Empirical Analysis of Price Impact Based on Stochastic Liquidity

Yuqian Li
Courant Institute of Mathematical Sciences
New York University
yl12057@nyu.edu

Srujitha Ambati
Courant Institute of Mathematical Sciences
New York University
sa8220@nyu.edu

Rui Yang
Courant Institute of Mathematical Sciences
New York University
ry2424@nyu.edu

Instructor: Prof. Merrell Hora

1 Introduction

As researched by Muhle-Karbe, Johannes et al. 2023 explores a more sophisticated scenario where price impacts are non-linear and concave with respect to trade sizes. This reflects real-world observations, where large trades tend to have diminishing marginal impacts on prices. The authors introduce stochastic liquidity parameters that approximate these nonlinear effects in a high-frequency trading regime. They use the propagator model, which represents how the impact of a trade decays over time, and propose a reduced-form limit that simplifies the nonlinear impact into a more tractable linear model with stochastic features. This enables the use of high-frequency data and analytical techniques that are typically reserved for simpler linear models.

Thus, we conclude that in financial markets, large trades often come with price impacts. By leveraging the impact multiplier λ, investors can optimize execution strategies based on varying market conditions. Intraday analysis improves trade execution and minimizes adverse price effects, providing actionable insights to minimize costs in algorithmic trading.

Our project aims to bridge the gap between empirical and theoretical models to improve algorithmic trading strategies. In addition, we focus on conducting extensive empirical comparisons of various price impact models, exploring intraday patterns, and refining models through Bayesian Regression.

2 Data Preparation

2.1 Notations

Symbol	Description	Formula
Δt	time bin	5-min
I_t	price difference	trade last price at the end minus beginning
F_t	trading volume	$\sum$(buy volume) - $\sum$(sell volume) of the entire time bin
ADV	average daily volume	average daily total trade volume (9:30am - 15:59pm) for whole sample (1Q)
σ	volatility	standard deviation of the daily log return of trade last price (15:59pm) for whole sample (1Q)

Table 1: Notations and Descriptions

2.2 Models

We mainly apply three models to do empirical analysis. As in Almgren et al.(2005) and many subsequent studies, ADV and σ normalize the price impact models across stocks and thereby make the impact coefficient λ comparable in the cross section.

- **The Original OW Model** - Chen el al.(2019) fit an Obizhaeva- Wang model with constant parameters:

$$I_{t+\Delta t} - I_t = -\beta I_t \Delta t + \lambda\sigma \frac{\Delta F_t}{ADV}$$

- **The Concave Propagator Model** - Bouchaud et al.(2009) review empirical studies of concave propagator models, which suggest $g(x) \propto x^c$ with $c \in [0.2, 0.5]$.We consider the most well-known model specification where scales with the square root of trade sizes:

$$I_{t+\Delta t} - I_t = -\beta I_t \Delta t + \lambda\sigma sgn(\Delta F_t)\sqrt{\frac{|\Delta F_t|}{ADV}}$$

- **The Depth Model** - Cont el al.(2013) propose an alternative model for stochastic liquidity:

$$\lambda_t \propto \frac{1}{D_t}$$

  Here, D_t is the current 'depth' on the best bid and ask queues. The corresponding linear price impact model is:

$$I_{t+\Delta t} - I_t = -\beta I_t \Delta t + \lambda\sigma \frac{\Delta F_t}{D_t}$$

2.3 Data Preprocessing

- **For entire data:**
  1. **Data:**. 2022 Q1 trade data of 261 tickers. Four of them contain too much NaN data, so after data preparation we abandon them and only use 257 tickers' data.
  2. **Filtering the dataframe**: The original DataFrame is filtered to extract records where the 'time' column falls within the range of 09:30:00 to 15:59:00.
  3. **Dropping NaN values**: Following the filtering step, any rows containing NaN values in the columns 'trade volume', 'trade last', 'buy volume', and 'sell volume' are removed.
  4. **Calculate ADV and σ:**
     ADV: average daily total trade volume (9:30am - 15:59pm).
     σ: standard deviation of the daily log return of trade last price (15:59pm).

- **For 5-min bin data:**
  1. **Define time bins**: Group data into 5-min time bins.
  2. **Check bin length**: After grouping the data into time bins, evaluate the length of each.
  3. **Filter out bins**: For each bin, if the number of rows is exactly 5, retain the bin. If a bin contains fewer or more than 5 rows, discard that entire bin from the analysis.
  4. **Construct final dataframe**: After filtering, compile the remaining bins into a new DataFrame that contains only the valid time bins with exactly 5 rows.
  5. **Calculate F and I for each time bin:**
     F: $\sum$(buy volume) - $\sum$(sell volume) of the entire time bin.
     I: trade last price at the end minus trade last - price at the beginning.

3 Ordinary Least Squares

We did the following work to all the 257 tickers and 3 models. But due to space limit, below we only show the analysis and results for AAPL OW as an example.

3.1 Summary

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.580
Model:                            OLS   Adj. R-squared (uncentered):              0.579
Method:                 Least Squares   F-statistic:                              3117.
Date:                Tue, 10 Dec 2024   Prob (F-statistic):                        0.00
Time:                        14:56:18   Log-Likelihood:                         -4458.0
No. Observations:                4523   AIC:                                      8920.
Df Residuals:                    4521   BIC:                                      8933.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.6276      0.010    -59.796      0.000      -0.648      -0.607
x2             0.2493      0.010     23.752      0.000       0.229       0.270
==============================================================================
Omnibus:                      359.711   Durbin-Watson:                   2.111
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2121.824
Skew:                           0.024   Prob(JB):                         0.00
Kurtosis:                       6.355   Cond. No.                         1.52
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 1: Result Dataframe

3.2 Residuals

3.2.1 Plot:

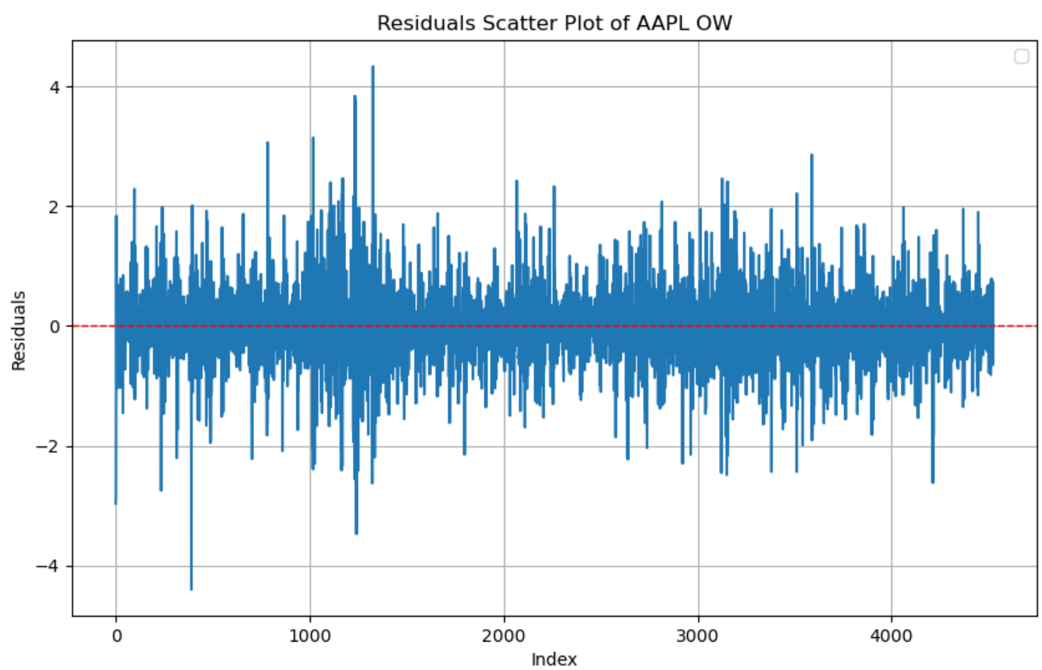

Figure 2: OLS Regression Residuals Plot

3.2.2 Normality:

- Using Kolmogorov-Smirnov Test:

Model	**KS stat**	**KS p_{value}**
OW	0.053151	$1.546174e-11$
Propagator	0.051613	$6.528337e-11$
Depth	0.049189	$5.978048e-10$

Table 2: KS Test Results

- Set the significance level α to be 0.01.
- Compare calculated p_{value} with α.
  - If $p_{value} < \alpha$: reject H_0, i.e. normal distribution assumption.
- According to the Table 2, all these three models do not follow normal distribution.

• Using QQ plot:

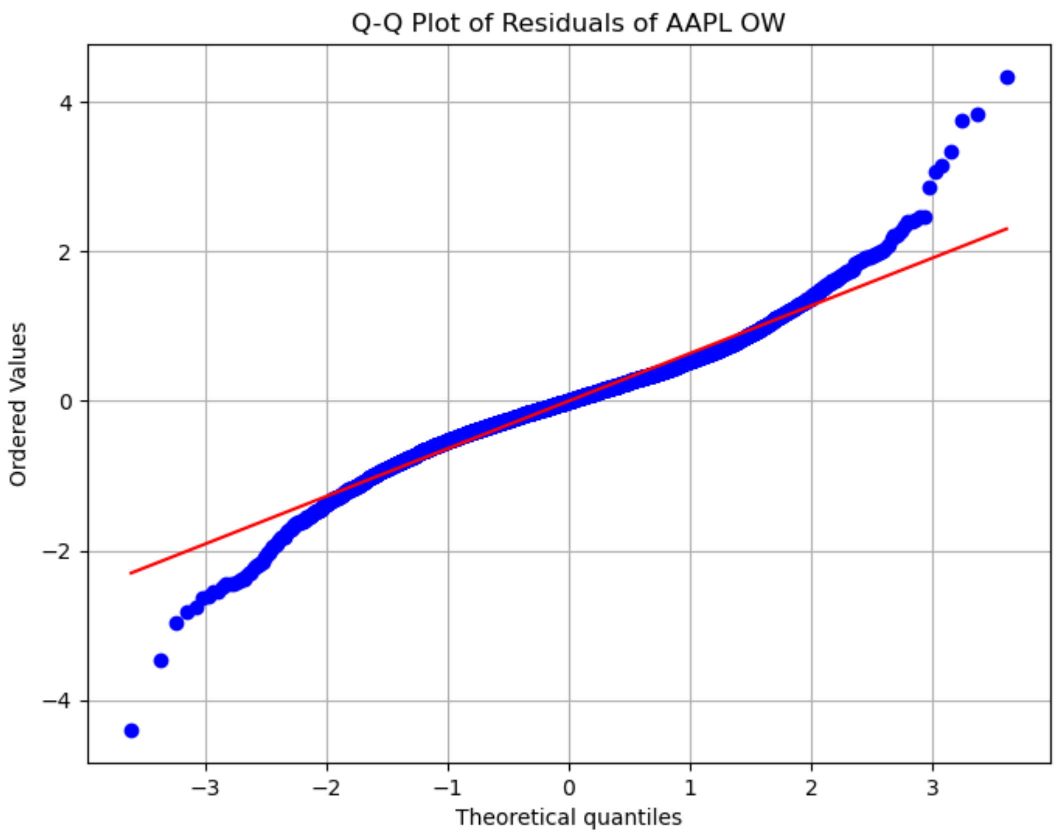

Figure 3: OLS Regression QQ Plot

- The trend of QQ plot could further validate our conclusion.

3.2.3 Autocorrelation:

• Using Augmented Dicky-Fuller Test:
  - Set the significance level α to be 0.01.
  - Compare calculated p_{value} with α.
    - If $p_{value} < \alpha$: reject H_0, i.e. autocorrelation assumption.
  - According to the Table 3, all these three models do not have autocorrelation.

• Using ACF and PACF:

Model	**ADF stat**	**ADF p_{value}**
OW	-51.542307	0.0
Propagator	-51.061731	0.0
Depth	-51.094781	0.0

Table 3: ADF Test Results

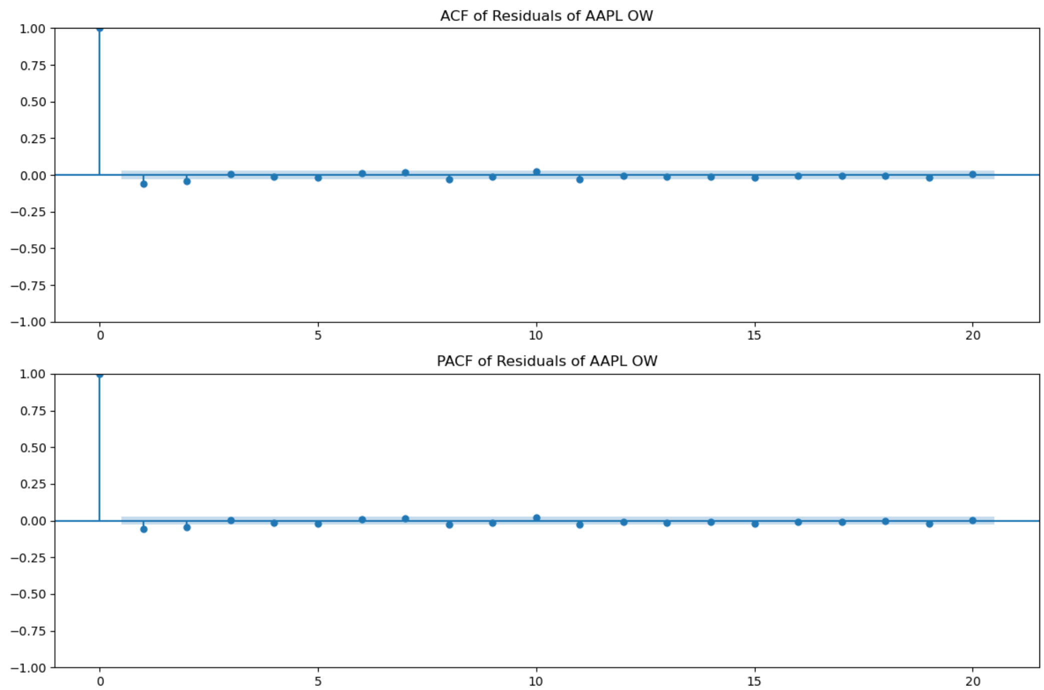

Figure 4: OLS Regression ACF/PACF Plot

– The trend of ACF/ PACF plots could further validate our conclusion.

3.3 Validity

- Set the significance level α to be 0.01.
- Compare calculated values with α.

Then we have two criteria:

- F $p_{value} < \alpha$ to show the linear relationship.
- ADF $p_{value} < \alpha$ to show no info in residuals.

If both criteria are met, we claim that the model is valid.

3.4 Best-fit Model

- We set the R^2_{adj} to be the criterion.
- Not use other metrics such as MSE, AIC, etc. because by definition formula, they are highly correlated with R^2_{adj} already.
- We will see it is reasonable to just use R^2_{adj} in section 5 Bayesian Regression.

3.5 Results

- Result 1: For all tickers with all models, the residuals are not normal but white noise, and the OLS is valid.
- Result 2: Best fit
  - Original OW model for 212 / 257 tickers
  - Depth model for 45 / 257 tickers
  - Propagator model for 0 / 257 tickers.

	ticker	model	beta	lambda	R-squared	Adj. R-squared	F-statistic	F-pvalue	MSE (Model)	MSE (Residual)	...	adf_stat	adf_pvalue	normality	white noise	valid
0	XLE	OW	-0.662227	0.152384	0.516682	0.516468	2416.544072	0.0	1168.476143	0.483532	...	-68.829960	0.0	0	1	1
1	XLE	Propagator	-0.669097	0.138489	0.513162	0.512947	2382.732643	0.0	1160.516883	0.487053	...	-68.560155	0.0	0	1	1
2	XLE	Depth	-0.667731	0.143046	0.514349	0.514134	2394.077725	0.0	1163.200422	0.485866	...	-68.350139	0.0	0	1	1
3	ALC	OW	-0.676363	0.154721	0.518023	0.517810	2429.022897	0.0	1171.250576	0.482190	...	-67.978168	0.0	0	1	1
4	ALC	Propagator	-0.694246	0.086034	0.502135	0.501914	2279.380893	0.0	1135.326671	0.498086	...	-67.436228	0.0	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
766	CP	Propagator	-0.698471	0.054332	0.501064	0.500843	2264.622061	0.0	1130.401459	0.499157	...	-32.720800	0.0	0	1	1
767	CP	Depth	-0.680455	0.110778	0.509795	0.509578	2345.117857	0.0	1150.097900	0.490422	...	-32.786032	0.0	0	1	1
768	BR	OW	-0.687481	0.196036	0.558321	0.558124	2835.348002	0.0	1252.872177	0.441876	...	-37.489916	0.0	0	1	1
769	BR	Propagator	-0.715142	0.096546	0.530348	0.530139	2532.879941	0.0	1190.101648	0.469861	...	-31.947586	0.0	0	1	1
770	BR	Depth	-0.694109	0.174566	0.550776	0.550575	2750.052088	0.0	1235.940819	0.449425	...	-37.213673	0.0	0	1	1

771 rows × 24 columns

Figure 5: Result Dataframe

4 Intraday Regression

4.1 Phenomenon

Markets exhibit significant variations in liquidity and activity throughout the trading day due to market thickness or thinness:

- **Market Thickness/Thinness:** Markets are typically thinner during the opening and closing hours, meaning there are fewer participants and less liquidity. Conversely, mid-day periods often experience more stable liquidity.
  - **Open:** Price discovery occurs; the market is often more volatile as participants respond to overnight news or global market movements.
  - **Midday:** Liquidity stabilizes as trading slows down, and the market becomes less sensitive to small trades.
  - **Close:** Liquidity and activity pick up again, often driven by portfolio rebalancing, algorithmic trading, and position closing.
- **Price Impact Differentials:** Trades executed during thin liquidity periods (e.g., early morning or late afternoon) may have a larger price impact compared to mid-day trades in thicker markets.

Because of these observed phenomenon, we are performing intraday regression analysis on our dataset to better understand and quantify these effects.

4.2 Results

Figure 6 shows the intraday pattern of parameters by OLS.

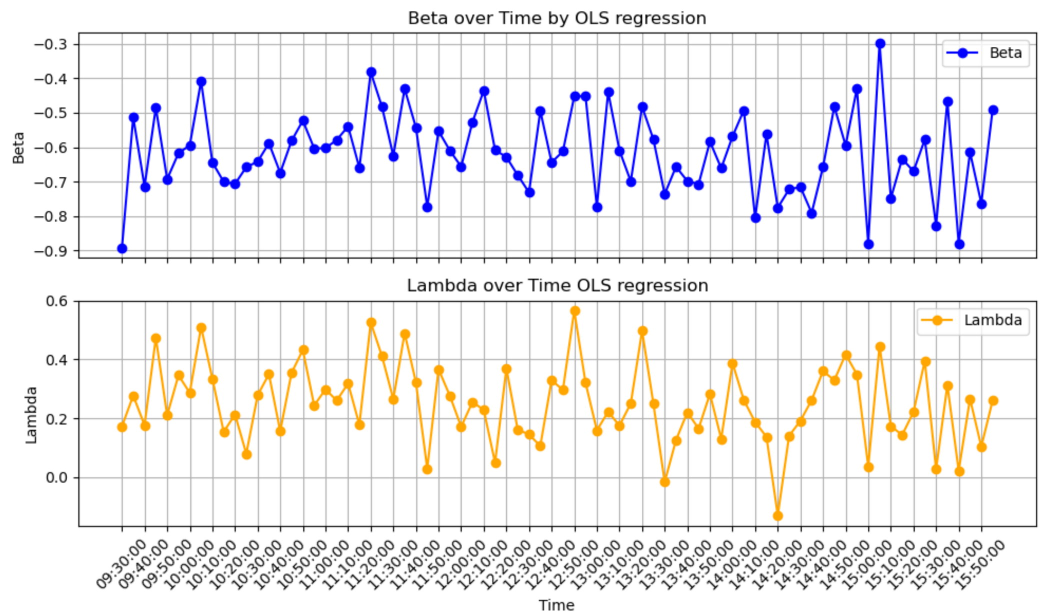

Figure 6: Intraday Pattern of Parameters By OLS

5 Bayesian Regression

5.1 Bayesian Framework

- Formula:

$$P(\theta|Data) = \frac{P(Data|\theta)P(\theta)}{\int P(Data|\theta)P(\theta)d\theta}.$$

  - Prior distribution: $P(\theta)$.
  - Likelihood function: $P(Data|\theta)$.
  - Posterior distribution: $P(\theta|Data)$.

5.2 Model Assumptions

- Prior distributions:
  $\beta \sim \mathcal{N}(0, 10)$
  $\lambda \sim \mathcal{N}(0, 10)$
  $\sigma \sim HalfNormal(1)$
- Observed data follows a Student T distribution:

$$y \sim StudentT(\mu, \sigma, \nu = 3), \quad where \quad \mu = \beta X_1 + \lambda X_2.$$

5.3 Sampling Method

- Due to the lack of analytical solutions in high-dimensional or complex models, MCMC is used with the No-U-Turn Sampler (NUTS) to converge to the posterior distribution.
- PyMC3 library is utilized for implementation.

5.4 Results

- Result 1: Statistics show that Bayesian Regression has a stable result.
  - Standard deviation is low.
  - HDI (High Density Interval) is narrow.

Parameter	Mean	SD	HDI 2.5%	HDI 97.5%	MCSE Mean	MCSE SD	ESS Bulk	ESS Tail	R-Hat
beta	-0.619	0.010	-0.640	-0.599	0.0	0.0	7337.0	6211.0	1.0
lambda	0.247	0.011	0.225	0.269	0.0	0.0	6810.0	5974.0	1.0
sigma	0.443	0.007	0.430	0.456	0.0	0.0	9578.0	6109.0	1.0

Table 4: Bayesian Regression Result

- Result 2: Statistics show that Bayesian Regression has a reliable result.
  - MC sampling error is low - all 0.0.
  - Effective sample size is very high - in thousands.
  - Convergence is great - all 1.0.
- Result 3: Posteriors show a normal distribution shape, suggesting it is a well-behaved model.

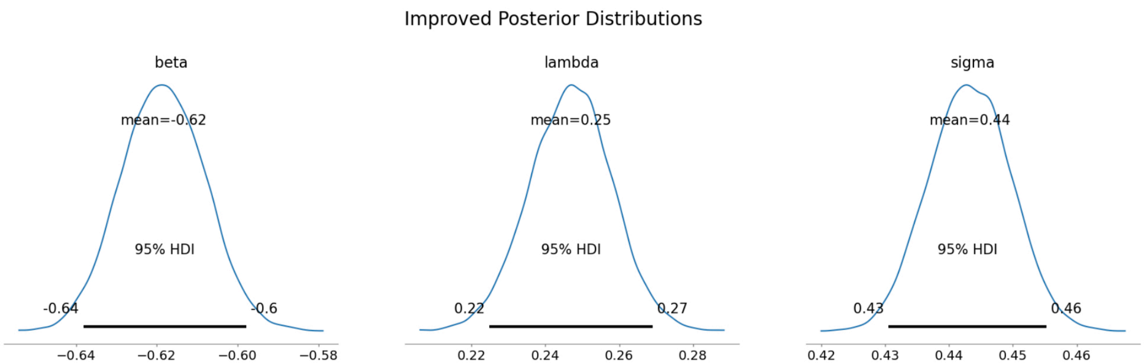

Figure 7: Improved Posterior Distribution

- Result 4: Mean of parameters in Bayesian Regression are close to those in OLS Regression, suggesting that the "informal, easy, cheap" OLS and R^2_{adj} is good enough.

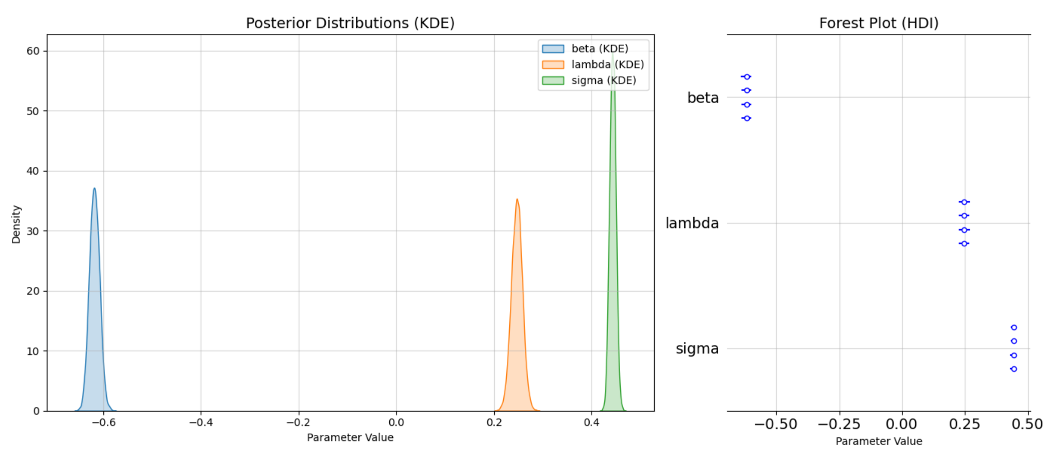

Figure 8: Kernel Density Estimation (KDE)

- Result 5: Comparison Between Intraday Pattern Between OLS and Bayesian Regression.

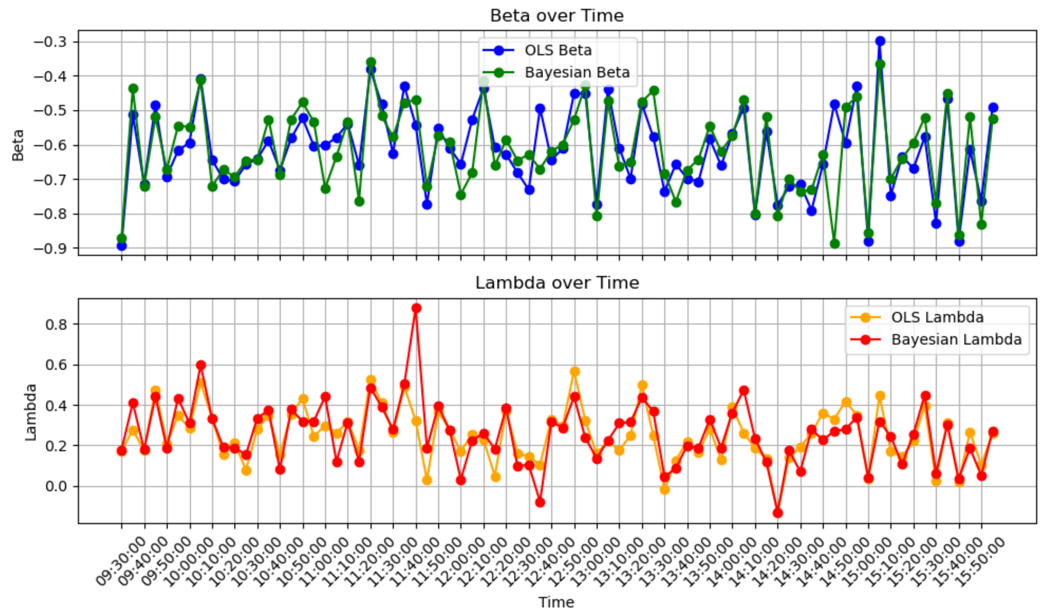

Figure 9: Intraday Pattern Comparison between OLS and Bayesian Regression

6 Conclusion

Our project investigated the price impact of trades through stochastic liquidity models, offering valuable insights into market dynamics.

We applied three models—the Original OW Model, the Concave Propagator Model, and the Depth Model—across 257 tickers. Classic OLS demonstrated a strong and reliable fit. Intraday regression highlighted significant variations in market liquidity, underscoring the importance of strategic trade execution. Bayesian regression yielded results comparable to classic OLS, confirming the latter as an effective and sufficient method.

Future work could focus on developing an enhanced empirical model to better represent stochastic liquidity, supported by robust theoretical justification.

7 References

1. J. Muhle-Karbe et al. "Stochastic Liquidity as a Proxy for Nonlinear Price Impact." *Oper. Res.* 72: 444-458,2023
2. R. Almgren and N. Chriss. "Optimal execution of portfolio transactions." *Journal of Risk*, 3:5–40, 2001.
3. Y. Chen, U. Horst, and H. Hai Tran. "Portfolio liquidation under transient price impact – theoretical solution and implementation with 100 NASDAQ stocks." Preprint, available at `arxiv.org`, 2019.
4. R. Cont, A. Kukanov, and S. Stoikov. "The price impact of order book events." *Journal of Financial Econometrics*, 12(1):47–88, 2013.
5. J.P. Bouchaud, J. Farmer, and F. Lillo. "How markets slowly digest changes in supply and demand." In T. Hens and K. R. Schenk-Hoppe, editors, *Handbook of Financial Markets: Dynamics and Evolution*, pages 57–160. North Holland, 2009.