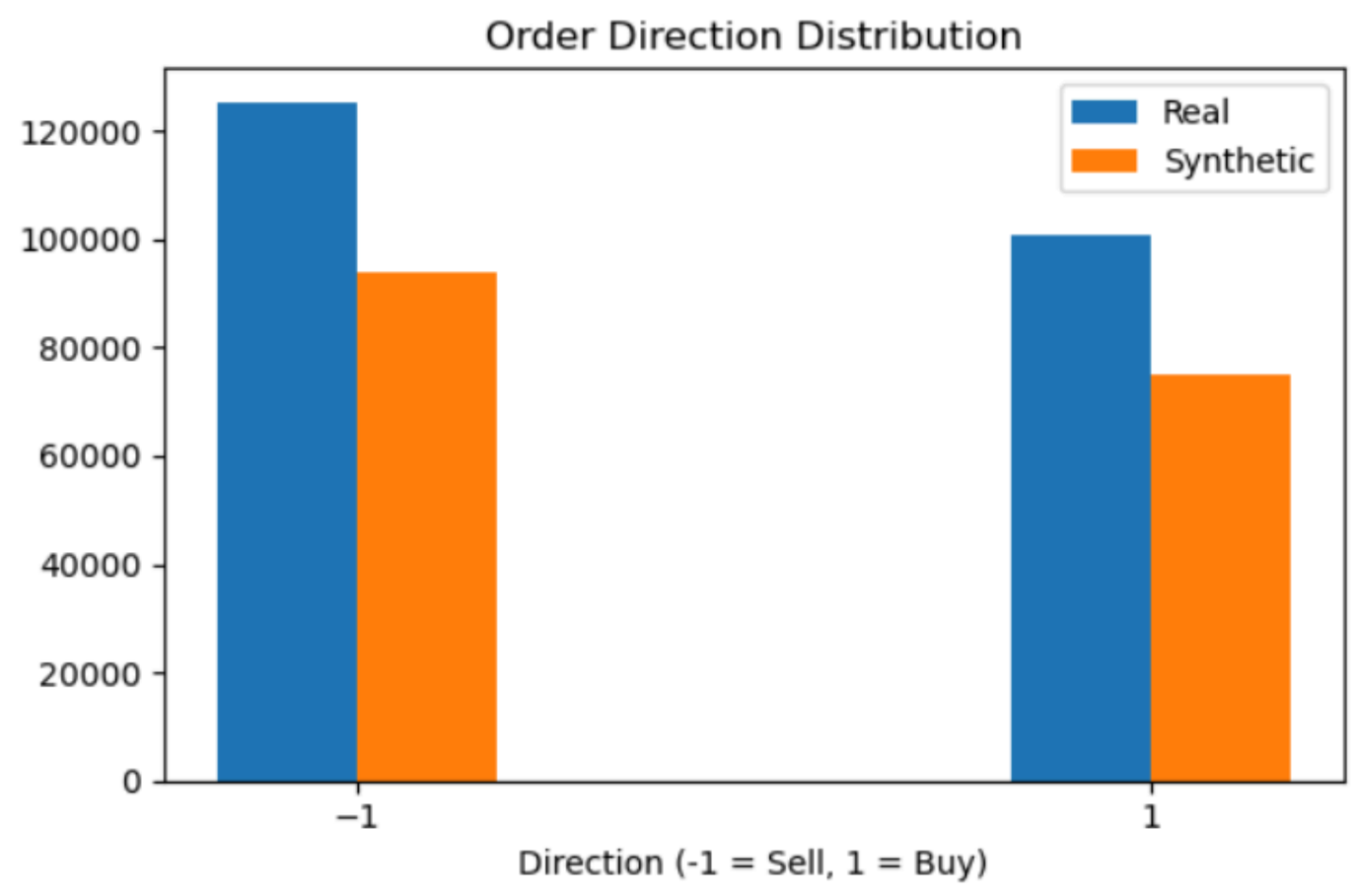
- Simulated arrival rates generally follow the trend of real data, indicating that the model captures the overall time-varying structure of order arrivals.
- Limit order simulation (top plot) closely tracks real arrival rates with smaller deviations, suggesting better model performance for limit orders.
- Market order simulation (bottom plot) shows more fluctuation and divergence from the real data, indicating that market order arrival behavior may be harder to model or less dependent on the chosen features.

Results - ACF

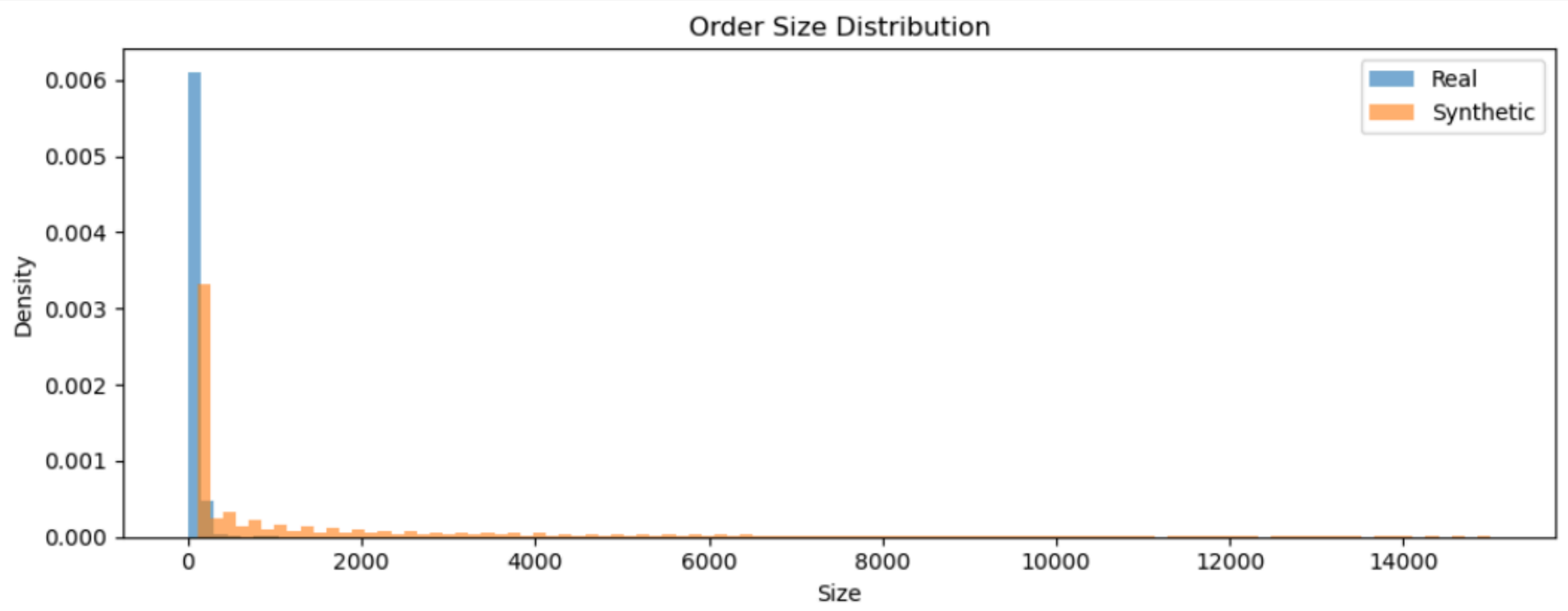


Shows short-term time dependency in Market Order Arrivals

Results - Order Direction

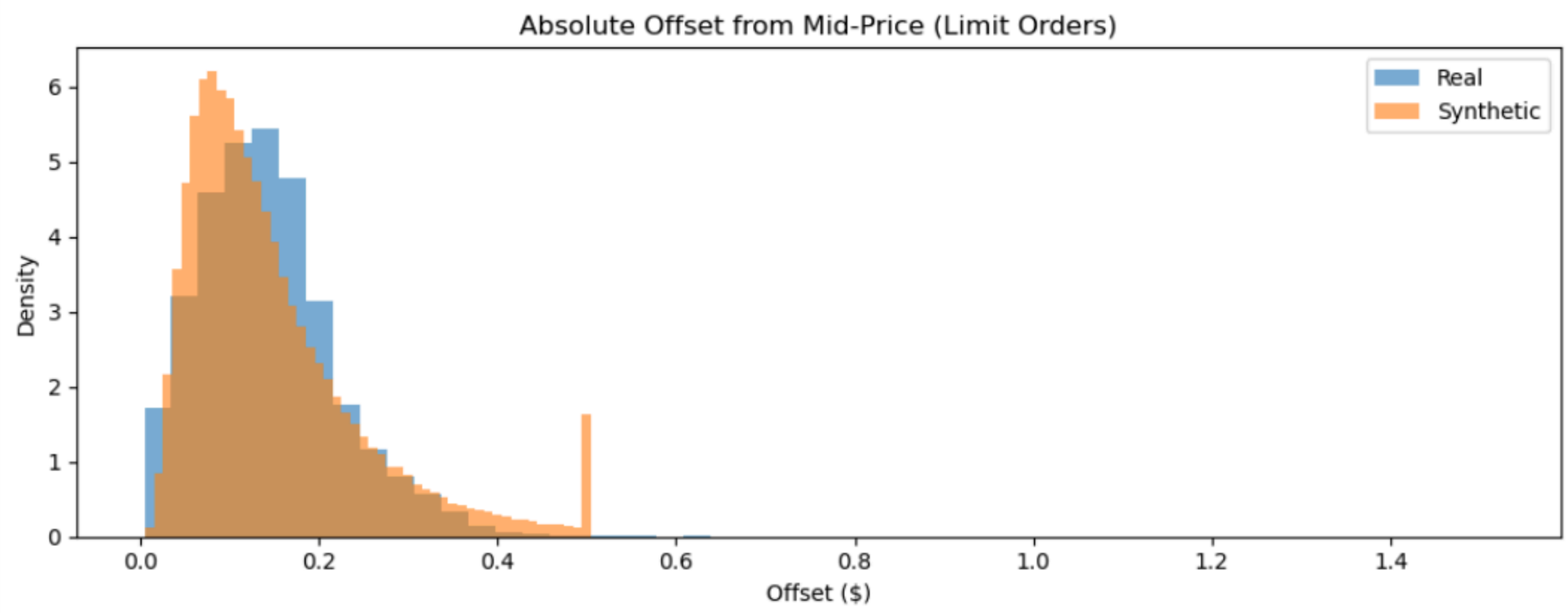


Results - Order Size



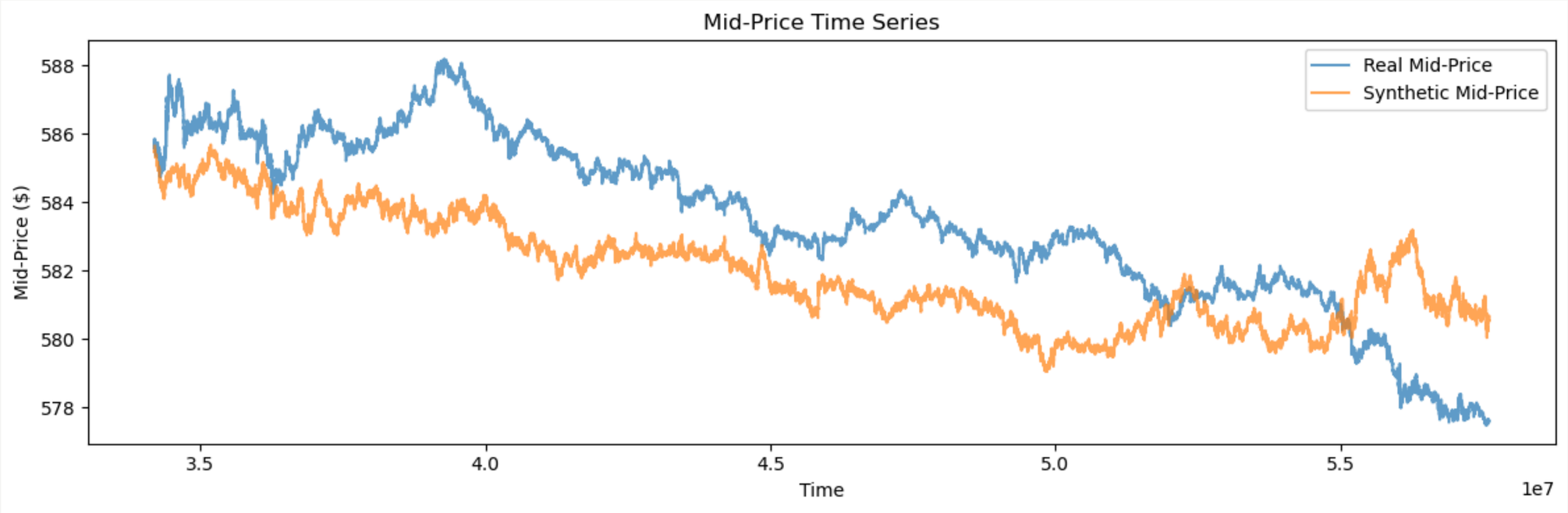
The synthetic order sizes capture the heavy-tailed shape of the real data but underrepresent small sizes, especially around the sharp peak near zero, suggesting that the fitted distribution (Power Law) may not fully capture the high frequency of small-lot orders seen in real trading.

Results - Mid-Price



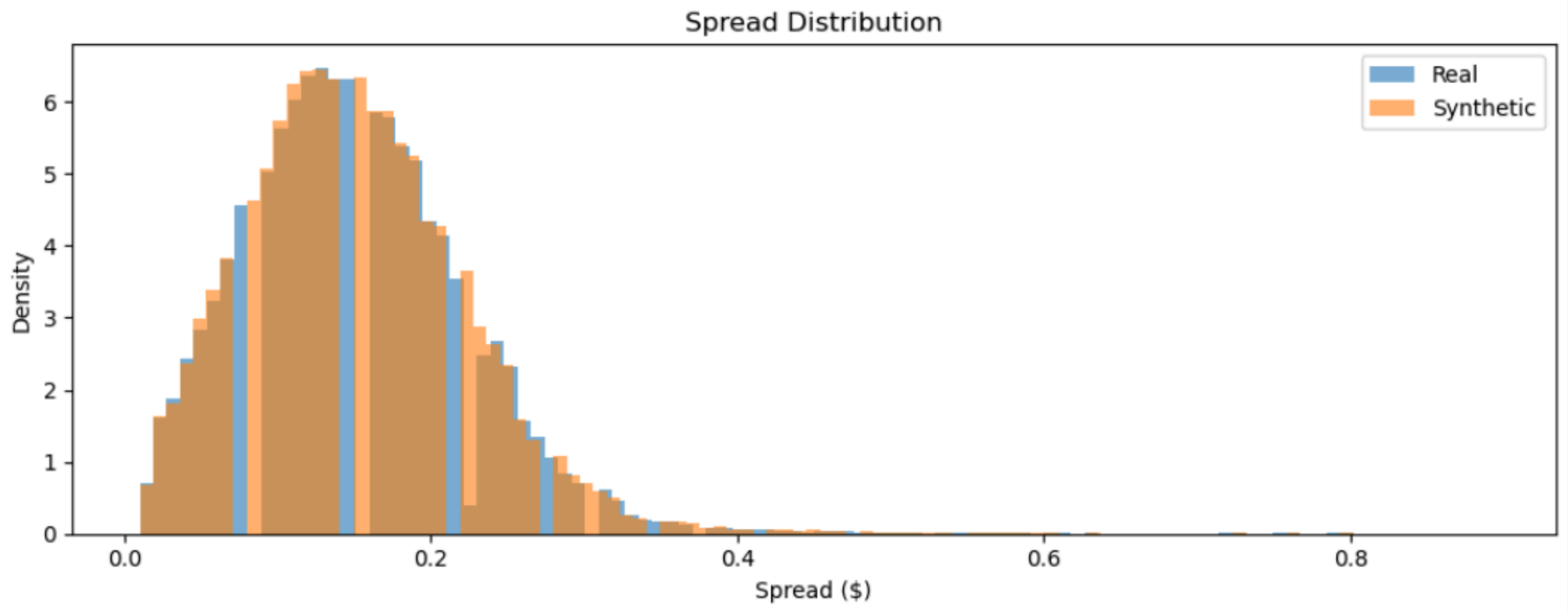
The synthetic offset distribution roughly captures the overall shape of the real distribution but overestimates the frequency of large offsets, especially around the 0.5 mark, likely due limitations in the fitted distribution.

Results - Mid-Price



The synthetic mid-price follows a similar downward trend as the real mid-price but at a lower overall level, suggesting that the random walk model captures general directional behavior but may be miscalibrated in terms of drift or starting point.

Results - Spread



The synthetic spread distribution closely matches the real data, especially around the mode and central region, indicating that the model accurately captures typical spread behavior.

Results - Imbalance

Both real and synthetic imbalance distributions show spikes at specific values, showing how buy/selll pressure is distributed throughout the day. The spike at 0.5 show that bid and ask volumes are equal, representing a balanced orderbook in these occasions.

Thank You