Feature Engineering from Level 2 Limit Order Book Data for Volatility Prediction

Ayush Singh

Abstract

This document provides an in-depth exploration of feature engineering techniques using Level 2 Limit Order Book (LOB) data, specifically aimed at predicting future market volatility. Accurate volatility prediction is vital for risk management and trading strategies. Level 2 LOB data offers a rich, high-frequency view of market supply and demand. We detail numerous potential features, ranging from basic statistics to more complex imbalance and shape measures, including their mathematical definitions, underlying financial intuition, variations, and potential complexities. The goal is to generate a comprehensive set of candidate features that capture different facets of market microstructure, which can then be potentially refined through dimensionality reduction before being fed into predictive models.

1 Introduction

The Limit Order Book (LOB) is a fundamental component of modern electronic markets, recording outstanding buy (bid) and sell (ask) orders at different price levels. Level 2 data typically provides aggregated volume information for a fixed number of price levels away from the best bid and ask. This data offers a snapshot of market participants' intentions and the available liquidity.

Volatility prediction aims to forecast the magnitude of future price fluctuations. Since the LOB reflects the forces driving price changes (supply and demand), extracting meaningful signals (features) from it is a promising approach for volatility forecasting. High-frequency changes in LOB structure often precede periods of heightened volatility.

This document focuses on constructing features from Level 2 LOB snapshots. We assume that at any given timestamp t, we have access to the prices and aggregated volumes for the K best bid levels and K best ask levels.

2 Understanding Level 2 LOB Data

A snapshot of the LOB at time t typically provides:

- **Ask Prices:** $P^a_{k,t}$ for levels $k = 1, \ldots, K$. ($P^a_{1,t} < P^a_{2,t} < \cdots < P^a_{K,t}$)
- **Ask Volumes:** $V^a_{k,t}$ (total volume at price $P^a_{k,t}$) for levels $k = 1, \ldots, K$.
- **Bid Prices:** $P^b_{k,t}$ for levels $k = 1, \ldots, K$. ($P^b_{1,t} > P^b_{2,t} > \cdots > P^b_{K,t}$)
- **Bid Volumes:** $V^b_{k,t}$ (total volume at price $P^b_{k,t}$) for levels $k = 1, \ldots, K$.

Here, $k = 1$ represents the best price level (Best Ask $P^a_{1,t}$ and Best Bid $P^b_{1,t}$). K is the depth of the LOB data provided (e.g., K=5, 10, or 20).

3 Core LOB Concepts and Basic Features

These form the building blocks for many other features.

Best Bid and Offer (BBO)

- **Best Ask Price ($P^a_{1,t}$):** The lowest price at which sellers are currently willing to sell.
- **Best Bid Price ($P^b_{1,t}$):** The highest price at which buyers are currently willing to buy.
- **Best Ask Volume ($V^a_{1,t}$):** Volume available at the best ask price.
- **Best Bid Volume ($V^b_{1,t}$):** Volume available at the best bid price.

Usefulness: The BBO directly represents the immediate trading cost and available liquidity at the tightest prices. Changes in BBO prices drive transaction prices.

Mid-Price

Definition: The average of the best bid and ask prices.

$$P^{mid}_t = \frac{P^a_{1,t} + P^b_{1,t}}{2} \tag{1}$$

Usefulness: Often used as a reference price, less noisy than the last traded price due to bid-ask bounce. Its movement reflects the underlying consensus price drift. High fluctuation in mid-price indicates instability. **Variations:** Can be calculated using deeper levels, but the BBO-based definition is standard.

Spread

Absolute Spread

Definition: The difference between the best ask and best bid prices.

$$S_t = P^a_{1,t} - P^b_{1,t} \tag{2}$$

Usefulness: Represents the cost of executing a round-trip market order. A key indicator of liquidity and uncertainty. Wider spreads often correlate with higher volatility and lower liquidity. **Variations:** Spreads at deeper levels ($P^a_{k,t} - P^b_{k,t}$) might indicate liquidity further down the book.

Relative Spread

Definition: Absolute spread normalized by the mid-price.

$$S^{rel}_t = \frac{P^a_{1,t} - P^b_{1,t}}{P^{mid}_t} = \frac{2S_t}{P^a_{1,t} + P^b_{1,t}} \tag{3}$$

Usefulness: Provides a scale-free measure of the spread, allowing comparison across different assets or price levels. Important as absolute spreads naturally increase with price.

Spread Volatility

Definition: Standard deviation of the spread (S_t or S^{rel}_t) over a recent time window (e.g., last M snapshots). **Usefulness:** Measures the stability of the trading cost. Rapidly changing spreads can indicate market nervousness or fluctuating liquidity, potentially preceding higher price volatility.

Weighted Average Price (WAP)

Definition: A price measure weighted by the volumes at the best bid and ask. Often considered more robust than the mid-price.

$$P_t^{WAP} = \frac{P_{1,t}^b V_{1,t}^a + P_{1,t}^a V_{1,t}^b}{V_{1,t}^a + V_{1,t}^b} \tag{4}$$

Usefulness: Reflects the price level considering the immediately available liquidity. If $V_{1,t}^b \gg V_{1,t}^a$, WAP is closer to $P_{1,t}^a$, indicating pressure towards the ask side. Its fluctuations can be used to calculate high-frequency volatility. **Variations:** Can be extended to include multiple levels, e.g., WAP based on the first k levels:

$$P_t^{WAP(k)} = \frac{\sum_{i=1}^k (P_{i,t}^b V_{i,t}^a + P_{i,t}^a V_{i,t}^b)}{\sum_{i=1}^k (V_{i,t}^a + V_{i,t}^b)} \quad \text{(Conceptual; needs careful definition of matching levels)} \tag{5}$$

A more common multi-level WAP considers the price needed to execute a certain volume.

4 Depth and Volume Features

These features quantify the amount of liquidity available at various price levels.

Volume at Best Levels

Definition: $V_{1,t}^a$ and $V_{1,t}^b$ as defined earlier. **Usefulness:** Immediate liquidity. Low volumes can mean the price is sensitive to small market orders.

Accumulated Volume (Depth)

Definition: Total volume available within the first k levels on each side.

$$V_t^{a,cum(k)} = \sum_{i=1}^k V_{i,t}^a \tag{6}$$

$$V_t^{b,cum(k)} = \sum_{i=1}^k V_{i,t}^b \tag{7}$$

Usefulness: Measures the total liquidity buffer within a certain price range from the BBO. Low accumulated volume suggests fragility and potential for larger price swings if significant orders arrive. Commonly calculated for $k = 5$ or $k = 10$. **Variations:** Can calculate volume within a fixed price distance (e.g., X ticks or Y basis points) from the mid-price or BBO, rather than a fixed number of levels.

Volume Ratios

Definition: Ratio of volumes at different levels or sides.

- Ask vs Bid Volume (Level 1): $V_{1,t}^a / V_{1,t}^b$
- Ask vs Bid Volume (Accumulated k levels): $V_t^{a,cum(k)} / V_t^{b,cum(k)}$
- Level 1 vs Deeper Levels: $V_{1,t}^b / V_t^{b,cum(k)}$ or $V_{1,t}^a / V_t^{a,cum(k)}$

Usefulness: These ratios can indicate relative supply/demand pressure and how concentrated liquidity is at the best prices versus deeper in the book. A high $V_{1,t}^b / V_t^{b,cum(k)}$ means most buy support is right at the top.

Price Level Density

Definition: How many price levels are occupied within a certain price range, or the average price step between levels.

$$\text{Ask Density}(k) = \frac{k}{P_{k,t}^a - P_{1,t}^a} \quad (\text{for } k > 1) \tag{8}$$

Similarly for the bid side. **Usefulness:** High density might indicate more competition or fragmented orders, potentially leading to slower price movements unless density is very low overall. Low density (large gaps) might lead to jumpier price action.

5 Imbalance Features

These are often considered highly predictive as they measure the relative pressure between buying and selling interest.

Order Book Imbalance (OBI) - Level 1

Definition: Difference between bid and ask volume at the best levels, often normalized.

$$OBI_t^{(1)} = \frac{V_{1,t}^b - V_{1,t}^a}{V_{1,t}^b + V_{1,t}^a} \tag{9}$$

Ranges from -1 (only asks) to +1 (only bids). **Usefulness:** Powerful predictor of short-term price direction. Positive imbalance suggests upward pressure, negative suggests downward. Extreme imbalances might precede price moves and potentially higher volatility as the imbalance resolves.

Accumulated Order Book Imbalance (Depth Imbalance)

Definition: Imbalance calculated using accumulated volumes over k levels.

$$OBI_t^{(k)} = \frac{V_t^{b,cum(k)} - V_t^{a,cum(k)}}{V_t^{b,cum(k)} + V_t^{a,cum(k)}} \tag{10}$$

Usefulness: Captures broader supply/demand pressure beyond just the BBO. May be more robust than Level 1 imbalance. Choice of k is important (e.g., $k = 5, 10$).

Weighted Order Book Imbalance

Definition: Similar to OBI, but volumes are weighted, typically giving more importance to levels closer to the BBO. A common weighting scheme uses the inverse of the distance from the mid-price or opposite BBO, or decays exponentially. Example using exponential decay α^i:

$$OBI_t^{W(k)} = \frac{\sum_{i=1}^k w_i V_{i,t}^b - \sum_{i=1}^k w_i V_{i,t}^a}{\sum_{i=1}^k w_i (V_{i,t}^b + V_{i,t}^a)} \quad \text{where } w_i \text{ decreases with } i. \tag{11}$$

For instance, $w_i = (1/\text{distance from mid-price})^\delta$ or $w_i = e^{-\delta(i-1)}$. **Usefulness:** Acknowledges that liquidity closer to the current price is more relevant for immediate price movements. Can potentially provide a more sensitive measure of pressure.

Imbalance Slope / Profile Shape

Definition: Measures how imbalance changes across levels. E.g., the difference between $OBI_t^{(1)}$ and $OBI_t^{(k)}$, or fitting a slope to the imbalance calculated level-by-level.

$$\text{Imbalance Slope}(k) \approx \frac{OBI_t^{(k)} - OBI_t^{(1)}}{k-1} \quad \text{(simplistic)} \tag{12}$$

More formally, calculate $Imb_i = (V_{i,t}^b - V_{i,t}^a)/(V_{i,t}^b + V_{i,t}^a)$ for $i = 1...k$ and fit a linear regression: $Imb_i \sim \beta_0 + \beta_1 i$. The slope β_1 is the feature. **Usefulness:** Captures whether pressure is concentrated at the BBO or builds up deeper in the book. A steep positive slope might indicate strong underlying buy interest even if the BBO is balanced.

6 Volatility and Price Movement Features (from LOB)

While the target is future volatility, measures of *current* micro-volatility derived from LOB states can be predictive.

Mid-Price / WAP Volatility

Definition: Standard deviation of P_t^{mid} or P_t^{WAP} over a recent window of M snapshots.

$$\sigma^{mid}(M)_t = \text{StdDev}(P_{t-M+1}^{mid}, \ldots, P_t^{mid}) \tag{13}$$

Usefulness: Direct measure of current price instability at the micro-level. High current micro-volatility often persists.

BBO Stability / Flickering

Definition: Frequency of changes in the $P_{1,t}^a$ or $P_{1,t}^b$ over a recent window. Can be measured as the number of times the best price level changes, or the number of times the BBO prices themselves change. **Usefulness:** High frequency of BBO changes indicates active price discovery and potential uncertainty or high trading activity, which might lead to higher macro volatility.

High-Frequency LOB Return

Definition: Log return of the mid-price or WAP.

$$r_{t,\Delta t}^{mid} = \log(P_t^{mid}) - \log(P_{t-\Delta t}^{mid}) \tag{14}$$

$$r_{t,\Delta t}^{WAP} = \log(P_t^{WAP}) - \log(P_{t-\Delta t}^{WAP}) \tag{15}$$

Usefulness: These high-frequency returns are the basis for calculating realized volatility using LOB data (e.g., sum of squared r^{WAP} over 1-minute intervals). Statistics of these returns (e.g., standard deviation, kurtosis) over a recent window can be features.

7 More Complex / Combined Features

Liquidity Consumption Ratio

Definition: Ratio of traded volume (requires trade data) within a small interval to the standing volume at the BBO just before the interval. **Usefulness:** Measures how quickly available liquidity is being consumed. High consumption suggests aggressive order flow, potentially leading to volatility. (Note: Requires matching LOB snapshots with trade data).

Book Pressure Decay

Definition: Measures how quickly volume drops off as one moves deeper into the book. Can be modeled by fitting an exponential decay function to $V_{k,t}^b$ and $V_{k,t}^a$ as k increases. The decay parameter is the feature. **Usefulness:** A rapid decay ('thin book') suggests fragility and higher potential volatility compared to a 'deep' book where volume is substantial across many levels.

Normalized Price Levels

Definition: Expressing deeper LOB price levels ($P_{k,t}^a$, $P_{k,t}^b$) relative to the mid-price or BBO, possibly normalized by the spread or recent volatility. **Usefulness:** Creates scale-free measures of where liquidity is positioned relative to the current market price.

8 Temporal Aggregation

LOB features are generated at very high frequency (every snapshot). To predict volatility over a longer horizon (e.g., next 10 minutes, next hour, next day), these high-frequency features need to be aggregated over a suitable lookback interval (e.g., previous 1 minute, 5 minutes, 1 hour). Common aggregation functions include:

- Mean
- Standard Deviation (Volatility of the feature itself)
- Sum (especially for volumes or counts)
- Minimum / Maximum
- Median / Quantiles
- Skewness / Kurtosis
- Last value in the interval
- Trend (e.g., slope of the feature over the interval)

For example, instead of using the instantaneous $OBI_t^{(5)}$, one might use the average $OBI^{(5)}$ over the last 60 seconds, or the standard deviation of $OBI^{(5)}$ over the last 60 seconds. Creating features using multiple aggregation functions and multiple lookback intervals significantly expands the feature set.

9 Feature Selection and Dimensionality Reduction

Generating features across multiple types, levels (k), aggregation functions, and lookback windows can lead to a very high-dimensional feature space. Many features might be highly correlated (e.g., different imbalance measures) or contain noise. This necessitates dimensionality reduction or feature selection techniques before feeding them into a predictive model (especially linear models or tree-based models sensitive to correlation; deep learning might handle this better internally but can still benefit). Common approaches include:

- **Correlation Analysis:** Removing features that are highly correlated (e.g., ¿ 0.9 or 0.95) with another feature.

- **Principal Component Analysis (PCA):** Linearly transforming features into a smaller set of uncorrelated principal components that capture most of the variance. Loses direct interpretability.
- **Feature Importance Ranking:** Using models like Random Forests or Gradient Boosting (e.g., LightGBM, XGBoost) to estimate the contribution of each feature to prediction accuracy. Keep the top N features. LASSO regression also performs implicit feature selection.
- **Domain Knowledge:** Prioritizing features known to be important based on financial literature or trader intuition.

The user's belief that feature engineering followed by dimensionality reduction can boost performance is well-founded, as it aims to provide the model with potent, non-redundant signals.

10 Conclusion

Level 2 LOB data is a treasure trove of information about market microstructure dynamics. By carefully engineering features that capture spread, depth, imbalance, micro-volatility, and shape characteristics of the book, we can create powerful inputs for volatility prediction models. The features detailed here range from simple statistics to complex, derived measures. The process typically involves generating a large set of candidate features based on these ideas, aggregating them over relevant time intervals, and then applying selection or dimensionality reduction techniques to arrive at a final feature set optimized for predictive performance. The choice of features, aggregation methods, and reduction techniques should ideally be guided by empirical validation on the specific dataset and prediction task.