Heavy-Tailed Features and Empirical Analysis of the Limit Order Book Volume Profiles in Futures Markets

Kylie-Anne Richards\textsuperscript{a,1}, Gareth W. Peters\textsuperscript{a,b,c,d}, William Dunsmuir\textsuperscript{a,b,c}

\textsuperscript{a}School of Mathematics and Statistics, University of NSW
\textsuperscript{b}Boronia Capital Pty. Ltd.
\textsuperscript{c}Department of Statistical Science, University College London
\textsuperscript{d}CMIS-Commonwealth Scientific and Industrial Research Organisation

Abstract

This paper poses a few fundamental questions regarding the attributes of the volume profile of a Limit Order Books stochastic structure by taking into consideration aspects of intraday and interday statistical features, the impact of different exchange features and the impact of market participants in different asset sectors. This paper aims to address the following questions:

1. Is there statistical evidence that heavy-tailed sub-exponential volume profiles occur at different levels of the Limit Order Book on the bid and ask and if so does this happen on an intra or interday time scale?

2. In futures exchanges, are heavy tail features exchange (CBOT, CME, EUREX, SGX and COMEX) or asset class (government bonds, equities and precious metals) dependent and do they happen on ultra-high ($<$1sec) or mid-range (1sec -10min) high frequency data?

3. Does the presence of stochastic heavy-tailed volume profile features evolve in a manner that would inform or be indicative of market participant behaviors, such as high frequency algorithmic trading, quote stuffing and price discovery intra-daily?

4. Is there statistical evidence for a need to consider dynamic behavior of the parameters of models for Limit Order Book volume profiles on an intra-daily time scale?

Progress on aspects of each question is obtained via statistically rigorous results to verify the empirical findings for an unprecedentedly large set of futures market LOB data. The data comprises several exchanges, several futures asset classes and all trading days of 2010, using market depth (Type II) order book data to 5 levels on the bid and ask.

Keywords: Limit Order Book, Futures Markets, High Frequency Volume Profiles,

1. Introduction

Studying the Limit Order Book (LOB) for a given asset in a given exchange involves the statistical analysis of a challenging continuous time multivariate event driven stochastic process for both price and volume on the bid and ask at different levels of “market depths”. Often the analysis of such a process involves distortions of the underlying process through transformations either by aggregation of volumes or prices, truncation, averaging and smoothing, matching or rounding into alternative class of stochastic processes to be studied either on more traditional discrete state spaces in price or volume, and or time. The analysis
of a LOB is further compounded by the impact on this process of market and exchange mechanisms, endogenous and exogenous macro-economic factors, “news impacts”, regulatory requirements, asymmetric information flows between market participants and the limited understanding currently available on the structure of noise in this process which is often changing as a function of the sampling rate at which one partially observes the process (micro-structure noise). In this context we propose to consider a time series based analysis of the volume profiles using market depth (Type II) order book data to 5 levels on the bid and ask.

The majority of the literature concerning the statistical properties of high frequency data primarily focuses on returns series of trade data. These studies have produced a set of broadly recognized stylized features such as fat tails, volatility clustering, autocorrelation and with various suggestions for the form of the distribution: student’s-t, hyperbolic, normal inverse Gaussian and exponentially truncated. As highlighted by Chakrabort et al. (2010), there is no general consensus for the best statistical model. Whilst these studies are not the direct focus of this research, they have formed the building blocks from which LOB research is based. In terms of LOB modeling, typically statistical models estimated for LOB data in the literature are primarily focused on relative prices. The limited number of studies that do consider volumes do so in the framework of data being aggregated at a fixed grid of price points (ticks) away from the best bid and ask, typically not assessing intra and inter day volume features. The deterministic price point aggregation commonly used in the limited studies on the LOB volume profiles can introduce two possible artificial features or distortions to the LOB multivariate stochastic process that may not be of relevance. The first involves a zero inflation if the price grid at which volumes are aggregated is too fine and the second is the opposite extreme in which an aggregation distortion is introduced if a course price grid is selected. To avoid this question, we will undertake the analysis in a structured manner through a principled transformation of the data to a time series structure based on Market Depth data to 5 levels on the bid and ask, thereby avoiding the volume aggregation to a deterministically set grid of price points.

Since we are interested in multiple exchanges and multiple asset classes effectively we have several LOB multivariate stochastic structures to explore. We aim to make inference on common features of these structures, primarily related to the potential for sub-exponential behavior in the tail features on the marginal distributions of the volume profiles for each level of depth on the bid and ask. Studying the LOB volume profiles in this fashion and its behavior between asset classes, exchanges and over time is still a relatively under explored. The LOB provides a rich and interesting source of data for development of financial models. Such analysis will aid in the process of attempting to address the key questions posed in the abstract to the paper.

The development of accurate statistical models and the estimation of such models for the volume profile in the LOB are also relatively under explored. It is a non-trivial task given the massive volumes of data available for each LOB structure for each asset. A range of different futures market assets, at a selection of sampling frequencies intra daily for the year 2010 are considered. The scope of this study and the findings for this study is to our knowledge unprecedented in its coverage in the literature on such LOB structures for volume or price profile modeling from a statistical perspective. The approach of this study considers volumes on individual price levels and allows for intraday and interday modeling, with consideration of characteristics such as the importance of considering time varying parameters. The intention of this work is to provide a solid statistical understanding of these stochastic structures to aid future model development.

The range of flexible statistical models estimated for the volume profile include the families: Generalized Extreme Value distributions (GEV), the Generalized Pareto Distributions (GPD) and univariate α-stable distributions. To ensure the accuracy of the statistical and financial conclusions drawn from the analysis we consider several parameter estimation approaches for each model which include: generalized method of moment based approaches; empirical percentile based approaches; mixed maximum-likelihood and moment based methods; as well as L-moment based estimators. In the process we comment on the suitability for
practical estimation of such models using ultra-high frequency LOB data sets.

Undertaking such a comprehensive study involved 100’sGb to terabytes of data to be analyzed; this posed a significant challenge to any statistical analysis. We clearly detail how such analysis can be performed including analysis of the applicability and attributes of different statistical estimation procedures for the models considered for the volume profile when applied in this context. In the process we comment on the statistical features and suitability for practical estimation of such models over such massive data sets under these statistical procedures.

We begin with a detailed analysis of the empirical features of the level 1 bid and ask, studied over a range of high frequency sampling rates, with sub-sampling frequencies of 10 seconds, 5 seconds, 2 seconds and 1 second intraday for each trading day of 2010. These trading intervals are considered in practice mid-range high frequency for hedge funds and traders operating in this area. By considering such sampling rates we have two main aims, the first is to disambiguate the real stochastic behavior of the heavy tailed features of the LOB volume profile structures from the high-frequency ”micro-structure noise”, see recent work in al Dayri (2011). The second is to understand for a given market and asset class, whether basic statistical features such as heavy tailed attributes are persistent in the stochastic processes and possibly arising at different sampling rates to different degrees. This understanding will provide the ability to make qualitative conclusions regarding potential for volume impacts from the high frequency traders in terms of the portions of the volume coming from legitimate trading activity versus “malicious” or “disruptive”, potentially destabilizing activities, such as ”quote stuffing” and price discovery mechanisms. Where quote stuffing is known as the practice of quickly entering and withdrawing large orders in an attempt to flood the market with quotes that competitors have to process, thus causing them to lose their competitive edge in high frequency trading. Examples of price discovery mechanisms include the periodic swamping of single share orders in the exchange over very short time frames with the intention to slow down the system and increase the volume profile depth artificially. The understanding of the impacts of such market participant behavior is a primary interest to regulatory authorities currently investigating regulation for such high frequency trading.

The LOB can be defined from two primary sources of financial instruments traded on financial markets: those traded through a physical trading floor or an electronic exchange system. In addition different types of markets are defined by the type of instruments that are traded on them, for example, equity markets, government and corporate bond markets, foreign exchange markets, derivative markets, and so on. The markets considered in this study are electronic in nature and the focus will be on the derivatives and futures markets.

Examples of electronic markets include the EUREX, Chicago Mercantile Exchange (CME) and the Singapore Exchange (SGX). The activities of these markets vary, but generally speaking, they allow for the trading of various financial instruments, such as options and futures contracts which we consider in this study. A futures contract can be an exchange traded financial instrument, whereby the product has an underlying asset associated with it. For example, if a person was to trade an equity futures contract, they would be trading a futures contract written over an equity or equity index. For this study we consider futures over government bonds, equity indices and precious metals traded across five different exchanges.

When studying in a formal statistical manner the attributes of the LOB previous studies of LOB characteristics on equity or futures markets have typically involved one or two assets over a limited time period with all data being aggregated at price points (ticks) away from the level 1 bid and ask, see Biais et al. (1995), Bouchaud et al. (2002), Challet and Stinchcombe (2001), Gu et al. (2008), Chakrabort et al. (2010). Any statistical models estimated to LOB data in the literature in this setting fail to consider the inter and intraday characteristics such as the important consideration of time varying parameters that we observe in all assets under consideration.

This work extends from papers such as, Biais et al. (1995), Challet and Stinchcombe (2001), Maslov and
Mills (2001), Bouchaud et al. (2002), Gu et al. (2008) and Chakrabort et al. (2010), Gould et al. (2012) as we significantly broaden both the scope of assets considered, and the range of flexible statistical models estimated. Previous studies only consider assets from a single exchange, whereas this study considers futures with different asset classes for the underlying instrument, across five different exchanges. The time frame of estimation of both the inter and intraday level is across one year, a much more extensive study in terms of the cross-section of the sample used to test these novel distributional features. In addition from the perspective of the scope of models considered, previous studies have primarily focused on the gamma distribution which as we have found, is limited in its ability to accurately capture the skewness and kurtosis features in the LOB volume data. Considerations of tail properties of the volume profile have not been explored in previous studies on LOB volume data. Further to tail properties, we also consider the dynamical nature of the time varying parameters and the impact of different estimation procedures for each distribution; addressing numerous computational constraints when applying these advanced distributions and multiple estimation methods to ultra-high frequency data.

2. Understanding the Limit Order Book and Volume Profile

It is important to consider several different exchanges in the analysis since the volume profile of the LOB's stochastic structure for a given asset will be directly influenced by the fact that each market is defined by its market structure. This structure may vary between markets depending on the trading rules and trading systems (Harris (2003)) which determine who can trade and how and what instruments are traded. In particular we consider markets in this study which are classified as order-driven markets which constitute the most common form of market. Gould et al. (2012) give a good description of the alternative type of market place, a quote-driven market and provide a mathematical description of the more flexible order driven market which we consider in this study.

Order driven markets operate according to pre-specified trading rules and orders typically have a price and time priority. A limit order is defined as any order with a specified maximum buy price or a specified minimum sell price. A limit order can be submitted during the opening rotation, normal trading and closing rotation, given the market has an opening and closing auction. When a limit order is submitted, it enters a queue where priority is given to the highest priced limit order for buy side, and lowest priced limit order for sell side. When two orders are identically priced, the order submitted first (older order) is given priority. A number of parameters must be specified when submitting a limit order: the limit price, buy (or sell) and order volume.

A limit order book (LOB), otherwise known as the queue, refers to a list of unexecuted limit orders with a specified buy or sell price of a particular asset. There are two sides to a LOB, the bid (limit order to buy an asset) and the ask (limit order to sell an asset). The orders are displayed at different price levels in the LOB. If a trader wants to immediately buy an asset at the best ask price, they submit a market order, which means buyer and seller matching via submission of a buy market order matching a sell limit order. The result is a trade. In this context, the LOB can be described as a multilevel stochastic process of orders at different price points on the bid and ask side. Incorporated within this complex process are elements such as cancellation of orders, price and volume amendments and the dynamics of market orders on limit orders.

In addition we take into consideration the interesting fact that several exchange markets allow for two distinct modes of trade placement on the exchange and therefore different volume profile dynamics intra and inter day on each exchange. There are the high-frequency trading algorithms (automated algorithmic trading of hedge funds and investment banks) which place hardware next to the exchange, minimizing the milliseconds it takes for trades to be placed on an exchange. Then there are the more traditional, slower trading dynamics coming from investment banking activity, insurance funds and private traders which go
through a broker to place trades on the market. This creates differences in LOB volume profile dynamics, where there is an increasing belief that although the high frequency traders may add liquidity (depth in the volume profile) to LOB stochastic structures, they also increase volatility in both share price and volume. Importantly, we will consider limit orders only during “liquid market hours”, excluding auction periods and excluding days when the exchange is open on a public holiday. This minimizes the effect on the modeling of the volume dynamics attributed to short periods related to exchange specific pre-market open auction mechanisms, instead focusing the modeling on the stochastic volume profile attributes. However, market specific attributes, such as trading hours, minimum tick size, minimum lot size, exclusion of special trades (ie crossings) and various rules around limit order placement, need to be considered when constructing the data sets used for modeling.

2.1. Background on Limit Order Book Modelling and Scope of the Volume Profile Analysis

The first group of studies we consider, relate to the shape of the LOB. Most of these empirical studies consider the order book in it’s entirety, rather than looking at the volume on the individual levels. Using equity LOB data from Paris Bourse (the historical Paris stock exchange), two studies by Biais et al. (1995) and Bouchaud et al. (2002) originally proposed that the shape of the order book was monotonically increasing away from the best bid/ask with the highest level of flow occurring at the best bid/ask. Consistent with these findings, a study by Challet and Stinchcombe (2001) considered the LOB for the Island ECN (which is an alternative trading system for US equities, where the prices should match those on the NASDAQ) and described the shape of the LOB as convex and peaking at the best bid/ask. However, more recent studies contradict these findings. An investigation by Potters and Bouchaud (2003) on the statistical properties of the NASDAQ used a zero-intelligence model to reproduce the empirical results and described the shape of the order book as being humped (ie the maximum being away from the best bid/ask) and this being due to the non-uniform, power-law distributed flow of incoming orders. This study raises the important consideration of non-uniform cancellation rates, due to the higher frequency of amendments of orders at, or around the best bid/ask. These findings have been further supported in recent studies by Gu et al. (2008) on the Shenzhen Stock Exchange considering 23 stocks and Chakrabort et al. (2010) on the Paris Bourse, both observing the maximum away from the best bid/ask.

The U-shaped curve (volume smile) which is described by Biais et al. (1995), is well known for trade data but is also shown to be the case for limit order data. Both Challet and Stinchcombe (2001) and Chakrabort et al. (2010) showed that the U-shaped curve does exist for limit orders. Challet and Stinchcombe (2001) also showed that orders have a tendency to cluster both in size and position, they tend to have a size of a multiple of 10, 100 or 1000, and to be placed at round prices, or at halves. Other research on the LOB volumes has not indicated this feature and it may be specific to the ECN data used, rather than a general feature of a LOB.

When considering the distributional properties of the volume on the order book, we categorize the studies into those considering the volume at each price point (tick) away from the best bid/ask and those considering volume at each consolidated price level. Bouchaud et al. (2002) describe a Gamma distribution to be the best fit for volumes at each price point. They found that the Gamma distribution (for three stocks) had a scale parameter of between (0.7, 0.8) and a shape parameter of approximately 2700. Gu et al. (2008) also consider volume at a price point and found that the volumes on the best bid/ask are best represented by a log-normal distribution. For order sizes that are smaller than average, they found that the distribution deviates from a log-normal distribution and exhibits power-law behavior with exponents for different levels being: 4.19 ± 0.09 for level 1, 2.61 ± 0.03 for level 2, 2.67 ± 0.05 for level 3. Using NASDAQ level II data, Maslov and Mills (2001) conducted an empirical study of statistical properties of the LOB considering volume at each consolidated price level. They initially investigated a power law distribution for limit order sizes and found that it was consistent with an exponent of 1.0 ± 0.3. They
also implemented a better fitting log-normal distribution to the data which has an effective power law exponent equal to 2 in the middle of the observed range.

Another important feature demonstrated by Gu et al. (2008), is the assessment of temporal dependency. They characterize the temporal dependency by the Hurst index and compare the first 3 levels of the LOB. The results are consistent with long memory, with Hurst indices significantly larger than 0.5 for both bid and ask side. For example, $H_1 = 0.76 \pm 0.01$, $H_2 = 0.83 \pm 0.01$, $H_1 = 0.81 \pm 0.01$ where \{1, 2, 3\} are the first three levels on the buy side.

3. Volume Profile Limit Order Book Data Characterizations

The source of the data used for this analysis involves the product known as the Thompson Reuters Tick History (TRTH), obtained as a collaborative analysis with the quantitative team at Boronia Managed Funds Pty. Ltd. This product provides data such as trades, level I and level II LOB data, although the level II data is consolidated by price and only a specified number of price levels are available for each market. The list below, describes the notation relating to data used throughout the analysis in this paper.

- **Market Depth Data** also known as *Type II* data refers to LOB data that contains all data for a pre-specified number of price levels (ie typically 5-10 price levels) with consolidation of volume at each price level. For example, for 5 levels of data you will only see 5 lines of orders on the bid and 5 lines of orders on the ask even though the bid may be made up of X individual orders and the ask made up of Y individual orders.

- **Trade data**, also known as *Time and Sales data* refers to the orders that have been executed, realized as trades and forming the traded price of the instrument.

- **Matched data** is comprised of LOB data (either level I, level II or level III data) that is time matched with trade data. Because the data is event driven, there will not always be a new trade every time the LOB changes. If the LOB changes and a new record is made but no trade has been executed at the time instant, then nothing will be recorded for the trade.

- **Time series data** is a version of transformed data and consistent with the above example, it is data that is aggregated in some way over evenly spaced time intervals.

In this paper we consider the Market Depth (Level II) LOB volume profile time series data for a range of futures market assets as detailed in Table 1.

| Asset Name          | Acronym | Market Hours (local time) | Exchange |
|---------------------|---------|---------------------------|----------|
| **Interest rate derivatives** |         |                           |          |
| 1. 5 Year T-Note    | 5YTN    | 7:20:00 to 13:50:00       | CBOT     |
| 2. Euro-BOBL        | BOBL    | 8:02:00 to 18:50:00       | EUREX    |
| **Equity derivatives** |         |                           |          |
| 3. SIMEX Nikkei 225 | NIKKEI  | 8:05:00 to 13:50:00       | SGX      |
| 4. E-mini S&P 500   | SP500   | 08:30:00 to 14:50:00      | CME      |
| **Precious Metals** |         |                           |          |
| 5. Gold             | GOLD    | 06:20:00 to 13:25:00      | COMEX    |
| 5. Silver           | SILVER  | 08:25:00 to 13:10:00      | COMEX    |

Table 1: Asset description used in the analysis and modelling. Market hours refer to the liquid market hours in local trading time of the exchange. Note - liquid market hours need not match the extent of the hours for which the exchange is open for trading.
4. Empirical Analysis of Limit Order Book Volume Profiles

In the majority of the analysis that follows we use time series data constructed by taking a sub-sample of data, whereby the last volume of the LOB at a specified time increment of, for example 10 seconds is recorded. The volume is defined as $X_{i,j}^{t,d} \in \mathbb{R}^2$, where $i$ represents the level the order is placed on the order book, $i \in \{-5, \ldots, -1, 1, \ldots, 5\}$, $j$ is the asset, $t$ is the intraday time in for example 10 second time increments and $d$ is the trading day.

4.1. Shape of LOB Volume Profiles

To assess the shape of the LOB volume profile, we develop a graphical representation of the volume on the LOB, highlighting particular empirical features that should be considered when developing statistical models for such stochastic structures. We construct a visualization that we denote as the volume profile for each asset obtained by taking the median of the 10 second volumes for each hourly time increment throughout each trading day of the year 2010. In addition the median volume per level on the LOB, levels 1 - 5, per day across the year of trading was considered. With this information we developed an understanding of the general volume features of the order book, inclusive of depth considerations.

Through a study of the heat map of these intraday volume profiles for the year, several observable features are present. For instance, from Figures 1, 2 and 3 a common feature appears between the 5YTN, BOBL, NIKKEI and SILVER whereby they demonstrate a hump shaped LOB which is contrary to the originally proposed idea of an LOB that is monotonically decreasing away from the best bid and ask (Biais et al. (1995), Bouchaud et al. (2002), Challet and Stinchcombe (2001)), but consistent with the more recent findings from Potters and Bouchaud (2003), Gu et al. (2008) and Chakrabort et al. (2010).

The SP500 and GOLD appears to have monotonically increasing volumes in the first 5 levels of the order book. As shown in Figure 1, the heat chart for BOBL volumes are significantly higher at the start of the year and drop off towards the end of 2010. This feature is also present in NIKKEI and to a lesser extent, the 5YTN and SILVER. The drop in volumes during the mid-year period of 2010 can be attributed to investors trading less with the EU crisis unfolding and Greek bailouts, leading investors to flee from assets such as the BOBL which would have been heavily affected by debt crisis and uncertainty in the EU sector, moving into the perceived security offered by bonds such as the US treasury T-Notes. The SP500 and GOLD volumes tend to be relatively consistent throughout the year, contrary to the clear change in volume profile dynamic throughout the year of 2010 for several of the other key futures assets.

The feature that we observe as common to all assets we studied is the inherent symmetry in the median volumes at each level of the LOB that exists between the bid and ask volumes residing at level one and to some extent higher levels of the LOB. In addition, it is interesting to note that we also observed these basic symmetry relationships between the bid and ask volume profiles at each level also reflected in the volume heat maps for the right tail quantile heat maps for the 75-th percentile and 95-percentiles that we examined. This is a feature that will be explored in detail in the statistical models developed in this paper.

Additionally, we see the attributes of the precious metals of GOLD and SILVER in which the volume profiles of the LOB are far less actively traded compared to the other assets, however both the symmetry properties of the bid and ask are present.

Another feature of the LOB that has been studied in previous literature by (Biais et al. (1995), Challet and Stinchcombe (2001) and Chakrabort et al. (2010)) is the existence of the U-shaped curve (volume smile) for limit orders. We do not see this U-shaped feature appear in any of the assets with exception to the equity derivative, NIKKEI and this is primarily due to the lunch break observed between trading sessions. This suggests the importance of revisiting the findings from the previous studies on LOB structures in the literature on a wide range of assets as carried out in this paper.
Figure 1: Heat maps of the volume for the first 5 levels of the LOB and the median volume on each level of the LOB on the best bid/ask for 2010. **Left Sub-Plots:** Asset - 5YTN; **Right Sub-Plots:** Asset - BOBL.

Figure 2: Heat maps of the volume for the first 5 levels of the LOB and the median volume on each level of the LOB on the best bid/ask for 2010. **Left Sub-Plots:** Asset - SP500; **Right Sub-Plots:** Asset - NIKKEI.
Figure 3: Heat maps of the volume for the first 5 levels of the LOB and the median volume on each level of the LOB on the best bid/ask for 2010. **Left Sub-Plots:** Asset - GOLD; **Right Sub-Plots:** Asset - SILVER

The features observed for the median volume profiles for each level of the LOB on each trading day throughout the year also indicate strongly the importance of considering flexible parametric models to account for wide variations in volume profile both intraday and interday. This will be explored in more depth in the following sections.

4.2. *Descriptive statistics for the LOB Volume Profiles Best Bid and Ask Level 1*

A detailed analysis using several descriptive statistics of the volumes at level 1 of the LOB for 2010 of each asset is developed to provide information on the distributional aspects required for the models considered in this paper. For each asset we calculate the descriptive statistic quantity each day and then take the mean and standard deviation of these quantities across all trading days for the year 2010.

From Table 2, we observe that the mean volume, the variance and the total daily spread of volumes on level 1 of the LOB for 5YTN, BOBL, SP500, NIKKEI, GOLD and SILVER can be large, making a continuous distributional approximation a reasonable choice. Parametric models developed for heavy tails in the continuous case, with respect to the tail index and parameters specifying these distributional forms have a well understood statistical interpretation which will directly inform the statistical attributes of these stochastic volume processes on the LOB. All assets show high levels of positive skewness, with GOLD demonstrating the highest level $7.61 \pm 4.88$ on the bid side and $8.74 \pm 7.04$ on the ask side. The mean level of kurtosis is high for all assets, however the standard deviation of kurtosis may indicate that high kurtosis is not always present in the data. For example, NIKKEI kurtosis is $6.5 \pm 3.9$ on the bid side and similarly for the ask side, $6.5 \pm 3.4$. As discussed in the section above on zero volumes, the volumes for all assets on level 1 of the LOB start at $1, X_{t,d}^0 > 0$. 

9
Table 2: Descriptive statistics for all trading days for 2010, using sub-sample data for level 1 of the LOB

| Asset   | Side | Max   | Min   | Median | Mean  | Std  | Kurtosis | Skew |
|---------|------|-------|-------|--------|-------|------|----------|------|
| 5YTN    | Bid  | 2553.05 | 1.05  | 390.88 | 420.19 | 324.95 | 8.73     | 1.30 |
|         | Ask  | 2514.93 | 1.04  | 399.45 | 424.80 | 322.76 | 7.58     | 1.20 |
| BOBL    | Bid  | 2633.98 | 1.04  | 430.67 | 465.35 | 298.44 | 14.42    | 1.59 |
|         | Ask  | 2606.95 | 1.04  | 429.43 | 462.18 | 293.88 | 14.42    | 1.55 |
| SP500   | Bid  | 4040.79 | 5.01  | 540.34 | 635.55 | 436.54 | 13.30    | 2.00 |
|         | Ask  | 4448.10 | 4.69  | 541.34 | 650.44 | 474.15 | 16.13    | 2.29 |
| NIKKEI  | Bid  | 531.62 | 1.17  | 103.65 | 115.43 | 75.35  | 6.55     | 1.33 |
|         | Ask  | 536.11 | 1.20  | 105.14 | 116.20 | 75.83  | 6.54     | 1.32 |
| GOLD    | Bid  | 172.96 | 1     | 4.62   | 6.78   | 8.66   | 139.65   | 7.61 |
|         | Ask  | 206.85 | 1     | 4.62   | 6.80   | 9.64   | 184.71   | 8.74 |
| SILVER  | Bid  | 79.17  | 1     | 6.56   | 7.83   | 6.39   | 46.77    | 3.63 |
|         | Ask  | 76.02  | 1     | 6.56   | 7.81   | 6.30   | 38.63    | 3.38 |

4.3. Hurst Exponent (Long Memory) for the LOB Volume Profiles

To study empirically the possibility of long memory in the LOB volume profiles at each of the 5 levels of volume on the bid and the ask, we first considered the autocorrelation function which suggested the presence of long memory for all assets. Gu et al. (2008), utilized the Hurst index to test for long memory in the volume of the LOB in a crude manner in which they utilized detrended fluctuation analysis (DFA) to estimate the Hurst index on the 1-min averaged volumes at the first three tick levels on the buy LOB. We implement the Hurst index estimation as detailed below and extend significantly this analysis to a wider range of assets at present results at a 10sec sampling rate for every trading day of 2010, for all five levels of the bid and the ask. Thus providing a richer data analysis with findings more relevant to the analysis in this paper.

Figure 4 shows the Hurst exponents for each intraday volume profile at level 1 to level 5 of the bid and ask in a sequence of box plots comprised of estimates for each trading day of 2010. This provides an “index of dependence” or “index of long-range dependence”, giving a quantitative measure of the relative tendency of a time series to either regress strongly to the mean or cluster. It provides a numerical estimate of the predictability of a time series, which we have disaggregated according to each level of the bid and ask intra-day.

As a guide, values for the index in the range $0.5 < H < 1$ indicate a time series with long-term positive autocorrelation. This indicates momentum in the intraday volume profile, whereby high volume in the series is likely to be succeeded by another high volume period. Values of the Hurst exponent between $0 < H < 0.5$ indicates a time series with long-term switching between high and low volumes in adjacent 10 second time increments. In the analysis performed for each of the assets, we observed a range of results between (0.67, 1) across all 5 levels on the bid and ask side volumes for all market sectors, exchanges and assets which is strongly consistent with what we would expect from data exhibiting long memory. This demonstrates an important aspect of LOB volume profile data that should be considered in developing statistical parametric models for the LOB dynamics that has not previously been explored at this level of detail.

Hence, we see several futures markets in which intraday high frequency statistical features relating to long memory are present. Such features in general financial time series have previously been noted in equity markets by Cont (2009), where it is shown that the stylized feature often present is a “volume/volatility correlation” in which the trading volume will be positively correlated with the market volatility. Again, in the context of equity market discussions, such features are also present in the context of long memory in Lobato and Velasco (2000) where they show that trading volume and volatility show the same type of long memory behavior. We have now detailed such features for a range of different asset classes in the futures markets.
4.4. Empirical Evidence for Heavy Tails on LOB Volumes

There is abundant anecdotal evidence that has been publicized in the media in a number of countries to suggest, prima facie, that volumes in high frequency traded markets will exhibit heavy tailed behavior. We first review some of the anecdotal evidence and follow that by an empirical assessment across a range of markets and assets of the tails of the volume distributions at the best bid and ask. We conclude from our analysis that the LOB volumes typically (but not always) have volumes which are heavier tailed than exponential. We then propose several statistical distribution models that are capable of capturing such heavy tailed behavior.

We start by considering the short term intraday influences that can result in volume profile heavy tail features. To understand this, consider the large sector of self-automated algorithms (algorithmic trading hedge funds) which can generate trades directly on the exchange in micro-seconds often via a direct link to the exchange. For instance in 2006 the London Stock Exchange estimated that more than 40% of all orders were placed by algorithmic trading and this has considerably grown since those early days. In addition, it is well known that American and European markets have higher proportions of algorithmic trading participants of the order of around 80% of market participants in some of these exchanges in 2008. However, it is noteworthy to observe that since the peaks of 2009 through to 2011, many speculate a decline in the volumes and activities of HFT due to regulations currently being discussed for such exchange trading as well as declining profitability. In particular one can observe that the total volume peaked in 2009 and 2010 with 61% and 56%, respectively, of all volume traded in the US stocks to around 51% in 2012. With the dual impact of a decline in volumes and prices, total market capitalization in the US has dropped from $5bil USD in 2009 to around $1.3bil USD in 2012, see discussion in Securities (2012) and Topics (2012).

Due to such automated traders there have being several high consequence and high profile events that have occurred in markets resulting in large swings in price and volumes in short periods of time intraday, therefore affecting the daily volume profile modeling being studied.

Firstly, we note the high profile event in which high frequency algorithmic trading contributed significantly to the Dow Jones Industrial Average’s second largest intraday price and volume shift in history.
on May 6, 2010. During this event, the Dow Jones plunged 1000 points in 20 minutes and promptly rebounding, in the process significantly altering volume profiles in many international exchanges. It is widely acknowledged by both the U.S. Securities and Exchange Commission, the Commodity Futures Trading Commission and the International Organization of Securities Commissions (IOSOC) that a significant portion of the volatility during the 2010 “Flash Crash” resulted from the algorithmic traders, see Lauricella, Mehta, Bowley (2010) and the IOSOC report of 2011.

Other examples include the event known as the “one-minute massacre” of the share price of the global insurer known as QBE Insurance around two years ago, where a robotic trading system shifted the share price from $15.70 AUD to less than a cent before trading was halted Potts (2012). Such an event would significantly alter the volume profile of the stock and therefore indirectly other associated stocks with “correlated” portfolio weights in different sectors and exchanges.

Another recent event occurred on Wall Street which arose due to an algorithmic trading group, who’s algorithmic trading routine recently placed market orders for $440 million USD in an hour and in the process made a significant impact on the share prices and volume profiles of 150 stocks.

In addition to these specific cases there is also the daily activity of many of these algorithmic funds to consider which involves “price discovery” with its simplest form being the process of flooding the market with artificial market and limit orders. This supposedly influences people to respond to the flood of market and limit orders by playing on the momentum driven behavior of many market participants. However, such trades are then canceled within a nanosecond prior to market open, again influencing the volume and price profiles regularly on an intraday time scale.

Whilst these examples help to illustrate the plausibility of the existence of heavy tailed intraday behavior of volume profiles of the LOB in the short term intraday setting, there is also longer term effects to consider for the volume profile of algorithmic trading participants. A fairly recent phenomenon has occurred in order driven markets which involves the shift of non-algorithmic trading participants from traditional exchanges to new private exchanges. There is a growing body of evidence suggesting that brokers are beginning to shift their transactions from traditional markets into private exchanges termed dark pools, where some analysts claim that algorithmic trading has “pushed up to 43 per cent of trades into such exchanges”, see Potts (2012). Though we do not address this directly in this analysis, it is a hypothesis that may be explored and related to the heavy tailed features under study in this paper.

An understanding of these features is critical to the study of the stochastic processes that make up the LOB structure. For example a statistical study of the tail attributes of the LOB volume profiles will help to characterize the behavior of the volume, trading and market depth attributes in times of dislocation in the markets. This knowledge can then be incorporated into the statistical models constructed to capture the LOB features, either intraday or interday or both. Alternatively, if systematic evidence of sub-exponential tails is present regularly, then perhaps there is something more fundamental occurring on the exchange mechanism related to processing/order filling rules, or the manner in which algorithmic trading strategies exploit such rules for executions when arbitrage opportunities present themselves.

The literature on LOB modeling to date has considered only simple classes of two parameter shape-scale models for the statistical modeling of LOB data for price or volume. In this section we demonstrate the need for more sophisticated, flexible parametric models. To motivate the use of such flexible parametric models, we initially consider basic shape-scale parametric models as a benchmark model.

For all assets in Table 1 across all time sampling rates for each trading day, the normal distribution when fitted to both log transformed data and Box-Cox transformed data provided, in all instances a poor

\[1\text{http://www.smh.com.au/money/investing/share-wars-how-the-robots-are-robbing-you-20120825-24t4t.html#ixzz241AJB3bd}\]
fit. Papers by Bouchaud et al. (2002) and Gu et al. (2008) consider the distributional features of the LOB volumes and conclude that a two parameter shape-scale model given by a Gamma distribution is most suitable. To obtain the estimators of the Gamma distribution, we equated the population moments with the sample moments (moment matching). We found that the gamma distribution provides a better fit than the normal distribution. To assess the stability of the parameters and to assess how well the Gamma distribution represents the skewness and kurtosis in the data, via moment matching, we estimated the parameters for all assets, every time segment across an entire year of trading. From these parameters we estimated the mean, variance (which should be consistent with the sample estimates), skewness and kurtosis. For all assets, the Gamma distribution provided a poor estimate of skewness and kurtosis. The results presented in Table 2 demonstrated the high levels of positive skewness and kurtosis for all assets.

To further emphasize the point that it is in fact, the right tail attributes of the volume profile that is driving this behavior, we present Figures 5, 6, 7,(Technical Appendix A.14, A.15, A.16), the QQ-plot for the LOB data relative to a Generalized Pareto Distribution with a tail index of $\gamma = 0$, making this a comparison between the right tail of the empirical CDF and the exponential distribution right tail. This therefore acts as a comparison between a ‘medium sized tail’ and allows one to identify a relatively ‘fat-tailed’ distribution. If the empirical CDF lie on the ‘dashed’ straight line then the volume profile is consistent intra-daily with an exponential distribution, however, the presence of a concave relationship indicates a fat-tailed distribution in the sub-exponential class. The results presented are estimated intra-daily for every 25th trading day of the year (as an illustration of the general results we observed consistently for each trading day of the year) and for each of the 5 levels of the LOB on the bid and ask sides, for each asset. The majority of these plots are provided in the Technical Appendix with selected examples presented in the main text.

The findings for the results in the QQ-plots comparing the empirical CDF to the exponential model, demonstrate several interesting features. Firstly, starting with the 5YTN, in general at all levels of the volume profile there is a convex relationship relative to the exponential quantiles, indicating light tails for these profiles, see Figure 5. However, occasionally on some days there is a clear evidence for a concave relationship between the empirical CDF and the quantiles of the exponential distribution, indicating the existence of a power law relationship. To help distinguish these days, we have emphasized examples of these particular trading days with a thicker solid line on the QQ plots. It is also clear from this analysis that there is a stronger tendency for power law relationships for the right tail of the volume profile in the ask side, relative to the bid side for the 5YTN. The BOBL also has occasional trading days which indicated the presence of heavy right tails for the intraday volume profile indicative of a power law relationship. In this case it is clear that there is a stronger tendency for such heavy tailed features to occur on all 5 levels of the bid side throughout 2010, as opposed to the ask side which indicates far fewer examples of power law tails, see Figure A.14. The NIKKEI in Figure A.16 demonstrated occasional concave relationships for the right tail of the volume profile for example on level one of the ask and level 3 of the bid and ask.

For the SP500 it is apparent that the power law relationship in the tails is prominent more often in the volume profile at all levels 1 to 5 in both the bid and ask sides see Figure A.15. The attributes of futures contracts on GOLD indicated strongly the presence of heavy tailed relationships on every trading day analyzed on all levels of the order book for the bid and ask. SILVER similarly had consistent evidence of heavy tailed relationships in the right tail of the intraday volume profile.

We also consider two further tools for assessing and motivating the need to consider heavy tailed behavior which are based on the mean excess plot and the Hill plot, see Kratz and Resnick (1996). As with the QQ-plot analysis we present these results as estimated intra-daily for every 25th trading day of the year and for each of the 5 levels of the LOB on the bid and ask sides, for each asset. The sample mean excess function defined by Equation 1, represents the sum of the excesses over a threshold $u$ divided by the number of data points which exceed the threshold $u$. It approximates the mean excess function.
describing the expected exceedence amount for a particular threshold $u$ given an exceedence in the volume profile has occurred. If the empirical mean exceedence function estimate has a positive slope for large thresholds $u$ then this indicates that the observed volume profile data is consistent with a Generalized Pareto Distribution with a positive tail index parameter Beirlant et al., Chapter 1. The sample mean excess is given by,

$$e_n(u) = \frac{\sum_{i=1}^{n} (X_i - u) I\{X_i > u\}}{\sum_{i=1}^{n} I\{X_i > u\}}. \quad (1)$$

which estimates the conditional expectation $e(u) = \mathbb{E}[(X - u)|X > u]$.

Results for this analysis are provided in Technical Appendix B and we provide some illustrations in the main text. In Figure 8 we observe the Mean Excess plot versus threshold for the 5YTN. It indicates a clear upward trend as the threshold increases for all of the trading days explored on the bid at level 1, again consistently indicating the presence of heavy tailed power law relationships in the volume profile. At level 1 of the ask there is a mix between evidence for some days having heavy tailed attributes in the right tail of the volume profile and other trading days with lighter tails. This is also present throughout the other levels of the LOB on the bid and ask. The results for the BOBL and the SP500 in terms of the mean excess plots in Figures B.17-B.18 demonstrate on several of the trading days a strong indication of power law relationships in the right tail of the volume profile. SP500 this is very pronounced at level 1 of both the bid and ask. The NIKKEI results in Figure B.19 also indicate the occasional presence of power law right tail. As expected from the QQ plots, the GOLD and SILVER futures indicate strong power law relationships consistently in the intraday volume profiles on the majority of trading days presented.

Figure 5: 5YTN: Quantiles of for exponential distribution model versus sample order statistics for intra-daily volume data every 25th trading day of 2010. **Top Row Bid** from left to right is Level 1 to Level 5 of LOB; **Bottom Row Ask** from left to right is Level 1 to Level 5 of LOB
The Hill-plot was developed by Hill in 1975, and is based on an estimate of the inverse tail index for the Generalized Extreme Value (GEV) model (specified in Definition 3), under \( n \) observations for a number of exceedences \( k \), given in Equation 2.

\[
\hat{\gamma}^{-1} = \frac{1}{k-1} \sum_{i=1}^{k-1} \ln X_{(i,n)} - \ln X_{(k,n)}, \quad \forall k > 2.
\]  

where the first order statistic is the minimum for \( n \) observations, that is, \( X_{(1,n)} = \min\{X_1, \ldots, X_n\} \). Similarly, the \( n \)th order statistic is the maximum in the set of \( n \) observations, that is, \( X_{(n,n)} = \max\{X_1, \ldots, X_n\} \).

In addition we provide analysis for the Hill plot at each level of the LOB on the bid and ask sides. The Hill plot represents the estimated inverse tail index as a function of the upper order statistics \( k \).
This plot provides feedback on the suitability of the selected threshold utilized in the estimation of the Extreme Value Theory (EVT) models (see Section 4.5.2), in particular the Peaks Over Threshold (POT)’s method (see Technical Appendix E.1 and Beirlant et al. (2004) for details) used for the Generalized Pareto Distribution (GPD) model (specified in Definition 4).

The results for the Hill plot analysis for each asset along with 95% confidence intervals are detailed in the Technical Appendix C. As discussed in McNeil et al. (2005, Chapter 7) one can interpret the Hill plot easily in an ideal scenario in which the observational data is assumed i.i.d. and follows a law with a regularly varying tail. In such situations the Hill estimator can be an accurate estimator of the tail index of the EVT model. Therefore in the Hill plot one studies the Hill estimates for the (inverse) tail index versus the data order statistics. If a stable region is found where the estimates constructed from a small number of order statistics give broadly comparable estimates, then there is some confidence in the estimate of the tail index. The conclusion of this analysis is that for several days both with heavy tailed and light tailed volume profile behavior, the estimated tail index for the EVT models will be accurate.

Figure 8: 5YTN: Mean Excess plot for intra-daily volume data every 25th trading day of 2010. Top Row Bid from left to right is Level 1 to Level 5 of LOB; Bottom Row Ask from left to right is Level 1 to Level 5 of LOB.
Capturing these features when building statistical models for LOB volumes has not been achieved in previous studies using basic shape-scale parametric models. The purpose of this section is to formally study the features of the volume profiles of level 1 bid and level 1 ask in the LOB for futures markets using more flexible families of parametric models. We firstly, assess the appropriateness of statistical models and fit (at different sampling frequencies) for the $\alpha$-stable family of distributions. Then we also consider the appropriateness of a heavy tailed assumption for the volume profiles each day as captured by considering evidence of suitability of sub-exponential family models for volume profile tails given by: GPD and GEV distributional families.

In the processes of estimating these models for the LOB volume profile data, we also assess, study and recommend sophisticated statistical estimation procedures and their performance for each model for practitioners when estimating such models for the large volume of data in a LOB model analysis which included: Maximum Likelihood Estimations (MLE); Generalized Method of Moments and L-Moment Estimation; Mixed methods of MLE and Generalized Moment Matching; Empirical Percentile Estimation and Quantile based estimators. The details of these estimation approaches for each model are presented in Section 5.

4.5. Sub-exponential Models for the Right Tail of the LOB Volume Profile

In this section we question whether sub-exponential tails, as defined in Definition 1 are suitable candidates for the extremes of the volume profiles on the level 1 bid and ask across exchanges and across futures asset sectors. The motivation for this analysis is derived from the characteristics of LOB volume profiles which are directly affected by market participants on the different electronic exchanges and their trading approaches. We hypothesize that although high frequency traders are argued to be capable of increasing liquidity in markets, they are also capable of drastically changing volume profiles and price profiles in the LOB over short periods of time and over long periods of time, thus influencing the tail features of the volume profile.

We note that the following analysis regarding the presence of extreme value tails is often taken under the assumption that the data is i.i.d. As in most real data settings, we will not have this assumption satisfied and indeed we have shown this in our empirical analysis. To test the effect this would have on the
estimators for the tail features, we also performed inference with lower sampling rates at which the auto-
correlation in the time series and the long memory was not a problem and samples were approximately
independent at 1min, 10min and 1hr sampling rates. We studied both synthetic and real data examples
of the bias-variance trade-off between the variance in the parameter estimates due to significantly reduced
sample sizes at lower sampling rates compared to the bias introduced at higher sampling rates due to the
assumption of independence being violated. From the findings we decided to proceeded with our analysis
at higher sampling rates due to improved precision in the estimation of the tail parameters, otherwise
sample sizes were too small in uncertainty in such estimates was deemed too large.

Definition 1 (Sub-exponential Distributions). Let $X_1,\ldots,X_n,\ldots$ be independent positive random
variables with distribution $F(x) = \mathbb{P}(X_k < x)$, $\forall k \in \{1,2,\ldots,n,\ldots\}$. Then the class of sub-exponential
distributions ($F(x) \in \mathcal{F}$) satisfy the limits

$$
\lim_{x \to \infty} \frac{1 - F^{\alpha}(x)}{1 - F(x)} = n, \quad \text{iff} \quad \lim_{x \to \infty} \frac{1 - F^{2\alpha}(x)}{1 - F(x)} = 2
$$

where, $F^{2\alpha}(x)$ is the convolution of $F$ with itself, defined using Lebesgue-Stieltjes integration, by:
$\mathbb{P}(X_1 + X_2 \leq x) = F^{2\alpha}(x) = \int F(x - y)dF(y)_{-\infty}^{\infty}$. The n-fold convolution $F^{n\alpha}(x)$ is defined in the same way.

It is well known that the sub-exponential family of distributions $\mathcal{F}$ defines a class of heavy-tailed models that we consider for volume profiles, see conditions for membership of distributions in this family as classified in Pitman (1980). The tail distribution function $\overline{F}(x)$ is defined by the survival probability $\overline{F}(x) = 1 - F(x)$. The class $\mathcal{F}$ includes all severity models in which the tail distribution under the log transformed r.v., $\overline{F}(\log(x))$, is a slowly varying function of $x$ at infinity. That is the limit in the right tail of the ratio

$$
\lim_{x \to \infty} \frac{\overline{F}(\log(kx))}{\overline{F}(\log(x))} = 1, \quad \forall k > 0.
$$

As a result of this slow variation of the transformed random variable, it also therefore holds that for $\alpha > 0$, $\lim_{x \to \infty} x^{\alpha}\overline{F}(\log(x)) \to \infty$, and it follows that,

$$
\lim_{x \to \infty} \frac{e^{-\alpha x}}{F(x)} = 0,
$$

resulting in the naming of this class as the “sub-exponential” distributions. We begin with assessment
of heavy tailed features of the volume profile LOB data by considering the subexponential sub-family of
$\alpha$-Stable, GPD and GEV families of models. Each is a well-known parametric family of models that are
members of the sub-exponential class of distributions.

4.5.1. Alpha-Stable Models

Models constructed with $\alpha$-Stable distributions possess several useful properties, including infinite vari-
ance, skewness and heavy tails (Zolotarev (1986); Alder et al. (1998); Samorodnitsky and Taqqu (1994);
Nolan (2007)). Considered as generalizations of the Gaussian distribution, they are defined as the class
of location-scale distributions which are closed under convolutions. $\alpha$-Stable distributions have found application in many areas of statistics, finance and signal processing engineering as models for impulsive,
heavy tailed noise processes (Mandelbrot (1960); Fama (1965); Fama and Roll (1968); Nikias and Shao
(1995); Godsill (2000); Melchiori (2006), Peters et al. (2010); Peters et al. (2011), Peters et al. (2012) and
Gu et al. (2012)). Here we consider development of this class for the modeling of the volume profiles in a
LOB stochastic process.
The univariate \(\alpha\)-Stable distribution is typically specified by four parameters, see (Levy 1924). The tail index \(\alpha \in (0, 2]\) determining the rate of tail decay, \(\beta \in [-1, 1]\) determining the degree and sign of asymmetry (skewness), \(\gamma > 0\) the scale (under some parameterizations) and \(\delta \in \mathbb{R}\) the location. The parameter \(\alpha\) is typically termed the characteristic exponent, with small and large \(\alpha\) implying heavy and light tails respectively. Gaussian \((\alpha = 2, \beta = 0)\) and Cauchy \((\alpha = 1, \beta = 0)\) distributions provide the only simple analytically tractable sub members of this family. In general, as \(\alpha\)-Stable models admit no closed form expression for the density which can be evaluated pointwise (excepting Gaussian and Cauchy members), inference typically proceeds via the characteristic function. Hence, we begin by defining the family of univariate \(\alpha\)-Stable models that we consider for the modeling of the volume profiles for the level 1 bid and ask.

**Definition 2 (\(\alpha\)-Stable).** A random variable \(X\) is stable if and only if \(X \overset{d}{=} aZ + b\), where \(0 < \alpha \leq 2, -1 \leq \beta \leq 1, a > 0, b \in \mathbb{R}\) and \(Z\) is a random variable with characteristic function

\[
\mathbb{E}[\exp(i\theta Z)] = \begin{cases} 
\exp \left(-|\theta|^\alpha \left[1 - i\beta \tan \frac{\alpha \pi}{2} \left(\text{sign}(\theta)\right)\right]\right) & \alpha \neq 1, \\
\exp \left(-|\theta|^\alpha \left[1 + i\beta \frac{2}{\pi} \left(\text{sign}(\theta)\right) \log |\theta|\right]\right) & \alpha = 1.
\end{cases}
\]

where \(\text{sign}(u) = -1\) if \(u < 0\), \(\text{sign}(u) = 0\) if \(u = 0\) and \(\text{sign}(u) = 1\) if \(u > 0\).

\(\alpha\)-Stable distributions provide no general analytic expressions for the density, median, mode or entropy, but are uniquely specified by their characteristic function, which has several parameterizations. In this paper we consider the following parameterization in which a random variable \(X\) is said to have a stable distribution, \(S_{\alpha}(\beta, \gamma, \delta; 0)\), if its CF has the following form:

\[
\mathbb{E}[\exp(i\theta X)] = \begin{cases} 
\exp\{-\gamma^\alpha |\theta|^\alpha (1 + i\beta(\text{sign}(\theta)) \tan \left(\frac{\alpha \pi}{2}\right) |\gamma|^{1-\alpha} - 1)\} + i\delta \theta & \text{if } \alpha \neq 1 \\
\exp\{-\gamma |\theta|(1 + i\beta \frac{2}{\pi}(\text{sign}(\theta)) \log |\gamma|^{1-\alpha})\} + i\delta \theta & \text{if } \alpha = 1.
\end{cases}
\]

In addition, we note in Technical Appendix Definition 7 the representations for the density and cdf available as series expansions, provided in Zolotarev (1983). Through these specifications of the \(\alpha\)-stable distributional model, the evaluation and estimation of this family of models for the volume profiles can be performed, even when the distribution function is not available in closed form to evaluate pointwise.

4.5.2. Extreme Value Models

Here we present the characterization of the Extreme Value Theory (EVT) families of parameter models considered for the modeling of the right tail of the volume profiles on the level 1 bid and ask. In particular we focus on the well-studied and widely utilized families of the Generalized Extreme Value (GEV) model and the related Generalized Pareto Distribution (GPD) model.

The parameterization considered for the GEV model is given in Definition 3 below. The GEV model allows one to capture the extreme volume profile events (tail events). The GEV model will represent the modeling of the extreme volume in the LOB level 1 bid or ask volume profile over a specified intraday or interday period.

It is also useful to note the definition for distributions of the same type as those that differ only in location and scale. Using these definitions, one can show that any EVT distribution can be represented according to the following distribution known generically as the Generalized Extreme Value (GEV) distribution that we denote by \(H\) and specified in Definition 3, see Beirlant et al..
Definition 3 (Generalized Extreme Value Distribution (GEV)). The generalized extreme value (GEV) distribution is defined by

\[ P(X < x; \mu, \sigma, \gamma) = \exp \left\{ -\left[ 1 + \gamma \left( \frac{x - \mu}{\sigma} \right) \right]^{-1/\gamma} \right\}, \]

for \( 1 + \gamma(x - \mu)/\sigma > 0 \), where \( \mu \in \mathbb{R} \) is the location parameter, \( \sigma > 0 \) the scale parameter and \( \gamma \in \mathbb{R} \) the shape parameter. Furthermore, the density function is given by,

\[ f(x; \mu, \sigma, \gamma) = \frac{1}{\sigma} \left[ 1 + \gamma \left( \frac{x - \mu}{\sigma} \right) \right]^{-(-1/\gamma)-1} \exp \left\{ -\left[ 1 + \gamma \left( \frac{x - \mu}{\sigma} \right) \right]^{-1/\gamma} \right\}, \]

again for \( 1 + \gamma(x - \mu)/\sigma > 0 \). In addition the support of a random variable \( X \sim H_\gamma \left( \frac{x-\mu}{\sigma} \right) \) is given by

\[ S_X = \begin{cases} [\mu - \sigma \frac{\gamma}{\gamma}, \infty], & \gamma > 0, \\ [\infty, \infty], & \gamma = 0, \\ [-\infty, -\mu - \sigma \frac{\gamma}{\gamma}], & \gamma < 0, \end{cases} \]

The derivation and estimation of the GEV model parameters involves a block maximum based analysis with its associated estimation procedures and properties, discussed in Section 5.2 and see Bensalah (2000) and Embrechts et al. (1999). An alternative specification of such EVT models is the Exceedences over a threshold (Peaks Over Threshold POT) based formulation which results in the GPD model and alternative data preparation and parameter estimation procedures, see Section Appendix E.1 and Beirlant et al. (2004) for details. The GPD model is given in Definition 4 below.

Definition 4 (Generalized Pareto Distribution (GPD)). A random variable \( X \sim GP(\gamma, \sigma) \) to have a distribution and density (conditional upon translation to the origin - location parameter \( \mu = 0 \)) given by

\[ F_X(x; \gamma, \sigma) = P(X < x | X \geq \mu) = \begin{cases} 1 - (1 + \frac{\gamma x}{\sigma})^{-\frac{1}{\gamma}}, & \gamma \neq 0 \\ 1 - \exp \left( -\frac{x}{\sigma} \right), & \gamma = 0, \end{cases} \]

\[ f_X(x; \gamma, \sigma) = \frac{\sigma^{\frac{1}{\gamma}}}{(\sigma + \gamma x)^{\frac{1}{\gamma}+1}} \]

with shape parameter \( \gamma \in \mathbb{R} \) and scale parameter \( \sigma \in (0, \infty) \). In addition the support of the density is given by

\[ S_X = \begin{cases} [\mu, \infty), & \gamma \geq 0, \\ [\mu, \mu - \sigma \frac{\gamma}{\gamma}], & \gamma < 0, \end{cases} \]

5. Parameter Estimation for Sub-exponential Models of LOB Volume Profiles

In this section we briefly detail the parameter estimation procedures for the proposed models. In particular we provide details of the less well known robust approaches and detail all other procedures in the Technical Appendix D and Technical Appendix E.

There are numerous approaches one can adopt for fitting heavy tailed and flexible families of models such as the GPD, GEV and \( \alpha \)-stable families. Each approach will have different merits related to statistical
efficiency, bias and variance trade-offs and importantly for the setting of analysis of LOB data (massive data sets) the computational robustness and efficiency. For the numerous statistical estimation approaches, each have different bias, variance, sample size behavior and computational efficiency when it comes to the massive LOB volume profile data we assess in this paper.

Note, though the GEV and GPD can be represented in certain cases as reparameterizations, the actual data preparation and estimation procedures for such models differs significantly and is the reason why we investigate each class carefully. For example under the GPD model estimation approach involving a Peaks Over Threshold (POT’s) procedure the amount of discarded data differs to that of the Block Maxima approach of the GEV model estimation, see discussion in Beirlant et al. (2004).

Table 3 provides a summary of the methods utilized for each of the distributions. This table shows the distributions and methods used for volume data on level 1 of the LOB for each of the six assets: 5YTN, BOBL, SP500, NIKKEI, GOLD and SILVER. For each asset we consider varying time resolutions of 1 second, 2 seconds, 5 seconds and 10 seconds across each trading day in 2010. The discussion that follows will provide insight into the dynamics of the parameters over time, varying resolutions, different estimation methodologies and associated implementation issues. A diagramatic representation of results will be provided for each distribution for one asset only, with full results available for all assets in Technical Appendix H, Technical Appendix I, Technical Appendix J, Technical Appendix K, Technical Appendix L and Technical Appendix M. All technical details relating to the processing of the data and the estimation equations and numerical procedures undertaken can be obtained in the following subsections.

To assess the quality of the statistical model estimations, we considered a number of approaches. The first involves an exploration of the Goodness-of-fit utilizing the two-sample Kolmogorov-Smirnov (KS) statistical test which aims to test the compatibility of the theoretical probability distribution with the empirical probability distribution. Since our goal is not to specifically propose one particular statistical model in preference to the others, rather we intend to demonstrate that for a range of heavy tailed models and associated robust statistical estimation procedures for the parameters of these models, there is overwhelming evidence of intraday and interday heavy tailed behavior which varies over time in its magnitude. With this in mind, we note that when considering results of the KS test, the classical KS test is limited in that it is only weakly sensitive to the quality of the fit in the tails of the tested distribution, as investigated in Chickeportiche and Bouchaud (2012). Therefore, in the context of the distributions, \(\alpha\)-stable, GEV and GPD, the tail events are the primary consideration, not the central distribution. As a consequence we performed several different analysis and the results of these studies are provided in the Technical Appendix H - Technical Appendix M showing the KS test statistics for each asset for each trading day of 2010 under the intra-daily estimation of the GPD, GEV and \(\alpha\)-stable models per day.

Under the null hypothesis, the KS test quantifies a distance between the empirical distribution function and the theoretical distribution function for the considered model. The results of the \(\alpha\)-stable analysis showed that the mean test statistic across all trading days for the bid side is 0.0540 and for the ask side, 0.0542. This results in the rejection of the model fit at a significance level of 10% for 90% of trading days. This is however known to be misleading results obtained from this test due to the fact that one ought to adjust for the fact that infinite mean and infinite variance models are considered. That is sub-exponential
models will require an adjustment to the distribution of the test statistic to account for the significant contribution in the tails see discussion in Chickeportiche and Bouchaud (2012).

Furthermore, we note that this analysis is further complicated by the massive data sets obtained for all intra-daily samples at 10 second time increments for each trading day, resulting in each hypothesis being formed from thousands of samples. The larger the data set, the higher the chance of rejecting the null hypothesis. Since the test is not correctly attributing appropriate weights to the sub-exponential tails of the model, the test will incorrectly reject the null as the sample increase at a disproportionate rate to reality. This is due primarily to the test expecting an exponential rather than a power law decay in the tail probabilities, see detailed discussion on this and possibly more sophisticated alternative tests Koning and Peng (2008). Therefore we instead utilize the KS test statistic merely as a guide to the relationship between the parametric model cdf and the empirical cdf. Future studies will consider the generalized version of the KS test discussed in Chickeportiche and Bouchaud (2012), but for now we proceed with caution when discussing the fit of tail distributions in relation to the tail events under consideration.

5.1. Statistical Estimation of $\alpha$-Stable Distribution Model Parameters

The estimation of parameters for the $\alpha$-Stable class has a long history with a variety of methods being proposed over the years. The most popular of these parameter estimation procedures includes the transformation approach of Zolotarev where one maps from a characteristic parameterization in $(\alpha, \beta, \gamma, \delta)$ to the $(\nu, \eta, \tau)$ parameterization. This is advantageous due to the fact that logarithmic moments have simple expressions in terms of parameters to be estimated. There is also the generalized method of moments approach and numerous approaches involving matching the empirical characteristic function with the true characteristic function point-wise, see discussion on these approaches in Peters et al. (2009).

The choice adopted in this paper is based on the approach of McCulloch (1998) which was selected for two reasons: firstly, the approach is known to be robust and efficient computationally, a very important requirement when processing the massive amount of data studied in the LOB structures in this paper. Secondly, the consistency and bias properties of this estimation approach have been studied and so it provides a reliable means of estimation of the $\alpha$-Stable model parameters.

To undertake the estimation, the preparation of the volume profile data for fitting the $\alpha$-Stable model involved taking intra-daily data for each asset over the period 2010 and considering varying time resolutions of 1 second, 2 seconds, 5 seconds and 10 seconds across each trading day in 2010. To each of these data sets per day the following estimation procedure was applied.

The four parameters in the $\alpha$-Stable model under the parameterization presented were determined from a set of five pre-determined quantiles for the parameter ranges $\alpha \in [0, 2.0]$, $\beta \in [-1, 1]$, $\gamma \in [0, \infty)$ and $\delta \in \mathbb{R}$ as detailed in McCulloch (1986). The estimators adopted were similar to those for the symmetric stable setting of Fama and Roll (1971) and were developed for the general asymmetric setting after removing the asymptotic bias, in McCulloch (1986).

The volume profile data is sorted in ascending order in which the $i$-th order statistic is denoted by $x_{(i,n)}$. The method proposed by McCulloch (1986) and McCulloch (1998) estimates the model parameters based on sample quantiles, while correcting for estimator skewness due to the evaluation of the sample quantiles $\hat{q}_p(x)$, the $p^{th}$ quantile of $x$ from the order statistics of the sample. The stages of this estimation involve the following:

1. **Obtain a finite sample consistent estimator of quantiles**: with the $x_{(i,n)}$ arranged in ascending order, the skewness correction is made by matching the sample order statistics with $\hat{q}_{s(i)}(x)$ where $s(i) = \frac{2i-1}{2n}$. Then a linear interpolation to $p$ from the two adjacent $s(i)$ values is used to establish $\hat{q}_p(x)$ as a consistent estimator of the true quantiles. This corrects for spurious skewness present in finite samples and $\hat{q}_p(x)$ is a consistent estimator of $q_p(x)$.
2. Obtain estimates of tail index \( \alpha \) and skewness parameter \( \beta \): in McCulloch (1986) two non-linear functions of \( \alpha \) and \( \beta \) are provided in terms of the quantiles as detailed in Equation 11.

\[
\nu_\alpha = \frac{q_{0.95}(\cdot) - q_{0.05}(\cdot)}{q_{0.75}(\cdot) - q_{0.25}(\cdot)}, \quad \nu_\beta = \frac{q_{0.95}(\cdot) + q_{0.05}(\cdot) - 2q_{0.5}(\cdot)}{q_{0.95}(\cdot) - q_{0.05}(\cdot)}.
\]

One can therefore estimate these quantities \( \hat{\nu}_\alpha \) and \( \hat{\nu}_\beta \) using the sample estimates of the quantiles \( \hat{q}_p(x) \). Now to obtain the actual parameters estimates \( \hat{\alpha} \) and \( \hat{\beta} \) one must numerically invert the non-linear functions \( \nu_\alpha \) and \( \nu_\beta \), this can be done efficiently through a look up table provided for numerous combinations of \( \alpha \) and \( \beta \) and provided in tabulated form in McCulloch (1986).

3. Obtain estimates of \( \gamma \) given estimates of \( \alpha \) and \( \beta \): in McCulloch (1986) a third non-linear function which is explicit in \( \gamma \) and implicit in \( \alpha \) and \( \beta \) is provided in terms of the quantiles as detailed in Equation 12.

\[
\nu_\gamma(\alpha, \beta) = \frac{q_{0.75}(\cdot) - q_{0.25}(\cdot)}{\gamma},
\]

an estimate then follows given \( \hat{\alpha} \), \( \hat{\beta} \) and consistent sample quantiles \( \hat{q}_{0.75}, \hat{q}_{0.25} \).

5.2. Statistical Estimation of the Generalized Extreme Value (GEV) Model

The estimation of parameters for the GEV family has been studied under numerous approaches including the asymptotically efficient MLE approach for \( \gamma < 0.5 \) in Prescott and Walden (1980) and Smith (1985) and the Method of L-Moments from Hosking (1990a). There have been numerous studies theoretical and empirical studies looking at the asymptotic and small sample properties of these estimation techniques with and without the presence of parameter constraints.

In Hosking (1985) it was demonstrated that the MLE parameter estimators for the GEV model can be unstable and this will depend on the sample size. As a consequence, the probability weighted moment (PWM) approach was proposed. It was discussed later in Hosking (1990b) that PWM for the GEV model is equivalent to the L-Moments approach which each involves a linear combination of order statistics, see detailed discussion on L-Moments in Hosking (1990b).

In addition one can show that the L-moments have several theoretical advantages over ordinary moments, especially in the case of sub-exponential models, since for the L-moments of a probability distribution to be meaningful one only requires that the distribution have finite mean and no higher-order moments need be finite. The standard errors of the L-moments are finite if the distribution has finite variance. Although moment ratios can be arbitrarily large, sample moment ratios have algebraic bounds see Dalén (1987) and consequently the sample L-moment ratios can be shown to take any values that the corresponding population quantities can. The implications of this manifest themselves in numerically stable and robust L-skewness and L-kurtosis estimators which can be used directly in L-Moments estimation. It has also been observed by several authors Royston (2006) and Vogel and Fennessey (1993) that the L-moments are less sensitive to outlying data values. There has being recent developments of mixed methods for inference in such models. These developments combine attributes of each approach which include computational efficiency, statistical accuracy and robustness. In this section we develop one such mixed approach combining MLE and L-Moments.

The preparation of the volume profile data for fitting the GEV extremal value distribution model involved taking intra-daily data for each asset over the period 2010 and considering varying time resolutions of 1 second, 2 seconds, 5 seconds and 10 seconds across each trading day in 2010. These trading days were then split into blocks and a Block maxima approach was adopted in which the maximum volume submitted in the sub-sample time block is recorded, producing a set of \( K \) ordered realized observations \( \{x_{(1,K)}, \ldots, x_{(K,K)}\} \). Comparing this to the \( \alpha \)-Stable preparation of the data with \( K \) samples, we see the
key differences are that we use the maximum volume rather than the last volume recorded in the specified
time increment when constructing the observed time series data.

In fitting the GEV model, we utilized combined knowledge of the bias-variance properties of moment
based estimators and MLE estimators for this family of extreme value distributions. The approach adopted
is therefore known as a mixed approach combining MLE and L-moments based estimation of Morrison
and Smith (2002). There are several ways this can be undertaken for the estimation of the GEV model
parameters $\mu, \sigma, \gamma$, however all approaches aim to incorporate the individual statistical features of efficiency,
consistency and unbiasedness of the moment and MLE based estimators. The concept is to combine the
features of each estimation approach to tackle scenarios for different data sizes, where the asymptotic
assumptions may not be suitable to justify distributional properties of the estimators, which according to
numerical results discussed in the literature can require very large sample sizes.

To present the mixed MLE and L-moments based approach considered, we first define the L-moments,
given by Hosking (1990a), for the real valued random variable $X$ with distribution $F(x)$ and quantile
function $Q(p)$ according to Definition 5.

**Definition 5 (L-Moments).** The Population L-moments of a real valued random variable $X \sim F(x)$, for which there is a $K$ sample realization with order statistics given by $X_{(1,K)} \leq X_{(2,K)} \leq \ldots \leq X_{(n,K)} \leq \ldots \leq X_{(K,K)}$, is defined according to

$$\lambda_r = r^{-1} \sum_{k=0}^{r-1} (-1)^k C_k^{r-1} \mathbb{E} [X_{(r-k,r)}], \forall r = 1, 2, \ldots,$$

(13)

where the expectation of the $r$-th order statistic from a sample of size $K$ is given by David and Nagaraja
(1970, page 33) as,

$$\mathbb{E} [X_{(r,K)}] = \frac{K!}{(r-1)!(K-r)!} \int x F(x)^{r-1} (1 - F(x))^{K-r} dF(x).$$

(14)

Here $\lambda_r$ is a linear function of the expected order statistics.

**Remark 1.** It turns out that a consequence of this definition of L-moments is that we can easily find
estimators of these quantities based on an observed sample of data, simply by using linear combinations of
the ordered data values.

From the population L-Moment estimators and the fact that one can then obtain closed form expressions
for the estimators of the L-moments of the GEV model directly in terms of the model parameters. These
are conveniently expressed below through the first and second population L-Moments and the population
L-skewness to obtain the L-Moment GEV estimators given in Definition 6.

**Definition 6 (L-Method of Moments for GEV).** The population L-Method of Moment expressions
for the parameters of the GEV distribution are given by solutions, with respect to the L-skewness, by the
system of equations

$$\frac{1 - 3\hat{\gamma}}{1 - 2\hat{\gamma}} = \frac{\hat{\lambda}_3 + 3}{2},$$

$$\hat{\sigma} = -\frac{\hat{\lambda}_2 \hat{\gamma}}{(1 - 2\hat{\gamma}) \Gamma(1 - \hat{\gamma})} \quad \text{and} \quad \hat{\mu} = \hat{\lambda}_1 + \frac{\hat{\sigma}}{\hat{\gamma}} [1 - \Gamma(1 - \hat{\gamma})],$$

(15)

where $\hat{\lambda}_1$, $\hat{\lambda}_2$ and $\hat{\gamma}$ are the estimators of the first two L-Moments and the L skewness obtained
from the sample.
The asymptotic properties of these estimators are discussed in Morrison and Smith (2002).

**Remark 2.** The bias of the LMOM estimates of the shape parameter of the GEV distribution $\gamma$ increases with a decrease in the magnitude of $\gamma$ and can be significant. In contrast, the MLE method produces an almost unbiased estimate of $\gamma$ for small samples, however the variance of the MLE estimates is larger than those obtained from the biased LMOM approach. In addition, it has been reported in Martins and Stedinger (2000) that the MLE estimates of the shape parameter can regularly, in practice produce unrealistic estimates of $\gamma$, resulting in very large errors in the quantile estimates. The mixed type methods aim to overcome these difficulties.

In practice for such a mixed approach the parameter range of the extreme value index is restricted to $\gamma \in [-0.5, 0.5]$ to ensure the moments are finite and appropriate regularity conditions for the MLE are satisfied, see discussion on such items in the GPD setting and the GEV setting in Beirlant et al. (2004, Chapter 5). However, this can be relaxed through the consideration of the truncated Trimmed L-Moments (TL-Moments) approach. We focus on the simplest mixed approach based on the MLE for $\gamma$ and L-Moments for $\mu, \sigma$ detailed for a sample size of $K$ according to the following stages.

**Stage 1:** Reparameterize the GEV model likelihood in terms of only the Extreme Value Index $\gamma$. This is achieved by expressing the parameters $\mu$ and $\sigma$ as functions of $\gamma$ via constraints on the population L-moments given by

\[
\lambda_1 = E[X_{(1,1)}] = \mu - \frac{\sigma}{\gamma} (1 - \Gamma(1 - \gamma)) \\
\lambda_2 = \frac{1}{2} E [X_{(2,2)} - X_{(1,2)}] = -2 \frac{\sigma}{\gamma} (1 - 2^\gamma) \Gamma(1 - \gamma).
\]  

(16)

**Stage 2:** Estimate the population L-moments empirically via the sample L-moments which utilize the ordered data realizations $\{x_{n,K}\}_{n \in \{1,2,\ldots,K\}}$ given by

\[
\hat{\lambda}_1 = \frac{1}{K} \sum_{n=1}^{K} x_{(n,K)}, \text{ and } \hat{\lambda}_2 = \frac{1}{K(K-1)} \sum_{n=1}^{K} \left\{ \left(\frac{n-1}{1}\right) - \left(\frac{K-n}{1}\right) \right\} x_{(n,K)}.
\]  

(17)

Then utilize these estimates to obtain the estimators for $\mu$ and $\sigma$ with respect to $\gamma$ according to the expressions incorporating these L-moment estimates given by

\[
\hat{\sigma} = -\frac{1}{2} \gamma \hat{\lambda}_2 [(1 - 2^\gamma) \Gamma(1 - \gamma)]^{-1} \\
\hat{\mu} = \hat{\lambda}_1 + \frac{1}{\gamma} (1 - \Gamma(1 - \gamma)) \left\{ -\frac{1}{2} \gamma \hat{\lambda}_2 [(1 - 2^\gamma) \Gamma(1 - \gamma)]^{-1} \right\}.
\]  

(18)

**Stage 3:** Perform Maximum Likelihood Estimation for the EVI parameter $\gamma$ subject to the constraints on L-Moments imposed by the estimates in Stage 2, which involves maximization of the re parameterized likelihood given, for $\gamma \neq 0$ by

\[
\ln l(x_{(1,K)}; \hat{\mu}, \hat{\sigma}, \gamma) \approx -K \ln \left( -\frac{1}{2} \gamma \hat{\lambda}_2 [(1 - 2^\gamma) \Gamma(1 - \gamma)]^{-1} \right) \\
- (1 + 1/\gamma) \sum_{n=1}^{K} \ln [1 + \gamma S(x_{(n,K)}, \gamma)] - \sum_{n=1}^{K} [1 + \gamma S(x_{(n,K)}, \gamma)]^{-1/\gamma}.
\]  

(19)
where the function $S(x_{(n,K)}, \gamma)$ is defined according to

$$
S(x_{(n,K)}, \gamma) = \left( x_{(n,K)} - \left( \hat{\lambda}_1 + \frac{1}{\gamma} (1 - \Gamma(1 - \gamma)) \left\{ -\frac{1}{2} \gamma \hat{\lambda}_2 \left[ (1 - 2^\gamma) \Gamma(1 - \gamma) \right]^{-1} \right\} \right)
$$

$$
- \frac{1}{2} \gamma \hat{\lambda}_2 \left[ (1 - 2^\gamma) \Gamma(1 - \gamma) \right]^{-1}.
$$

(20)

If $\gamma$ is in the neighbourhood of the origin ($\gamma \in \text{n.e.}(0)$) the likelihood is given according the Gumbel limit of the GEV distribution, specified as

$$
\ln l(x_{(1:K,K)}; \mu, \sigma, \gamma) \approx -K \ln \left( -\frac{1}{2} \gamma \hat{\lambda}_2 \left[ (1 - 2^\gamma) \Gamma(1 - \gamma) \right]^{-1} \right)
$$

$$
- \sum_{n=1}^{K} S(x_{(n,K)}, \gamma) - \sum_{n=1}^{K} \exp \left[ -S(x_{(n,K)}, \gamma) \right].
$$

(21)

5.3. Generalized Pareto Distribution (GPD) Model Estimation

There are numerous parameter estimation procedures available for the GPD family, the more standard of these are detailed in the Technical Appendix E. The approaches we considered were the re-parameterized MLE estimation and those based on empirical percentiles such as the analytic Pickands estimators and robust versions of the Empirical Percentiles Method (EPM). Again intra-daily data for each asset during 2010 at time resolutions of 1, 2, 5 and 10 seconds was considered, and trading days were split into blocks and a Peaks over Threshold (POT) approach was adopted. Under the POT’s approach the last volume recorded in the specified time increment is retained only if it exceeds the threshold. Comparing this to the preparation of data in Sections 5.1 and 5.2, we can see that the $\alpha$-Stable estimation considers the full sub-sampled intraday data sets, GEV considers the maximum volume per sub-sample time increment and GPD considers only the largest percentage of volume defined by the specified threshold.

The standard MLE and reparameterized MLE approaches for the GPD model are detailed in the Technical Appendix E.2. The reparameterization makes the procedure numerically more robust since the main challenge with MLE based estimation here is the constraints required on the parameters $\gamma$ and $\sigma$ which for the $J$ data points obtained after the POT’s procedure are subject to:

1. $\sigma > 0$
2. $1 + \gamma x_{(J,J)}/\sigma > 0$.

Note, this second constraint is important since one observes that if $\gamma < -1$ as $-\sigma/\gamma \to y(J)$ then the likelihood approaches infinity. It is well known that one can in fact reparametrize the GPD likelihood to aid in the numerical stability of the parameter estimation via an MLE approach. The GPD estimation of parameters via Method of Moments, which is analytic for the parameterization considered can be found in Technical Appendix E.3.

5.3.1. Robust Parameter Estimation of GPD Models via Empirical Percentiles Methods

In this section we detail an additional parameter estimation procedures which may be developed with no restriction on the extreme value index shape parameter $\gamma \in \mathbb{R}$. This is based on the percentile based matching approach proposed in Castillo and Hadi (1997). The idea of the Empirical Percentile Method (EPM) approach to parameter estimation involves making the most of information contained in the order statistics obtained from the observations. One equates the model cdf evaluated at the observed order statistics to their corresponding percentile values. This system of equations can then be solved for the models distributional parameters. In the case of the GPD model for the volume profiles on the bid or ask
at level 1, there are two model parameters so one requires two distinct order statistics as a minimum to perform the estimation, since we fix the threshold at which we consider the POT’s method giving \( \mu \) and we need to estimate \( \gamma \) and \( \sigma \).

Consider a set of realized data obtained under a POT’s approach, where the \( J \) volumes that have exceeded a pre-specified threshold level \( u \) are denoted by the data \( \{ x_i \}_{i=1}^J \) with order statistics denoted by \( \{ x_{(i,j)} \}_{i=1}^J \). Given the CDF of the GPD model in Equation 22

\[
F(x; \gamma, \sigma) = \begin{cases} 
1 - (1 - \frac{x}{\sigma})^{\frac{1}{\gamma}} , & \gamma \neq 0, \sigma > 0, \\
1 - \exp\left(-\frac{x}{\sigma}\right) , & \gamma = 0, \sigma > 0,
\end{cases}
\]

we match the CDF at two of the selected order statistics \( i \neq j \in \{1,2,\ldots,J\} \) to the corresponding percentile values,

\[
F(x_{(i,j)}; \gamma, \sigma) = p_{(i,j)} \quad \text{and} \quad F(x_{(j,i)}; \gamma, \sigma) = p_{(j,i)},
\]

where the percentile is given for the GPD model with \( J \) observations by

\[
p_{(i,j)} = \frac{i - \eta}{J + \zeta}. \tag{24}
\]

It is recommended in Castillo and Hadi (1997) that choices of \( \eta = 0 \) and \( \zeta = 1 \) provide reasonable results, so these settings were utilized in the studies performed. The solution to this system of equations in terms of the parameters is obtained by solving the equations for \( \gamma \) and \( \sigma \) given by

\[
1 - \left(1 - \frac{\gamma x_{(i,j)}}{\sigma}\right) = \frac{i}{J+1} \quad \text{and} \quad 1 - \left(1 - \frac{\gamma x_{(j,i)}}{\sigma}\right) = \frac{j}{J+1}. \tag{25}
\]

Hence for any two pairs of order statistics \( i, j \) the solutions to these system of equations is

\[
\hat{\gamma}(i, j) = \frac{\ln \left(1 - \frac{x_{(i,j)}}{\delta_{(i,j)}}\right)}{C_i} \quad \text{and} \quad \hat{\sigma}(i, j) = \hat{\gamma}(i, j) \hat{\delta}(i, j), \tag{26}
\]

in terms of \( C_i = \ln \left(1 - p_{(i)}(J)\right) < 0 \) and \( \hat{\delta}(i, j) \). Here \( \hat{\delta}(i, j) \) is the solution to the equation,

\[
C_i \ln \left(1 - \frac{x_{(j,i)}}{\delta}\right) = C_j \ln \left(1 - \frac{x_{(i,j)}}{\delta}\right) \tag{27}
\]

which is obtained using a univariate root finding algorithm such as bisection. Note, that \( \delta \) corresponds to a reparameterization of the GPD distribution when \( \delta = \frac{\sigma}{\gamma} \).

**Remark 3 (EPM and Pickands Analytic Solution).** A special case of the EPM estimators is widely used in estimation of the GPD model parameters and known as the Pickands estimators. These correspond to the EPM setting in which \( i = \frac{J}{2} \) and \( j = \frac{3J}{4} \). In these special cases, the bisection method is not required as the system of equations can be solved analytically according to

\[
\hat{\gamma} = \frac{1}{\ln 2} \ln \left(\frac{x_{(j/2,j)}}{x_{(3j/4,j)} - x_{(j/2,j)}}\right) \quad \text{and} \quad \hat{\sigma} = \hat{\gamma} \left(\frac{x_{(j/2,j)}^2}{2x_{(j/2,j)} - x_{(3j/4,j)}}\right). \tag{28}
\]

In general we would not just pick two indexes \( i, j \) and instead we would combine the Algorithm’s one and two discussed in Castillo and Hadi (1997) to produce an estimate of the GPD parameters. This involves the following sequence of steps.

**Empirical Percentile Estimation for GPD Parameter Estimation:**
Combining Algorithm 1 and Algorithm 2 of Castillo and Hadi (1997).
1. Repeat for all order indexes \( \{i, j : i < j, \text{ for } i, j \in \{1, 2, \ldots, J\} \} \), such that \( x_{i,j} < x_{j,j} \), the following steps:

   (a) Compute for \( s \in \{i, j\} \) the values \( C_s = \ln \left( 1 - \frac{s - \eta}{J + s} \right) \).

   (b) Set \( d = C_j x_{i,j} - C_i x_{j,j}, \) if \( d = 0 \) let \( \delta(i, j) = \pm \infty \) and set the EVI estimate \( \hat{\gamma}(i, j) = 0 \), otherwise compute \( \delta_0 = x_{i,j} x_{j,j}(C_j - C_i)/d \).

   (c) If \( \delta_0 > 0 \), then \( \delta_0 > x_{i,j} \) and the bisection method can be used for the interval \( [x_{j,j}, \delta_0] \) to obtain a solution \( \hat{\delta}(i, j) \). Otherwise the bisection method is applied to the interval \( [\delta_0, 0] \).

   (d) Use \( \hat{\delta}(i, j) \) to compute \( \hat{\gamma}(i, j) \) and \( \hat{\sigma}(i, j) \) using

\[
\hat{\gamma}(i, j) = \hat{\gamma}(1, J) = \frac{\ln \left( 1 - \frac{x_{i,J}}{\delta(i,j)} \right)}{C_i} \quad \text{and} \quad \hat{\gamma}(i, j) = \hat{\gamma}(1, J) \hat{\delta}(i, j).
\]

2. Take the median of each of the sets of estimated parameters for the overall estimator to obtain

\[
\hat{\gamma}_{\text{EPM}} = \text{median} \{\hat{\gamma}(1, 2), \hat{\gamma}(1, 3), \ldots, \hat{\gamma}(J - 1, J)\} \quad \text{and} \quad \hat{\sigma}_{\text{EPM}} = \text{median} \{\hat{\sigma}(1, 2), \hat{\sigma}(1, 3), \ldots, \hat{\sigma}(J - 1, J)\}.
\]

6. Discussions on Model Estimation for LOB Volume Profiles

Having fitted each of these classes of parametric models to the volume profile data we assess the findings of the fits for the heavy tail analysis. Before detailing the results it is interesting to observe the following attributes of the global futures markets in 2010, as provided in the white paper report of Will Acworth (2010) and summarized below in Table 4 which is based on the number of contracts traded and cleared at 76 exchanges worldwide. We can also decompose this further to particular sectors.

| Global Futures and Options Volume | Jan-Jun 2009 | Jan-Jun 2010 | % Change |
|----------------------------------|--------------|--------------|----------|
| Futures                          | 3,868,238,401| 5,685,753,558| 47.0%    |
| Options                          | 4,649,547,118| 5,535,731,102| 19.1%    |
| Total Volume                     | 8,517,785,519| 11,221,484,660| 31.7%    |

| Volume By Category/Sector         | Jan-Jun 2009 | Jan-Jun 2010 | % Change | % of Total |
|----------------------------------|--------------|--------------|----------|------------|
| Equity Index                     | 3,137,652,330| 3,639,154,960| 16.0%    | 32.4%      |
| Individual Equity                | 2,801,412,186| 3,304,469,449| 18.0%    | 29.4%      |
| Interest Rate                    | 1,208,656,847| 1,659,815,977| 37.3%    | 14.8%      |
| Currency                         | 351,909,733  | 1,238,189,101| 251.8%   | 11.0%      |
| Agricultural                     | 410,749,029  | 582,754,069  | 41.9%    | 5.2%       |
| Energy                           | 323,529,689  | 358,938,047  | 10.9%    | 3.2%       |
| Non-Precious Metals              | 156,184,718  | 305,385,440  | 95.5%    | 2.7%       |
| Precious Metals                  | 76,869,990   | 85,066,263   | 10.7%    | 0.8%       |
| Other                            | 50,820,997   | 47,711,354   | -6.1%    | 0.4%       |
| Total                            | 8,517,785,519| 11,221,484,660| 31.7%    | 100.0%     |

Table 4: FIA Volume Webinar: Will Acworth, Editor, Futures Industry.

source by: http://www.futuresindustry.org/downloads/Volume_Webinar_2010_Final_2.pdf
These demonstrate the state of the market at a global economic scale, providing insight into the total volume changes in the futures market relative to other economic activity in the period of study for the LOB volume profiles. It will be of interest to understand the modeling in this context especially with respect to extreme volume profile attributes observed.

We note that in general we found little difference in the estimated model parameters at each of the sampling frequencies and on these grounds we proceed with discussion of the 10 seconds sampling rate. This provides some conclusions relating to the presence of heavy-tailed features as fundamental attributes of the volume profiles at each level of the LOB on the bid and ask. The importance of the interpretation of this result stems from the fact that the heavy tailed features are consistent at both the high sampling frequencies considered (1 second) and the lower sampling frequencies of 1 minute - 10 minutes. The consistent presence of these features at all these sampling rates allows one to conclude that in this range of high frequency trading on these exchanges, it is certainly not the trading activities, such as quote stuffing and price discovery mechanisms which are driving these heavy tailed features. The reason for this conclusion is that such activities tend to take place at the high frequency range of the sampling rates considered (< 1 second). This can have implications for future analysis, especially relating to the current perception that such high frequency trading activities may be artificially destabilizing and enhancing the volatility of electronic exchanges.

6.1. Alpha-Stable Model Results

The $\alpha$-Stable family of distributions were fitted to volume data on bid and ask level 1 that was scaled by the interquartile range (IQR). For each day of data we utilized the McCullochs method to estimate the parameters and we consider analysis of the parameter behavior intraday and interday. We begin the analysis with the dynamics of the parameters that define the $\alpha$-Stable distribution across the 6 assets analyzed, 5YTN, BOBL, SP500, NIKKEI, GOLD and SILVER, with all results being provided in the Technical Appendix H - Technical Appendix M.

6.1.1. Five Year T-Note Futures Contract

In terms of the total volume of contracts traded worldwide the 5YTN was 9th most heavily traded futures in the interest rate sector globally with an increase in total volume between 2009 and 2010 of 47.1%, see Will Acworth (2010). The tail index parameter estimates indicate for the majority of 2010, light tails for volumes on bid and ask, with a mean of 1.9895. However, a particularly interesting feature involved a few pronounced periods in which a heavy tail model is clearly appropriate for the volume profile. The days on which these heavy tailed volume profiles occurred did not correspond to the same days for the bid and ask volume profiles, indicating an asymmetry in the volume profile on the bid and ask over time, with respect to extreme volumes. It is difficult to pinpoint exact events driving this behavior. However, broadly speaking, in March of 2010 when the first heavy tailed profiles for the LOB bid and ask were present, was a period in which investors who had moved into the security of the US treasury notes due to the debt crisis in Europe may have been looking to move back to European markets. The reason for this is that in March some measure of relief in the European markets arose due to agreements for a Greek bailout by Germany and the International Monetary Fund (IMF). In addition sentiment was buoyed by the fact that US seemed to be recovering with the signing of the historic $938 billion health care bill. The next two periods of extreme movement in the volume profile in the LOB of the 5YTN correspond to periods in late July and late August during which times the events that may have affected the bond markets in the US included the fact that in mid-July Greece raised US $2.04 billion in an auction of government bonds, to appease the bailout that had occurred in May, which would have potentially drawn investors from the US markets and altered volume. Then in August in the US home loan data on foreclosures
continued to increase, resulting in a flight to safety of the US treasury notes by investors, see discussions on macro-economic factors affecting the economy in this period in Blog (2010).

Though statistically it is not possible to attribute significance directly to the relationship between the volume profile extreme movements and these events, they do however form a macro-economic perspective to help understand the sentiment of investors and conceivable reasons for such extreme volume profile events. Additionally, volumes for the 5YTN also appear to be heavily right skewed with a $\beta = 1$ a large portion of the year, (see results in Technical Appendix H).

6.1.2. Euro-BOBL Futures Contract

The BOBL was the 8th most heavily traded futures contract in the interest rate sector globally, with an increase in total volume between 2009 and 2010 of 36.7%, see Blog (2010). Contrasting the findings for the 5YTN to BOBL, from Figure 10 we can see a significantly greater variation in the tail index $\alpha$, daily and also the skewness parameter $\beta$, with the mean of the tail index parameter being 1.8195. This indicates that the daily volume profile on the bid and ask is consistently more heavy tailed than the daily behavior of the volume profile for the 5YTN. In addition, the extreme volume profile events observed occasionally in the 5YTN are not present in the BOBL volume profile, until the end of 2010 where an event caused the volume profile to demonstrate an infinite mean tail behavior for a few trading days in late November to early December. During this period of significantly heightened activity in the heaviness of the volume profile tails, there was a corresponding significant event in Europe which may have impacted the European traded BOBL. During this time, investors massively sold off Irish bonds which consequently drove prices on the actively traded 10 year bonds to a new low. In addition to this in mid-December of 2010 the rating agency Moody’s gave a 5 notch credit downgrade to Irish debt ratings, from Aa2 to Ba1 rating further driving an exodus from Irish and European bond markets. This is consistent with the observed extreme events estimated from the $\alpha$-Stable daily model fits during this period.

6.1.3. SIMEX Nikkei 225 Futures Contract

The NIKKEI in 2010 was globally the 12th most actively traded Equity index with an increase in total volume from 2009 to 2010 of 23.8%, see Blog (2010). The daily volume profiles for the LOB for this asset were similar in nature to the BOBL’s consistent heavy right tail attributes with strong skew and a mean tail index of 1.8022. In addition, we observed for the NIKKEI in Figure 11 that the daily volume profile on the bid and ask is not only consistently heavy tailed but asymptotically dominates the behavior in the volume profile of the BOBL and 5YTN. In addition, there is a marked relationship between asymmetry in the bid and ask volume profiles with the bid tending to produce a symmetric distributional fit when the ask is asymmetric and vice versa.

6.1.4. E-mini S&P 500 Futures Contract

Comparing the SP500 with the other assets considered, we see a shape parameter that is more pronounced, with a mean of 1.7474. In addition, we see a consistent daily tail profile which had a tail index away from $\alpha < 2$. What is also of interest here is that when considering the SP500, which was globally the 2nd most actively traded equity future in 2010, there was actually a total volume decrease between 2009 and 2010 of $-3.0\%$. However, this total change in volume did not affect the general attributes observed for the model estimation with regards to the heavy tailed behavior and the manner in which these contracts are traded on an intra-daily basis throughout the year.
Figure 10: $\alpha$-Stable daily parameter estimation for the year 2010 using McCulloch's method for BOBL at a time resolution of 10 seconds. The blue dashed line is the bid $L_1$ and red dashed line ask $L_1$. **Top Left Plot:** Tail index parameter $\alpha$ daily estimates. **Top Right Plot:** Asymmetry parameter $\beta$ daily estimates. **Bottom Left Plot:** Scale parameter $\gamma$ daily estimates. **Bottom Right Plot:** Location parameter $\delta$ daily estimates.

Figure 11: $\alpha$-Stable daily parameter estimation for the year 2010 using McCulloch's method for GOLD at a time resolution of 10 seconds. The blue dashed line is the bid $L_1$ and red dashed line ask $L_1$. **Top Left Plot:** Tail index parameter $\alpha$ daily estimates. **Top Right Plot:** Asymmetry parameter $\beta$ daily estimates. **Bottom Left Plot:** Scale parameter $\gamma$ daily estimates. **Bottom Right Plot:** Location parameter $\delta$ daily estimates.

6.1.5. Gold and Silver Futures Contract

For the precious metals explored, the most prominent of the heavy-tailed volume profiles is the GOLD futures which had a mean shape parameter of 1.5076 and consistently heavy tailed behavior intra-daily throughout all trading days in 2010. GOLD also demonstrated a strongly right skewed distribution for the volume profile at level 1 of the bid and ask. During 2010, the worldwide volume of futures contracts over precious metals, of which GOLD and SILVER are members, increased from 76,869,990 traded contracts in 2009 to 85,066,263 in 2010. This constituted an increase of 10.66% worldwide. However, by region the story was vastly different, with the asset GOLD analyzed in this study and which trades on the CME,
saw an increase of 48.0% in total volume of contracts between 2009 and 2010. Gold futures that trade on Tocom (not considered in this paper) saw an increase of 20.9% from 2009 to 2010 and conversely, Gold M Futures and Gold Futures that trade on the MCX saw a decrease of \(-14.4\%\) and \(-10.2\%\), respectively. Silver futures grew consistently on all exchanges throughout the 2009 to 2010 period, most notably on the CME exchange where the data was assessed with a very large increase in total volume of contracts of 65.9\%.\(^2\) The results for the Silver futures demonstrated a few marked periods in which the intra-daily volume profile on both the bid and ask became exceptionally heavy tailed in nature, most noticeably in the mid-year in which the ask side had tail index values around \(\alpha \approx 1\).

6.2. GEV Model Results

Consistent with the findings from the \(\alpha\)-Stable model, the estimation results for the 5YTN show heavy-tailed behavior is present in the volume profiles between the 50th and 60th trading day and the 140th and 150th trading days. Interestingly, the prominence of the heavy tailed features for BOBL is more pronounced under the GEV model fits compared to the \(\alpha\)-Stable model. Additionally, the structural changes in the behavior of the intra-daily volume profile on the bid and ask are observed in the location and scale parameters for the BOBL around the 100th trading day, where there is a marked regime shift in the estimated model parameters. This is just as prominent in the GEV fit as it was in the \(\alpha\)-Stable model, indicating its plausibility in a marked dynamic change in the intraday activity in this market mid trading year. A similar change is visible in the SP500 shape and scale parameters, but in this case the regime reverts gradually back to the behavior present intraday at the start of the year. This structural change is not as prominent in the NIKKEI.

When considering the tail index parameter, \((\gamma)\) for the MLE method, we can see that the mean intra-daily estimated value averaged over all trading days in 2010 for the 5YTN, BOBL (Figure 12) and NIKKEI is close to zero, respectively \(-0.0493, 0.0114, 0.0725\). SP500 has a higher mean level for the shape parameter, being 0.1495. Again, as in the \(\alpha\)-Stable case, the GEV model estimations do indicate a reasonable variation in the tail index throughout 2010, indicating a number of days in which heavy tailed attributes are appropriate. On a few days analyzed and for assets 5YTN, SP500 and NIKKEI we see instances where the shape parameter spikes \((\gamma > 1)\) indicating infinite mean-variance models are suitable. The occurrence of such events coincides with the trading days in which the \(\alpha\)-Stable model also indicated heavy tailed behavior as suitable. For all assets we see a correlation in the parameter estimations for the bid and ask side and all assets show time variation across the year for scale and location. The structural downward shift is only present in BOBL (Figure 12) at around the same time point as that discussed for the \(\alpha\)-Stable distribution.

The Mixed L-Moments approach is used to confirm our findings of the estimation of the tail index parameter, which is notoriously difficult to estimate. The results for all assets are provided in the Technical Appendix H - Technical Appendix M. Comparing the subplots in Figure 12 for BOBL we have a good representation of the features we found to be consistent across the assets 5YTN, SP500 and NIKKEI. The scale and location parameters are very similar for each method implemented. However we see that the shape parameter is systematically different when comparing the MLE and Mixed L-Moments methods, with the mean level for the shape parameter using the MLE method being between \((-0.0495, 0.1495)\) for all assets, whereas the Mixed L-Moments method produces a mean level of \((0.1951, 0.2321)\) for the shape parameter for all assets.

\(^2\)http://www.futuresindustry.org/downloads/Sept_Volume(rev).pdf
To further explore the discrepancies between the MLE and Mixed L-Moments, specifically the upward translation of the shape parameter by approximately 2, we performed a simulation study which considered sample sizes of 50 and 10,000 randomly generated GEV distributed data series. We consider 20 different simulations to estimate the parameters of the GEV distribution using MLE and Mixed L-Moments methods. Results are provided in a case study in the Technical Appendix F.1 showing the same discrepancies between the different estimation methods for the shape parameter. As $\gamma \to 0$, we see an increased bias of $\gamma$ under the Mixed L-Moments method, but the tradeoff is that the variance is reduced in the Mixed L-Moments method. Mixed L-Moments becomes more reliable as the sample size decreases, thus prompting us to recommend the use of MLE for higher frequency LOB volume data.

In summary, we observed that the MLE method provided a more stable fit daily compared with the Mixed L-Moments method for these applications. Interestingly, we found from the analysis that as the data becomes significantly heavier tailed, as was the case for GOLD with a mean intraday tail index parameter of 0.3564 for 2010, the results for the MLE estimation and the L-moments based solutions were in much closer in alignment. The observed bias present in the cases of light tailed volume profiles on certain trading days in the BOBL was not present in the consistently heavy tailed GOLD.

6.3. GPD Model Results

The GPD family utilize a Peaks Over Threshold preparation of the data, with a translation by a threshold corresponding to the 80th percentile of the data see discussion in Embrechts et al. (1999) for the location ($\mu$) parameter. The features obtained under the GPD model are consistent with the findings discussed for the $\alpha$-Stable and GEV models. In particular the prevalence for the heavy tailed attributes remain, as does the interesting features of the increased extreme intraday volume activities resulting in heavy tailed attributes appearing for the BOBL as observed in the GEV and $\alpha$-Stable fits.

The scale parameter, $\sigma$ for the GPD distribution, using the MLE method for estimation, shows some structural shifts for 5YTN and SP500 around the same time period as the structural shifts observed in the $\alpha$-Stable distribution. A detailed summary of the results for the intraday estimation of the GPD model per trading day of 2010 are presented for the MLE, Pickands and Empirical Percentile Methods approaches for all assets in the Technical Appendix H - Technical Appendix M.
In terms of the different estimation approaches, the scale parameter using the Pickands estimator shows similar trending features as the MLE method. However, it also exhibits much higher variability in parameter estimation compared with the MLE method. The Pickands estimator fails to capture the days where we see a significant spikes in the shape and scale parameters.

The results observed for the EPM method demonstrated a substantial deviation from those results from the MLE and the Pickand’s estimator. The implementation of the EPM approach that we adopted involved matching each pair of percentiles above a threshold of the median to obtain a solution for the model parameters, producing a very large number of solutions. Next we took the median of these solutions as a robust estimation of the model parameters under the EPM approach, as discussed in Castillo and Hadi (1997). Our findings indicated the sensitivity of this estimation approach to the inclusion of low percentile solutions into the estimation (median calculation) for the shape (tail index) parameter.

In the Technical Appendix F.2 a case study is provided to further explore why the EPM method gives substantially different results to the MLE and Pickands methods. We consider a simulation study with 20 data sets each of 500 randomly generated GP data points used to estimate the MLE, Pickands and EPM methods. We consider the impact of setting the shape parameter positive and negative.

The results show that for a positive shape parameter, the three methods appear to be consistent. However, when the actual shape parameter is negative we see a significant translation upwards, with increased variability in the estimator, consistent with results for the observed three methods when analyzing the real LOB volume data.

We note the EPM method of estimation had significant issues when attempting to apply this method to real data. As discussed in the simulation case studies, we found that the method produced a much more stable result when using higher starting percentiles for the grid search method for maximizing the log-likelihood function. However, for the simulated and real case, we could not resolve the issue of when the data appears to have an actual negative shape parameter, thus resulting in a shape parameter multiple magnitudes higher than what we achieve through MLE and Pickands estimation techniques. It should also be noted that the starting percentile used for this method was the 50th percentile.

7. Conclusion

This paper considered 3 distributions across 6 futures assets for an entire year of trading (2010). From this analysis we attempt to model and assess the tail features present in the data. In addition to the different distributional fits considered, we explored many different parameter estimation methods within
each distribution to ensure that we have a method that is robust to the ultra-high frequency data analyzed and a method that best captures the features present in the data.

We conclude by providing some initial answers to the questions posed from the onset of this analysis. With regard to the question of seeking statistical evidence for the presence of sub-exponential volume profile right tails at different levels of the order book, we confirmed that all models considered, the $\alpha$-Stable, GEV and GPD models all showed consistent results for the presence of intraday periods of heavy tails in which the tail index parameter spiked in an asymmetric manner on level 1 of the bid and ask volume profiles. With regard to the notion that perhaps the heavy tailed features may be exchange specific due to possible effects related to specific exchange mechanisms or market participants in such exchanges, we found that the potential for heavy tailed features in the intraday volume profiles was present in all exchanges considered in the analysis. In addition, the asymmetry present in the times of occurrence of the heavy tailed features of the bid and ask volume profiles was also not exchange specific. Interestingly, it was certainly the case that particular asset classes had a greater tendency to systematically display heavy tailed volume profiles at all levels of the volume profiles on the bid and ask compared to others. For instance we demonstrated that clearly the precious metals (GOLD and SILVER) contracts displayed heavy tailed intraday behaviour systematically throughout the trading year, whereas the 5YTN and BOBL displayed such features intermittently throughout the year.

Due to the range of high frequency sampling rates considered at 10 seconds, 5 seconds, 2 seconds and 1 second intraday for each trading day of 2010, we were able to disambiguate the real stochastic behavior of the heavy tailed features of the LOB volume profile structures from the high-frequency “micro-structure noise”, see recent work in al Dayri (2011). It was hypothesized that trading activities described as “malicious” or “disruptive”, for example quote stuffing and price discovery mechanisms, lead to potentially destabilizing high frequency trading activities and where thus, more prevalent in the 1 second sampling resolutions. We showed that for a given market and asset class, the presence of statistical features such as the heavy tailed attributes were persistent in the stochastic processes for the range of sampling resolutions. Thus, the “destabilizing” activities were not responsible for the heavy tailed attributes estimated in intraday series.

At qualitative level we found plausible arguments for why the presence of the heavy tailed features on particular market days in 2010 was present on certain exchanges and for certain assets. We were able to relate these occurrences to the macro-economic environment and news impacts in those portions of the 2010 trading year, and future work will be required to undertake this study in a formal statistical analysis investigating these as causal drivers for such features. During non-heavy tailed volume profile intraday time periods, we found the bid and ask LOB marginal dynamics to generally follow a strong common trend. However interestingly, during heavy tailed events, there was an asymmetry in the volume profiles on the bid and ask at all levels. It was also clear that there is strong statistical evidence from the model estimations undertaken to recommend that when developing parametric models for the volumes on the bid and ask, these should be parameter driven and dynamics for the parameter process should be considered both on an intraday and interday dynamics, incorporating short term and long term mean dynamics, time varying volatility and heavy tailed features. This was clear since all assets demonstrate time variation for the parameters across the year with intraday and interday variations present. For BOBL, SP500 and NIKKEI, the large number of times the shape parameter dropped below 1.8 was indicative of why we found the Normal and Gamma distributions to be a bad fit even after simple transformations were applied. Skewness for the SP500 is aligned with the findings for 5YTN, with a $\beta = 1$ a large portion of the year. Both BOBL (figure 10) and NIKKEI demonstrated left skewness, but less pronounced compared with 5YTN and SP500.
8. Acknowledgements

We thank Boronia Capital Pty. Ltd. for providing the data and in particular the expertise of Dr. Chris Mellen, Russell Grew and Dr David Oliver. Kylie-Anne Richards is also thankful for the support provided by the Boronia-QRSLab PhD scholarship that has support part of this research time.

References

Alder, R., Feldman, R., Taqqu, M.S., 1998. A practical guide to heavy-tails: Statistical techniques for analysing heavy-tailed distributions. Birkhäuser.

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., Waal, D., Ferro, C., 2004. Statistics of extremes: Theory and applications. Wiley, New York.

Bensalah, Y., 2000. Steps in applying extreme value theory to finance: A review. Bank of Canada.

Biais, B., Hillion, P., Spatt, C., 1995. An empirical analysis of the limit order book and the order flow in the paris bourse. The Journal of Finance 50, 1655 – 1689.

Blog, I., 2010.

Bouchaud, J.P., Mezard, M., Potters, M., 2002. Statistical properties of stock order books: Empirical results and models. Quantitative Finance 2, 251–256.

Bowley, G., 2010. (1-oct-2010) lone 4.1 billion sale led to flash crash in may. the new york times. http://www.nytimes.com/2010/10/02/business/02flash.html?ref=finance

Castillo, E., Hadi, A.S., 1997. Fitting the generalized pareto distribution to data. Journal of the American Statistical Association 92, 1609–1620.

Chakrabort, A., Toki, L.M., Patriarca, M., Aberg, F., 2010. Econophysics: Empirical facts and agent-based models. WP.

Challet, D., Stinchcombe, R., 2001. Analyzing and modelling 1+1d markets. Physica A: Statistical Mechanics and its Applications 300, 285–299.

Chickeportiche, R., Bouchaud, J.P., 2012. Weighted kolmogorov-smirnov test: accounting for the tails. WP.

Cont, R., 2009. Volatility clustering in financial markets: Empirical facts and agent-based models. Available at SSRN 1411462.

Dalén, J., 1987. Algebraic bounds on standardized sample moments. Statistics & probability letters 5, 329–331.

David, H., Nagaraja, H., 1970. Order statistics. Wiley Online Library.

al Dayri, K.A., 2011. Market microstructure and modeling of the trading flow. PhD Thesis - Docteur de l’Ecole Polytechnique.
Embrechts, P., Klüppelberg, C., Mikosch, T., 1997. Modelling extremal events for insurance and finance.

Embrechts, P., Resnick, S., Samorodnitsky, G., 1999. Extreme value theory as a risk management tool. North American Actuarial Journal 3, 30–41.

Fama, E., 1965. The behaviour of stock market prices. J. Business 38, 34–105.

Fama, E., Roll, R., 1968. Some properties of symmetric stable distributions. Journal of the American Statistical Association 63, 817–83.

Fama, E., Roll, R., 1971. Parameter estimates for symmetric stable distributions. Journal of the American Statistical Association , 331–338.

Godsill, S., 2000. Inference in symmetric alpha-stable noise using MCMC and the slice sampler, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3806–3809.

Gould, M.D., Porter, M.A., Williams, S., McDonald, M., Daniel J. Fenn, S.D.H., 2012. Limit order books. arXiv:1012.0349 [q-fin.TR] .

Gu, G.F., Chen, W., Zhou, W.X., 2008. Empirical shape function of limit-order books in the chinese stock market. Physica A 387, 5182–5188.

Gu, W., Peters, G., Clavier, L., Septier, F., Nevat, I., et al., 2012. Receiver study for cooperative communications in convolved additive α-stable interference plus gaussian thermal noise. Ninth International Symposium on Wireless Communication Systems , 1–5.

Harris, L., 2003. Trading & Exchanges. Oxford University Press.

Hosking, J., 1985. Algorithm as 215: Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution. Journal of the Royal Statistical Society. Series C (Applied Statistics) 34, 301–310.

Hosking, J., 1990a. L-moments: analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society. Series B (Methodological) , 105–124.

Hosking, J., 1990b. L-moments: analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society. Series B (Methodological) , 105–124.

Hosking, J., Wallis, J., 1987. Parameter and quantile estimation for the generalized pareto distribution. Technometrics , 339–349.

Koning, A., Peng, L., 2008. Goodness-of-fit tests for a heavy tailed distribution. Journal of statistical planning and inference 138, 3960–3981.

Kratz, M., Resnick, S., 1996. The qq-estimator and heavy tails. Stochastic Models 12, 699–724.

Lauricella, T., . (2 oct 2010). how a trading algorithm went awry. the wall street journal. http://online.wsj.com/article/sb10001424052748704029304575526390131916792.html .

Lobato, I., Velasco, C., 2000. Long memory in stock-market trading volume. Journal of Business & Economic Statistics 18, 410–427.
Mandelbrot, B., 1960. The Pareto-Levey law and the distribution of income. International Economic Review 1, 79–106.

Martins, E., Stedinger, J., 2000. Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. Water Resources Research 36, 737–744.

Maslov, S., Mills, M., 2001. Price fluctuations from the order book perspective - empirical facts and a simple model. Physica A: Statistical Mechanics and its Applications 299, 234–246.

McCulloch, J.H., 1986. Simple consistent estimators of stable distribution parameters. Communications in Statistics - Simulation and Computation 15, 1109–1136.

McCulloch, J.H., 1998. Numerical approximation of the symmetric stable distribution and density., in: Adler, E., R. Feldman, R., Birkhauser, T.M. (Eds.), A practical guide to heavy tails: statistical techniques and applications.

McNeil, A., Frey, R., Embrechts, P., 2005. Quantitative risk management: concepts, techniques, and tools. Princeton university press.

Mehta, N., . (1 oct 2010). automatic futures trade drove may stock crash, report says. bloomberg. http://www.bloomberg.com/news/2010−10−01/automatic−trade−of−futures−drove−may−6−stock−crash−report−says.html .

Melchiori, M.R., 2006. Tools for sampling multivariate archimedian copulas. www.YieldCurve.com .

Morrison, J., Smith, J., 2002. Stochastic modeling of flood peaks using the generalized extreme value distribution. Water resources research 38, 1305.

Nikias, C., Shao, M., 1995. Signal processing with alpha stable distributions and applications. Wiley, New York.

Nolan, J.P., 2007. Stable distributions: Models for heavy-tailed data. Birkhäuser.

Peters, G., Nevat, I., Septier, F., Clavier, L., 2012. Generalized interference models in doubly stochastic poisson random fields for wideband communications: the pns (alpha) model. Arxiv preprint arXiv:1207.1531 .

Peters, G., Shevchenko, P., Young, M., Yip, W., 2011. Analytic loss distributional approach models for operational risk from the α-stable doubly stochastic compound processes and implications for capital allocation. Insurance: Mathematics and Economics .

Peters, G.W., Sisson, S.A., Fan, Y., 2009. Simulation for Bayesian models constructed with alpha stable distributions utilising approximate Bayesian computation .

Peters, G.W., Wüthrich, M.V., Shevchenko, P.V., 2010. Chain ladder method: Bayesian bootstrap versus classical bootstrap. Insurance: Mathematics and Economics 47, 36–51.

Pitman, E., 1980. Subexponential distribution functions. J. Austral. Math. Soc. Ser. A 29, 337–347.

Potters, M., Bouchaud, J.P., 2003. More statistical properties of order books and price impact. Physica A: Statistical Mechanics and its Applications 324, 122 – 140.
Potts, D., 2012. Share wars: how the robots are robbing you. The Age newspaper Australia August 26, 2012.

Prescott, P., Walden, A., 1980. Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. Biometrika 67, 723–724.

Royston, P., 2006. Which measures of skewness and kurtosis are best? Statistics in Medicine 11, 333–343.

Samorodnitsky, G., Taqqu, M.S., 1994. Stable non-Gaussian random processes: Stochastic models with infinite variance. Chapman and Hall/CRC.

Securities, T.G.R., 2012. Declining u.s. high-frequency trading. The New York Times October 15.

Smith, R., 1985. Maximum likelihood estimation in a class of nonregular cases. Biometrika 72, 67–90.

Topics, T., 2012. High-frequency trading. The New York Times October 16.

Vogel, R., Fennessey, N., 1993. L moment diagrams should replace product moment diagrams. Water Resources Research 29, 1745–1752.

Will Acworth, Editor, F.I., 2010. Fia volume webinar. FIA Volume Webinar, Sponsored by Tokyo Commodity Exchange.

Zolotarev, V., 1983. Univariate stable distributions .

Zolotarev, V.M., 1986. One-Dimensional Stable Distributions. Translations of Mathematical Monographs, American Mathematical Society.
Online Technical Appendix:

Heavy-Tailed Features and Empirical Analysis of the Limit Order Book Volume Profiles in Futures Markets

Kylie-Anne Richards, Gareth W. Peters and William Dunsmuir.

In the following set of Technical Appendices we present the following features:

- **Appendix A**: Quantile-Quantile plots for level 1 to level 5 on the bid and ask LOB volume profiles for assets: BOBL, SP500, NIKKEI
- **Appendix B**: Mean exceedence plots for level 1 to level 5 on the bid and ask LOB volume profiles for assets: BOBL, SP500, NIKKEI, SILVER
- **Appendix C**: Hill plots for level 1 to level 5 on the bid and ask LOB volume profiles for assets: 5YTN, BOBL, SP500, NIKKEI
- **Appendix D**: Detailed representations for the density and distribution functions for the univariate \(\alpha\)-Stable distributions
- **Appendix E**: Parameter estimation techniques for the GPD model
  - **Appendix E.1**: Generalized Pareto Distribution (GPD) likelihood specification
  - **Appendix E.2**: Generalized Pareto Distribution (GPD) reparameterization of the likelihood, Maximum Likelihood Estimation (MLE), Fisher information matrix and asymptotic properties of the MLE for the reparameterized GPD likelihood.
  - **Appendix E.3**: Generalized Pareto Distribution (GPD) Method of Moments estimation
- **Appendix F**: Synthetic case studies for different estimation techniques under the GEV and GPD models
  - **Appendix F.1**: Synthetic case studies for GEV models
  - **Appendix F.2**: Synthetic case studies for GPD models
- **Appendix G**: Kolmogorov-Smirnov test case study to assess goodness of fit
- **Appendix H**: Model estimation and empirical results for volume profile of 5 Year T-Note
  - **Appendix H.1**: McCulloch’s quantile based estimation - Alpha-Stable model results
  - **Appendix H.2**: Generalized Extreme Value (GEV) distribution
    - **Appendix H.2.1**: Maximum Likelihood Estimation - reparameterized GEV model results
    - **Appendix H.2.2**: Mixed L-Moments Estimation - GEV model results
  - **Appendix H.3**: Generalized Pareto Distribution (GPD)
    - **Appendix H.3.1**: Maximum Likelihood Estimation - reparameterized GPD model results
    - **Appendix H.3.2**: Pickands Estimator - GPD model results
    - **Appendix H.3.3**: Empirical Percentile Method - GPD model results
• Appendix I: Model estimation and empirical results for volume profile of Euro-BOBL
  - Appendix I.1: McCulloch’s quantile based estimation - Alpha-Stable model results
  - Appendix I.2: Generalized Extreme Value (GEV) distribution
    * Appendix I.2.1: Maximum Likelihood Estimation - reparameterized GEV model results
    * Appendix I.2.2: Mixed L-Moments Estimation - GEV model results
  - Appendix I.3: Generalized Pareto Distribution (GPD)
    * Appendix I.3.1: Maximum Likelihood Estimation - reparameterized GPD model results
    * Appendix I.3.2: Pickands Estimator - GPD model results
    * Appendix I.3.3: Empirical Percentile Method - GPD model results

• Appendix J: Model estimation and empirical results for volume profile of E-mini S&P 500
  - Appendix J.1: McCulloch’s quantile based estimation - Alpha-Stable model results
  - Appendix J.2: Generalized Extreme Value (GEV) distribution
    * Appendix J.2.1: Maximum Likelihood Estimation - reparameterized GEV model results
    * Appendix J.2.2: Mixed L-Moments Estimation - GEV model results
  - Appendix J.3: Generalized Pareto Distribution (GPD)
    * Appendix J.3.1: Maximum Likelihood Estimation - reparameterized GPD model results
    * Appendix J.3.2: Pickands Estimator - GPD model results
    * Appendix J.3.3: Empirical Percentile Method - GPD model results

• Appendix K: Model estimation and empirical results for volume profile of SIMEX Nikkei 225
  - Appendix K.1: McCulloch’s quantile based estimation - Alpha-Stable model results
  - Appendix K.2: Generalized Extreme Value (GEV) distribution
    * Appendix K.2.1: Maximum Likelihood Estimation - reparameterized GEV model results
    * Appendix K.2.2: Mixed L-Moments Estimation - GEV model results
  - Appendix K.3: Generalized Pareto Distribution (GPD)
    * Appendix K.3.1: Maximum Likelihood Estimation - reparameterized GPD model results
    * Appendix K.3.2: Pickands Estimator - GPD model results
    * Appendix K.3.3: Empirical Percentile Method - GPD model results

• Appendix L: Model estimation and empirical results for volume profile of Gold
  - Appendix L.1: McCulloch’s quantile based estimation - Alpha-Stable model results
  - Appendix L.2: Generalized Extreme Value (GEV) distribution
    * Appendix L.2.1: Maximum Likelihood Estimation - reparameterized GEV model results
    * Appendix L.2.2: Mixed L-Moments Estimation - GEV model results
  - Appendix L.3: Generalized Pareto Distribution (GPD)
    * Appendix L.3.1: Maximum Likelihood Estimation - reparameterized GPD model results
    * Appendix L.3.2: Pickands Estimator - GPD model results
Appendix L.3.3: Empirical Percentile Method - GPD model results

Appendix M: Model estimation and empirical results for volume profile of Silver

- Appendix M.1: McCulloch’s quantile based estimation - Alpha-Stable model results
- Appendix M.2: Generalized Extreme Value (GEV) distribution
  * Appendix M.2.1: Maximum Likelihood Estimation - reparameterized GEV model results
  * Appendix M.2.2: Mixed L-Moments Estimation - GEV model results
- Appendix M.3: Generalized Pareto Distribution (GPD)
  * Appendix M.3.1: Maximum Likelihood Estimation - reparameterized GPD model results
  * Appendix M.3.2: Pickands Estimator - GPD model results
  * Appendix M.3.3: Empirical Percentile Method - GPD model results
Appendix A. GPD(γ = 0) Quantile-Quantile Plots for Level 1 to Level 5 of the Bid and Ask LOB Volume Profiles

Figure A.14: **BOBL**: QQ plot for exponential distribution model for intra-daily volume data every 25th trading day of 2010. **Top Row Bid** from left to right is Level 1 to Level 5 of LOB; **Bottom Row Ask** from left to right is Level 1 to Level 5 of LOB

Figure A.15: **SP500**: QQ plot for exponential distribution model for intra-daily volume data every 25th trading day of 2010. **Top Row Bid** from left to right is Level 1 to Level 5 of LOB; **Bottom Row Ask** from left to right is Level 1 to Level 5 of LOB
Figure A.16: **NIKKEI:** QQ plot for exponential distribution model for intra-daily volume data every 25th trading day of 2010. **Top Row Bid** from left to right is Level 1 to Level 5 of LOB; **Bottom Row Ask** from left to right is Level 1 to Level 5 of LOB.

**Appendix B. Mean Exceedence Plots for Level 1 to Level 5 of the Bid and Ask LOB Volume Profiles**

Figure B.17: **BOBL:** Mean Excess plot for intra-daily volume data every 25th trading day of 2010. **Top Row Bid** from left to right is Level 1 to Level 5 of LOB; **Bottom Row Ask** from left to right is Level 1 to Level 5 of LOB.
Figure B.18: **SP500**: Mean Excess plot for intra-daily volume data every 25th trading day of 2010. **Top Row Bid** from left to right is Level 1 to Level 5 of LOB; **Bottom Row Ask** from left to right is Level 1 to Level 5 of LOB

Figure B.19: **NIKKEI**: Mean Excess plot for intra-daily volume data every 25th trading day of 2010. **Top Row Bid** from left to right is Level 1 to Level 5 of LOB; **Bottom Row Ask** from left to right is Level 1 to Level 5 of LOB
Figure B.20: **SILVER**: Mean Excess plot for intra-daily volume data every 25th trading day of 2010. **Top Row Bid** from left to right is Level 1 to Level 5 of LOB; **Bottom Row Ask** from left to right is Level 1 to Level 5 of LOB

**Appendix C. Hill Plots for Level 1 to Level 5 of the Bid and Ask LOB Volume Profiles**

**Hill Plots**

Figure C.21: **5 Year TNote**: Hill plot for intra-daily volume data every 25th trading day of 2010. **Top Row Bid** from left to right is Level 1 to Level 5 of LOB; **Bottom Row Ask** from left to right is Level 1 to Level 5 of LOB
Figure C.22: BOBL: Hill plot for intra-daily volume data every 25th trading day of 2010. **Top Row Bid** from left to right is Level 1 to Level 5 of LOB; **Bottom Row Ask** from left to right is Level 1 to Level 5 of LOB

Figure C.23: Mini S&P: Hill plot for intra-daily volume data every 25th trading day of 2010. **Top Row Bid** from left to right is Level 1 to Level 5 of LOB; **Bottom Row Ask** from left to right is Level 1 to Level 5 of LOB
Figure C.24: Nikkei: Hill plot for intra-daily volume data every 25th trading day of 2010. Top Row Bid from left to right is Level 1 to Level 5 of LOB; Bottom Row Ask from left to right is Level 1 to Level 5 of LOB
Appendix D. Representations of the $\alpha$-Stable Distribution and Density Models

In this section we provide some more technical aspects of the representations for the $\alpha$-Stable distribution and density that can be of relevance when utilizing the models fitted and also in developing robust estimation procedures such as the quantile based estimation procedure we utilized from McCulloch (1998)

**Definition 7 ($\alpha$-Stable Severity Density and Distribution Representations).** w.l.o.g. the density function of an $\alpha$-Stable severity distribution (standardized such that $\gamma = 1$ and $\delta = 0$) can be evaluated pointwise according to the series expansions Zolotarev (1983) [Equation 2.4.6, p. 89]

\[
fx(x; \alpha, \beta, 1, 0; S(0)) = \begin{cases} 
\frac{1}{\pi} \sum_{n=1}^{\infty} (-1)^{n-1} \frac{\Gamma(n+1)}{\Gamma(n+1)} \sin(n\pi \rho x) x^{n-1}, & \text{if } \alpha > 1, \beta \in [-1, 1], x \in \mathbb{R}, \\
\frac{1}{\pi} \sum_{n=1}^{\infty} (-1)^{n-1} nb_n x^{n-1}, & \text{if } \alpha = 1, \beta \in (0, 1], x \in \mathbb{R}, \\
\frac{1}{\pi} \frac{\Gamma(n\alpha+1)}{\Gamma(n+1)} \sin(n\pi \rho \alpha) x^{-n\alpha-1}, & \text{if } \alpha < 1, \beta \in [-1, 1], x \in \mathbb{R}^+, 
\end{cases} \tag{D.1}
\]

where the coefficients $b_n$ are given by

\[
b_n = \frac{1}{\Gamma(n+1)} \int_0^\infty \exp(-\beta u \ln u) u^{n-1} \sin \left( (1 + \beta) u \frac{\pi}{2} \right) du. \tag{D.2}\]

In addition, the distribution function of an $\alpha$-Stable severity model can be evaluated pointwise according to

\[
Fx(x; \alpha, \beta, 1, 0; S(0)) = \begin{cases} 
C(\alpha, \theta) + \frac{\epsilon(\alpha)}{2} \int_0^1 \exp \left( -x^{\frac{\alpha}{\alpha-1}} U_\alpha(\varphi, \theta) \right) d\varphi, & \text{if } \alpha \neq 1 \text{ and } x > 0, \\
\frac{1}{2} \int_{-1}^1 \exp \left( -\exp \left( -\frac{x}{\beta} \right) U_1(\varphi, \beta) \right) d\varphi, & \text{if } \alpha = 1, \text{ and } \beta > 0,
\end{cases} \tag{D.3}
\]

otherwise in all other cases it suffices to utilize the duality principle of infinitely divisible stable distributions which has the consequence that

\[
Fx(-x; \alpha, \beta, 1, 0; S(0)) + Fx(x; \alpha, -\beta, 1, 0; S(0)) = 1. \tag{D.4}
\]

Note, the notation of Zolotarev (1983) [page 74] is adopted above in which

\[
\epsilon(\alpha) = \text{sgn}(1 - \alpha), \quad K(\alpha) = \alpha - 1 + \text{sgn}(1 - \alpha), \quad \theta = \beta \frac{K(\alpha)}{\alpha}, \quad C(\alpha, \theta) = 1 - \frac{1}{4} (1 + \theta) (1 + \epsilon(\alpha)), \\
U_\alpha(\varphi, \theta) = \left( \frac{\sin \frac{\pi}{2} \alpha (\varphi + \theta)}{\cos \frac{\pi}{2} \varphi} \right)^{\frac{1}{\alpha}} \frac{\cos \frac{\pi}{2} ((\alpha - 1) \varphi + \alpha \theta)}{\cos \frac{\pi}{2} \varphi}.
\]
Appendix E. Parameter Estimation for the GPD Model

In this section of the technical appendix we detail the reparametrized MLE method for estimation of the GPD model parameters and also the Method of Moments based estimators for the GPD parameterization we consider in the paper.

Appendix E.1. Generalized Pareto Distribution (GPD) Model Estimation

The preparation of the volume profile data for fitting the GPD model involved taking intra-daily data for each asset over the period 2010 and considering varying time resolutions of 1 second, 2 seconds, 5 seconds and 10 seconds across each trading day in 2010. These trading days were then split into blocks and a Peaks over Threshold (POT) approach was adopted in which the last volume recorded in the specified time increment is retained only if it exceeds the threshold. Comparing this to the preparation of data in Sections 5.1 and 5.2, we can see that the α-Stable estimation considered the full intraday data sets, GEV considered the maximum volume per sub-sample time increment and GPD considered only the largest percentage of volume defined by the specified threshold. From this data preparation the following estimation procedures were considered.

Under the assumption that the volumes collected from the exceedence data are i.i.d. in the POT’s approach, the likelihood for the GPD as a function of the absolute exceedence data is given for the case in which $1 + \frac{\gamma Y_j}{\sigma} > 0$, by

$$
\ln l(Y; \gamma, \sigma) = -J \log \sigma - \left( \frac{1}{\gamma} + 1 \right) \sum_{j=1}^{J} \log \left( 1 + \frac{\gamma Y_j}{\sigma} \right), \quad (E.1)
$$

where the condition $1 + \frac{\gamma Y_j}{\sigma} > 0$ ensures the log likelihood is finite.

If $\gamma = 0$, the likelihood is given according to the exponential based distribution,

$$
\ln l(Y; 0, \sigma) = -J \log \sigma - \frac{1}{\sigma} \sum_{j=1}^{J} Y_j. \quad (E.2)
$$

Given the likelihood and the moments of the GPD distribution or the Quantile function, there are numerous statistical approaches one could adopt to perform the parameter estimation. First we discuss how to perform maximum likelihood estimation for such models. Maximization of the GPD likelihood provided in Equations E.1 and E.2 with respect to the parameters $\gamma$ and $\sigma$ is subject to the constraints:

1. $\sigma > 0$
2. $1 + \gamma y_{(J)}/\sigma > 0$ where $y_{(J)} = \max \{y_1, y_2, \ldots, y_J\}$.

This second constraint is important since one observes that if $\gamma < -1$ as $-\sigma/\gamma \rightarrow y_{(J)}$ then the likelihood approaches infinity. Hence, to obtain maximum likelihood parameter estimates, one should maximize the likelihood subject to these constraints and $\gamma \geq -1$. It is well known that one can in fact reparametrize the GPD likelihood to aid in the numerical stability of the parameter estimation via an MLE approach. A reparameterized MLE version is detailed in Appendix E.2, along with the GPD estimation of parameters via Method of Moments which is analytic for the parameterization we consider.

Appendix E.2. Reparameterization of the GPD Log-Likelihood and Maximisation

In practice it is beneficial to consider a reparameterization of the GPD log-likelihood function according to

$$
(\gamma, \sigma) \rightarrow (\gamma, \tau) = \left( \gamma, \frac{\gamma}{\sigma} \right), \quad (E.3)
$$
producing a reparameterized log-likelihood model given by

$$\ln l(Y; \gamma, \tau) = -J \ln \gamma + J \ln \tau - \left( \frac{1}{\gamma} + 1 \right) \sum_{i=1}^{J} \ln \left(1 + \tau Y_i \right).$$  \hspace{1cm} \text{(E.4)} $$

This log-likelihood is then maximized subject to $\tau < 1/y(J)$ and $\gamma \geq -1$. Under the first partial derivative this produces,

$$\frac{\partial \ln l(Y; \gamma, \tau)}{\partial \gamma} = 0; \Rightarrow \gamma(\tau) = \frac{1}{J} \sum_{j=1}^{J} \ln(1 - \tau y_j).$$  \hspace{1cm} \text{(E.5)} $$

Hence, the estimation is performed in two steps:

1. Estimate $\hat{\tau}_{MLE} = \arg \max \ln l(\gamma(\tau), \tau)$ subject to $\tau < 1/y(J)$.
2. Estimate $\hat{\gamma}_{MLE} = \frac{1}{J} \sum_{j=1}^{J} \ln(1 - \hat{\tau}_{MLE} y_j)$. Then solve for the original parameterization via inversion

$$\hat{\sigma}_{MLE} = \hat{\sigma}_{MLE}.$$

Note that the log likelihood $\ln l(\gamma(\tau), \tau)$ is continuous at $\tau = 0$, hence if the estimator $\hat{\tau}_{MLE} = 0$ then one should consider $\hat{\gamma}_{MLE} = 0$ and

$$\hat{\sigma}_{MLE} = \frac{1}{J} \sum_{j=1}^{J} y_j.$$  \hspace{1cm} \text{(E.6)} $$

In addition, in practice to ensure that $\gamma \geq 1$, the condition that $\tau < 1/y(J)$ should be modified to $\tau < (1 - \epsilon)/y(J)$, where $\epsilon$ is found from the condition that $\gamma(\tau) \geq -1$.

The covariances of the MLE parameters can be estimate using the inverse of the “observed Fisher information matrix” given for $J$ realized data samples $y = [y_1, y_2, \ldots, y_J]$ for the $(k, m)$-th element by

$$\hat{I}(\hat{\theta})_{k,m} = -\frac{1}{J} \sum_{j=1}^{J} \frac{\partial^2 \ln l(y_j; \theta)}{\partial \theta_k \partial \theta_m} \bigg|_{\theta = \hat{\theta}}.$$  \hspace{1cm} \text{(E.7)} $$

In the case of the GPD MLE parameter estimation the required second order partial derivatives are given by:

$$\frac{\partial^2}{\partial \gamma^2} \ln l(y; \gamma, \sigma) = -2 \sum_{j=1}^{J} \frac{y_j}{\sigma + \gamma y_j} + \left(1 + \frac{1}{\gamma} \right) \sum_{j=1}^{J} \frac{y_j^2}{(\sigma + \gamma y_j)^2}$$

$$\frac{\partial^2}{\partial \sigma^2} \ln l(y; \gamma, \sigma) = -\frac{1}{\sigma} (1 + \gamma) \sum_{j=1}^{J} \frac{y_j}{(\sigma + \gamma y_j)^2}$$

$$\frac{\partial^2}{\partial \sigma \partial \gamma} \ln l(y; \gamma, \sigma) = \sum_{j=1}^{J} \frac{y_j}{(\sigma + \gamma y_j)^2} - \frac{1}{\sigma} \sum_{j=1}^{J} \frac{y_j^2}{(\sigma + \gamma y_j)^2}.$$  \hspace{1cm} \text{(E.8)} $$

**Remark 4.** It has been shown in Smith (1985) and Embrechts et al. (1997) [Section 6.5.1] that in the case in which $\gamma > -1/2$ the MLE vector $(\hat{\gamma}_{MLE}, \hat{\sigma}_{MLE})$ is asymptotically consistent and distributed according to
a bivariate Gaussian distribution with asymptotic covariance, obtained using the MLE parameter estimates and the inverse Fisher information matrix is given analytically for the \((k, m)\)-th element as

\[
I(\hat{\theta})_{k,m} = E \left[ \frac{\partial^2}{\partial \theta_k \partial \theta_m} \ln f(Y_1 | \theta) \right].
\]  

(E.9)

where the limit asymptotically is given by,

\[
\lim_{J \to \infty} I(\hat{\theta}) \bigg|_{\theta = \hat{\theta}_{MLE}} = \frac{1}{J} \left[ \frac{(1 + \gamma)^2}{(1 - \gamma)(1 - 2\gamma)} \right] \bigg|_{\theta = \hat{\theta}_{MLE}}.
\]

(E.10)

Appendix E.3. Moment Matching Solutions to GPD Parameter Estimation

The approach of moment matching (MOM) in the GPD model was first proposed in Hosking and Wallis (1987) and utilizes the fact that expressions for the moments of the GPD distribution are known and exist for the \(r\)-th moment when \(\gamma < \frac{1}{r}\) as non-linear functions of the GPD shape and scale parameters. It has been shown that, given they exist, the mean, variance, skew and kurtosis of a random variable with GPD distribution are given by,

\[
E[X] = \frac{\sigma}{1 - \gamma}, \quad \text{Var}(X) = \frac{\sigma^2}{(1 - \gamma)^2(1 - 2\gamma)},
\]

\[
\text{Skewness}(X) = \frac{2(1 + \gamma)(l - 2\gamma)^{1/2}}{(1 - 3\gamma)}, \quad \text{Kurtosis}(X) = \frac{3(1 - 2\gamma)(3 + \gamma - 2\gamma^2)}{(1 - 3\gamma)(l - 4\gamma)} - 3.
\]  

(E.11)

Hence, the estimators of the shape and scale parameters obtained from MOM are given according to Definition 8.

Definition 8 (Method Of Moments GPD Model). The estimators for the shape and scale parameters in the GPD severity distribution by the MOM are given, for observed realized LOB volumes prepared under the POT’s exceedences method and denoted as \(\{x_i\}_{i=1}^n\), by the estimators,

\[
\hat{\gamma}_{MOM} = \frac{1}{2} \left( 1 - \frac{\left( \frac{1}{n} \sum_{i=1}^n x_i \right)^2}{\left( \frac{1}{n-1} \sum_{i=1}^n \left( x_i - \frac{1}{n} \sum_{j=1}^n x_j \right) \right)^2} \right),
\]

\[
\hat{\sigma}_{MOM} = \frac{1}{2n} \sum_{i=1}^n x_i \left( 1 + \frac{\left( \frac{1}{n} \sum_{i=1}^n x_i \right)^2}{\left( \frac{1}{n-1} \sum_{i=1}^n \left( x_i - \frac{1}{n} \sum_{j=1}^n x_j \right) \right)^2} \right).
\]  

(E.12)
Appendix F. Model Estimation Synthetic Case Studies to Assess Different Estimation Procedures for GEV and GPD models

Here we present the model estimation results for the synthetic case studies used to assess the behaviors of the different statistical estimation procedures for the GEV and GPD models.

Appendix F.1. Synthetic Case Studies for GEV models

To further explore the discrepancies between the MLE and Mixed L-Moments, specifically the upward translation of the shape parameter by approximately 2, we performed a simulation study which considered sample sizes of 50 and 10,000 randomly generated GEV distributed data series. We consider 20 different simulations to estimate the parameters of the GEV distribution using MLE and Mixed L-Moments methods. These results demonstrate the same discrepancies between the different estimation methods for the shape parameter. As $\gamma \to 0$, we see an increased bias of $\gamma$ under the Mixed L-Moments method, but the tradeoff is that the variance is reduced in the Mixed L-Moments method. Mixed L-Moments becomes more reliable as the sample size decreases.

Figure F.25: GEV parameter estimation using MLE, Mixed L-Moments methods. 20 simulations of 50 randomly generated GEV data series.
Appendix F.2. Synthetic Case Studies for GPD models

To further explore why the EPM method gives substantially different results to the MLE and Pickands methods, we consider a simulation study with 20 data sets, each of 500 randomly generated GP data points using the MLE Pickands and EPM estimation methods. We consider the impact of setting the shape parameter positive and negative.

The results show that for a positive shape parameter for the three methods appear to be consistent, however when the actual shape parameter is negative we see a significant translation upwards, with increased variability in the estimator, consistent with results for the observed three methods when analyzing the real LOB volume data.

Figure F.27: GPD parameter estimation using MLE, Pickands and EPM (All percentiles solutions included) methods for 20 simulations of 500 randomly generated GP data series.
Figure F.28: GPD parameter estimation using MLE, Pickands and EPM (> 50th percentile solutions included) methods for 20 simulations of 500 randomly generated GP data series.

Figure F.29: GPD parameter estimation using MLE, Pickands and EPM (> 75th percentile solutions included) methods for 20 simulations of 500 randomly generated GP data series.
Appendix G. Kolmogorov-Smirnov Test Case Study to Assess Goodness of Fit

Under the null hypothesis, the KS test quantifies a distance between the empirical distribution function and the theoretical $\alpha$-Stable distribution function. The results from the KS test can be misleading due to the fact that one should adjust for the fact that infinite mean and infinite variance models are considered in this tail analysis. That is, sub-exponential models will require an adjustment to the distribution of the test statistic to account for the significant contribution in the tails, see discussion in Chickeportiche and Bouchaud (2012). Furthermore, we note that this analysis is further complicated by the massive data sample size obtained for all intra-daily samples at 10 sec for each trading day, resulting in each hypothesis being formed from thousands of samples. Therefore, the chance of rejection will significantly increase.

To investigate this attribute further, we perform a case study of the KS test with a sample size of 200 observations versus the full sample for each day, 3888 observations. We utilize the actual LOB volume data from BOBL and the results derived from the $\alpha$-Stable distribution discussed in Appendix I.1. We can see from Figure G.30 that the p-value changes significantly when using a subset of the observations, demonstrating the sample size effect in the KS test. Using a subset of the observations, we find that the average p-value across the entire year is 0.3388 for the bid side and 0.3132 for the ask side, rejecting the null hypothesis approximately 27% of the trading days for at the 0.1 level of significance. Whereas using the full set of observations, we have an average p-value of 0.0290 for the bid side and 0.0294 for the ask side, rejecting the null hypothesis 91% of the trading days.

![Figure G.30: $\alpha$-Stable KS test p-value for the year 2010 using McCullochs method for BOBL bid side at time resolution of 10 seconds. For each side we consider the p-value for the full sample for each day (sample size 3888) and a random selection of data points (sample size 200) within each day to highlight the sample size effect on the KS test.](image-url)
Appendix H. Model Estimation and Empirical Results for Volume Profile of 5 Year T-Note

Here we present the model estimation results for the 5 Year US T-Notes daily estimation and analysis in 2010.

Appendix H.1. McCulloch’s Quantile Based Estimation - Alpha-Stable Model Results

In this section we present results for the McCulloch quantile based approach to estimation of the $\alpha$-Stable model parameters and assessment of the quality of the fit of these models to the limit order volume profile data.

Figure H.31: $\alpha$-Stable daily parameter estimation for the year 2010 using McCulloch’s method for 5YTN at a time resolution of 10 seconds. The red dashed line is the bid L1 and blue dashed line and the ask L1. Top Left Plot: Tail index parameter $\alpha$ daily estimates. Top Right Plot: Asymmetry parameter $\beta$ daily estimates. Bottom Left Plot: Scale parameter $\gamma$ daily estimates. Bottom Right Plot: Location parameter $\delta$ daily estimates.

Figure H.32: $\alpha$-Stable daily CDF plots for the year 2010 using McCulloch’s method for 5YTN bid and ask side at time resolution of 10 seconds.

Figure H.33 presents boxplots for the estimated parameteric model cdf and the empirical cdf fitted intra-daily for each trading day of the year. Each boxplot represents the $i$th percentile where $i \in \{0.1, 0.2, \ldots, 0.9, 0.95, 0.99\}$, which is comprised of the CDF each day across the year for that particular percentile. We then utilise the boxplot charts to assist in the assessment of the goodness-of-fit in light of...
the limitations of the KS test discussed in the case study in Appendix G. From Figure H.33 we can see that the $\alpha$-Stable model provides a suitable fit across all trading days, particularly in the tails which is the region of interest to this analysis.

**Figure H.33:** $\alpha$-Stable KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using McCulloch's method for 5YTN bid and ask side at time resolution of 10 seconds.

**Appendix H.2. Generalized Extreme Value (GEV) Distribution**

In this section we present first the Maximum Likelihood Estimation results for the reparameterized likelihood model and then the results based on a mixed estimation approach combining MLE and L-method of moments based solutions for the GEV model parameters of the LOB volume profiles.

**Appendix H.2.1. Maximum Likelihood Estimation - Reparameterized GEV Model Results**

**Figure H.34:** GEV daily parameter estimation using MLE method for 5YTN bid and ask side at time resolution of 10 seconds.
Figure H.35: GEV daily CDF plots for the year 2010 using MLE method for 5YTN bid and ask side at time resolution of 10 seconds.

Figure H.36: GEV KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLE method for 5YTN bid and ask side at time resolution of 10 seconds.

Appendix H.2.2. Mixed L-Moments Estimation - GEV Model Results

Figure H.37: GEV daily parameter estimation using MLM method for 5YTN bid and ask side at time resolution of 10 seconds.
Figure H.38: GEV daily CDF plots for the year 2010 using MLM method for 5YTN bid and ask side at time resolution of 10 seconds.

Figure H.39: GEV KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLM method for 5YTN bid and ask side at time resolution of 10 seconds.

Appendix H.3. Generalized Pareto Distribution (GPD)

In this section we present first the Maximum Likelihood Estimation results for the reparameterized likelihood model and then the results based on both the Pickands estimator and also the Empirical Percentile Method estimation results for the GPD model parameters of the LOB volume profiles.
Appendix H.3.1. Maximum Likelihood Estimation - Reparameterized GPD Model Results

Figure H.40: GPD daily parameter estimation using MLE method for 5YTN bid and ask side at time resolution of 10 seconds.

Figure H.41: GPD daily CDF plots for the year 2010 using MLE method for 5YTN bid and ask side at time resolution of 10 seconds.

Figure H.42: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLE method for 5YTN bid and ask side at time resolution of 10 seconds.
Appendix H.3.2. Pickands Estimator - GPD Model Results

Figure H.43: GPD daily parameter estimation using Pickands method for 5YTN bid and ask side at time resolution of 10 seconds.

Figure H.44: GPD daily CDF plots for the year 2010 using Pickands method for 5YTN bid and ask side at time resolution of 10 seconds.

Figure H.45: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using Pickands method for 5YTN bid and ask side at time resolution of 10 seconds.
Appendix H.3.3. Empirical Percentile Method - GPD Model Results

Figure H.46: GPD daily parameter estimation using EPM method for BOBL bid and ask side at time resolution of 10 seconds.

Figure H.47: GPD daily parameter estimation using EPM method for 5YTN bid and ask side at time resolution of 10 seconds.

Figure H.48: GPD daily CDF plots for the year 2010 using EPM method for 5YTN bid and ask side at time resolution of 10 seconds.
Figure H.49: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using EPM method for 5YTN bid and ask side at time resolution of 10 seconds.
Appendix I. Model Estimation and Empirical Results for Volume Profile of Euro-BOBL

Here we present the model estimation results for the BOBL daily estimation and analysis in 2010.

Appendix I.1. McCulloch’s Quantile Based Estimation - Alpha-Stable Model Results

In this section we present results for the McCulloch quantile based approach to estimation of the α-Stable model parameters and assessment of the quality of the fit of these models to the limit order volume profile data.

![Graphs showing parameter estimation](image1)

**Figure I.50:** α-Stable daily parameter estimation for the year 2010 using McCullochs method for BOBL at a time resolution of 10 seconds. The red dashed line is the bid L1 and blue dashed line is the ask L1. **Top Left Plot:** Tail index parameter α daily estimates. **Top Right Plot:** Asymmetry parameter β daily estimates. **Bottom Left Plot:** Scale parameter γ daily estimates. **Bottom Right Plot:** Location parameter δ daily estimates.

![Graphs showing CDF plots](image2)

**Figure I.51:** α-Stable daily CDF plots for the year 2010 using McCullochs method for BOBL bid and ask side at time resolution of 10 seconds.
Figure I.52: $\alpha$-Stable KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using McCulloch's method for BOBL bid and ask side at time resolution of 10 seconds.

Appendix I.2. Generalized Extreme Value (GEV) Distribution

In this section we present first the Maximum Likelihood Estimation results for the reparameterized likelihood model and then the results based on a mixed estimation approach combining MLE and L-method of moments based solutions for the GEV model parameters of the LOB volume profiles.

Appendix I.2.1. Maximum Likelihood Estimation - Reparameterized GEV Model Results

Figure I.53: GEV daily parameter estimation using MLE method for BOBL bid and ask side at time resolution of 10 seconds.
Figure I.54: GEV daily CDF plots for the year 2010 using MLE method for BOBL bid and ask side at time resolution of 10 seconds.

Figure I.55: GEV KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLE method for BOBL bid and ask side at time resolution of 10 seconds.

Appendix I.2.2. Mixed L-Moments Estimation - GEV Model Results

Figure I.56: GEV daily parameter estimation using MLM method for BOBL bid and ask side at time resolution of 10 seconds.
Figure I.57: GEV daily CDF plots for the year 2010 using MLM method for BOBL bid and ask side at time resolution of 10 seconds.

Figure I.58: GEV KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLM method for BOBL bid and ask side at time resolution of 10 seconds.

Appendix I.3. Generalized Pareto Distribution (GPD)

In this section we present first the Maximum Likelihood Estimation results for the reparameterized likelihood model and then the results based on both the Pickands estimator and also the Empirical Percentile Method estimation results for the GPD model parameters of the LOB volume profiles.
Appendix I.3.1. Maximum Likelihood Estimation - Reparameterized GPD Model Results

Figure I.59: GPD daily parameter estimation using MLE method for BOBL bid and ask side at time resolution of 10 seconds.

Figure I.60: GPD daily CDF plots for the year 2010 using MLE method for BOBL bid and ask side at time resolution of 10 seconds.

Figure I.61: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLE method for BOBL bid and ask side at time resolution of 10 seconds.
Appendix I.3.2. Pickands Estimator - GPD Model Results

Figure I.62: GPD daily parameter estimation using Pickands method for BOBL bid and ask side at time resolution of 10 seconds.

Figure I.63: GPD daily CDF plots for the year 2010 using Pickands method for BOBL bid and ask side at time resolution of 10 seconds.

Figure I.64: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using Pickands method for BOBL bid and ask side at time resolution of 10 seconds.
Appendix I.3.3. Empirical Percentile Method - GPD Model Results

Figure I.65: GPD daily parameter estimation using EPM method for BOBL bid and ask side at time resolution of 10 seconds.

Figure I.66: GPD daily CDF plots for the year 2010 using EPM method for BOBL bid and ask side at time resolution of 10 seconds.

Figure I.67: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using EPM method for BOBL bid and ask side at time resolution of 10 seconds.
Appendix J. Model Estimation and Empirical Results for Volume Profile of E-mini S&P 500

Appendix J.1. McCulloch’s Quantile Based Estimation - Alpha-Stable Model Results

In this section we present results for the McCulloch quantile based approach to estimation of the $\alpha$-Stable model parameters and assessment of the quality of the fit of these models to the limit order volume profile data.

Figure J.68: $\alpha$-Stable daily parameter estimation for the year 2010 using McCulloch's method for SP500 at a time resolution of 10 seconds. The red dashed line is the bid L1 and blue dashed line is the ask L1. **Top Left Plot:** Tail index parameter $\alpha$ daily estimates. **Top Right Plot:** Asymmetry parameter $\beta$ daily estimates. **Bottom Left Plot:** Scale parameter $\gamma$ daily estimates. **Bottom Right Plot:** Location parameter $\delta$ daily estimates.

Figure J.69: $\alpha$-Stable daily CDF plots for the year 2010 using McCulloch's method for SP500 bid and ask side at time resolution of 10 seconds.
Figure J.70: $\alpha$-Stable KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using McCulloch's method for SP500 bid and ask side at time resolution of 10 seconds.

Appendix J.2. Generalized Extreme Value (GEV) Distribution

In this section we present first the Maximum Likelihood Estimation results for the reparameterized likelihood model and then the results based on a mixed estimation approach combining MLE and L-method of moments based solutions for the GEV model parameters of the LOB volume profiles.

Appendix J.2.1. Maximum Likelihood Estimation - Reparameterized GEV Model Results

Figure J.71: GEV daily parameter estimation using MLE method for SP500 bid and ask side at time resolution of 10 seconds.
Figure J.72: GEV daily CDF plots for the year 2010 using MLE method for SP500 bid and ask side at time resolution of 10 seconds.

Figure J.73: GEV KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLE method for SP500 bid and ask side at time resolution of 10 seconds.

Appendix J.2.2. Mixed L-Moments Estimation - GEV Model Results

Figure J.74: GEV daily parameter estimation using MLM method for SP500 bid and ask side at time resolution of 10 seconds.
Figure J.75: GEV daily CDF plots for the year 2010 using MLM method for SP500 bid and ask side at time resolution of 10 seconds.

Figure J.76: GEV KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLM method for SP500 bid and ask side at time resolution of 10 seconds.

Appendix J.3. Generalized Pareto Distribution (GPD)

In this section we present first the Maximum Likelihood Estimation results for the reparameterized likelihood model and then the results based on both the Pickands estimator and also the Empirical Percentile Method estimation results for the GPD model parameters of the LOB volume profiles.
Appendix J.3.1. Maximum Likelihood Estimation - Reparameterized GPD Model Results

Figure J.77: GPD daily parameter estimation using MLE method for SP500 bid and ask side at time resolution of 10 seconds.

Figure J.78: GPD daily CDF plots for the year 2010 using MLE method for SP500 bid and ask side at time resolution of 10 seconds.

Figure J.79: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLE method for SP500 bid and ask side at time resolution of 10 seconds.
Appendix J.3.2. Pickands Estimator - GPD Model Results

Figure J.80: GPD daily parameter estimation using Pickands method for SP500 bid and ask side at time resolution of 10 seconds.

Figure J.81: GPD daily CDF plots for the year 2010 using Pickands method for SP500 bid and ask side at time resolution of 10 seconds.

Figure J.82: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using Pickands method for SP500 bid and ask side at time resolution of 10 seconds.
Appendix J.3.3. Empirical Percentile Method - GPD Model Results

Figure J.83: GPD daily parameter estimation using EPM method for SP500 bid and ask side at time resolution of 10 seconds.

Figure J.84: GPD daily CDF plots for the year 2010 using EPM method for SP500 bid and ask side at time resolution of 10 seconds.

Figure J.85: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using EPM method for SP500 bid and ask side at time resolution of 10 seconds.
Appendix K. Model Estimation and Empirical Results for Volume Profile of SIMEX Nikkei 225

Appendix K.1. McCulloch’s Quantile Based Estimation - Alpha-Stable Model Results

In this section we present results for the McCulloch quantile based approach to estimation of the $\alpha$-Stable model parameters and assessment of the quality of the fit of these models to the limit order volume profile data.

Figure K.86: $\alpha$-Stable daily parameter estimation for the year 2010 using McCullochs method for NIKKEI at a time resolution of 10 seconds. The red dashed line is the bid L1 and blue dashed line is the ask L1. **Top Left Plot:** Tail index parameter $\alpha$ daily estimates. **Top Right Plot:** Asymmetry parameter $\beta$ daily estimates. **Bottom Left Plot:** Scale parameter $\gamma$ daily estimates. **Bottom Right Plot:** Location parameter $\delta$ daily estimates.

Figure K.87: $\alpha$-Stable daily CDF plots for the year 2010 using McCullochs method for NIKKEI bid and ask side at time resolution of 10 seconds.
Appendix K.2. Generalized Extreme Value (GEV) Distribution

In this section we present first the Maximum Likelihood Estimation results for the reparameterized likelihood model and then the results based on a mixed estimation approach combining MLE and L-method of moments based solutions for the GEV model parameters of the LOB volume profiles.

Appendix K.2.1. Maximum Likelihood Estimation - Reparameterized GEV Model Results

Figure K.89: GEV daily parameter estimation using MLE method for NIKKEI bid and ask side at time resolution of 10 seconds.
Appendix K.2.2. Mixed L-Moments Estimation - GEV Model Results

Figure K.90: GEV daily CDF plots for the year 2010 using MLE method for NIKKEI bid and ask side at time resolution of 10 seconds.

Figure K.91: GEV KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLE method for NIKKEI bid and ask side at time resolution of 10 seconds.

Figure K.92: GEV daily parameter estimation using MLM method for NIKKEI bid and ask side at time resolution of 10 seconds.
Appendix K.3. Generalized Pareto Distribution (GPD)

In this section we present first the Maximum Likelihood Estimation results for the reparameterized likelihood model and then the results based on both the Pickands estimator and also the Empirical Percentile Method estimation results for the GPD model parameters of the LOB volume profiles.
Appendix K.3.1. Maximum Likelihood Estimation - Reparameterized GPD Model Results

Figure K.95: GPD daily parameter estimation using MLE method for NIKKEI bid and ask side at time resolution of 10 seconds.

Figure K.96: GPD daily CDF plots for the year 2010 using MLE method for NIKKEI bid and ask side at time resolution of 10 seconds.

Figure K.97: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLE method for NIKKEI bid and ask side at time resolution of 10 seconds.
Appendix K.3.2. Pickands Estimator - GPD Model Results

Figure K.98: GPD daily parameter estimation using Pickands method for NIKKEI bid and ask side at time resolution of 10 seconds.

Figure K.99: GPD daily CDF plots for the year 2010 using Pickands method for NIKKEI bid and ask side at time resolution of 10 seconds.

Figure K.100: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using Pickands method for NIKKEI bid and ask side at time resolution of 10 seconds.
Appendix K.3.3. Empirical Percentile Method - GPD Model Results

Figure K.101: GPD daily parameter estimation using EPM method for NIKKEI bid and ask side at time resolution of 10 seconds.

Figure K.102: GPD daily CDF plots for the year 2010 using EPM method for NIKKEI bid and ask side at time resolution of 10 seconds.

Figure K.103: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using EPM method for NIKKEI bid and ask side at time resolution of 10 seconds.
Appendix L. Model Estimation and Empirical Results for Volume Profile of Gold

Appendix L.1. McCulloch’s Quantile Based Estimation - Alpha-Stable Model Results

In this section we present results for the McCulloch quantile based approach to estimation of the \(\alpha\)-Stable model parameters and assessment of the quality of the fit of these models to the limit order volume profile data.

---

**Figure L.104:** \(\alpha\)-Stable daily parameter estimation for the year 2010 using McCulloch’s method for GOLD at a time resolution of 10 seconds. The red dashed line is the bid L1 and blue dashed line is the ask L1. **Top Left Plot:** Tail index parameter \(\alpha\) daily estimates. **Top Right Plot:** Asymmetry parameter \(\beta\) daily estimates. **Bottom Left Plot:** Scale parameter \(\gamma\) daily estimates. **Bottom Right Plot:** Location parameter \(\delta\) daily estimates.

**Figure L.105:** \(\alpha\)-Stable daily CDF plots for the year 2010 using McCulloch’s method for GOLD bid and ask side at time resolution of 10 seconds.
Appendix L.2. Generalized Extreme Value (GEV) Distribution

In this section we present first the Maximum Likelihood Estimation results for the reparameterized likelihood model and then the results based on a mixed estimation approach combining MLE and L-method of moments based solutions for the GEV model parameters of the LOB volume profiles.

Appendix L.2.1. Maximum Likelihood Estimation - Reparameterized GEV Model Results

Figure L.106: $\alpha$-Stable KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using McCullochs method for GOLD bid and ask side at time resolution of 10 seconds.

Figure L.107: GEV daily parameter estimation using MLE method for GOLD bid and ask side at time resolution of 10 seconds.
Figure L.108: GEV daily CDF plots for the year 2010 using MLE method for GOLD bid and ask side at time resolution of 10 seconds.

Figure L.109: GEV KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLE method for GOLD bid and ask side at time resolution of 10 seconds.

Appendix L.2.2. Mixed L-Moments Estimation - GEV Model Results

Figure L.110: GEV daily parameter estimation using MLM method for GOLD bid and ask side at time resolution of 10 seconds.
Appendix L.3. Generalized Pareto Distribution (GPD)

In this section we present first the Maximum Likelihood Estimation results for the reparameterized likelihood model and then the results based on both the Pickands estimator and also the Empirical Percentile Method estimation results for the GPD model parameters of the LOB volume profiles.
Appendix L.3.1. Maximum Likelihood Estimation - Reparameterized GPD Model Results

Figure L.113: GPD daily parameter estimation using MLE method for GOLD bid and ask side at time resolution of 10 seconds.

Figure L.114: GPD daily CDF plots for the year 2010 using MLE method for GOLD bid and ask side at time resolution of 10 seconds.

Figure L.115: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLE method for GOLD bid and ask side at time resolution of 10 seconds.
Appendix L.3.2. Pickands Estimator - GPD Model Results

Figure L.116: GPD daily parameter estimation using Pickands method for GOLD bid and ask side at time resolution of 10 seconds.

Figure L.117: GPD daily CDF plots for the year 2010 using Pickands method for GOLD bid and ask side at time resolution of 10 seconds.

Figure L.118: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using Pickands method for GOLD bid and ask side at time resolution of 10 seconds.
Appendix L.3.3. Empirical Percentile Method - GPD Model Results

Figure L.119: GPD daily parameter estimation using EPM method for GOLD bid and ask side at time resolution of 10 seconds.

Figure L.120: GPD daily CDF plots for the year 2010 using EPM method for GOLD bid and ask side at time resolution of 10 seconds.

Figure L.121: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using EPM method for GOLD bid and ask side at time resolution of 10 seconds.
Appendix M. Model Estimation and Empirical Results for Volume Profile of Silver

Appendix M.1. McCulloch’s Quantile Based Estimation - Alpha-Stable Model Results

In this section we present results for the McCulloch quantile based approach to estimation of the $\alpha$-Stable model parameters and assessment of the quality of the fit of these models to the limit order volume profile data.

Figure M.122: $\alpha$-Stable daily parameter estimation for the year 2010 using McCulloch's method for SILVER at a time resolution of 10 seconds. The red dashed line is the bid L1 and blue dashed line is the ask L1. **Top Left Plot:** Tail index parameter $\alpha$ daily estimates. **Top Right Plot:** Asymmetry parameter $\beta$ daily estimates. **Bottom Left Plot:** Scale parameter $\gamma$ daily estimates. **Bottom Right Plot:** Location parameter $\delta$ daily estimates.

Figure M.123: $\alpha$-Stable daily CDF plots for the year 2010 using McCulloch's method for SILVER bid and ask side at time resolution of 10 seconds.
Figure M.124: $\alpha$-Stable KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using McCulloch's method for SILVER bid and ask side at time resolution of 10 seconds.

Appendix M.2. Generalized Extreme Value (GEV) Distribution

In this section we present first the Maximum Likelihood Estimation results for the reparameterized likelihood model and then the results based on a mixed estimation approach combining MLE and L-method of moments based solutions for the GEV model parameters of the LOB volume profiles.

Appendix M.2.1. Maximum Likelihood Estimation - Reparameterized GEV Model Results

Figure M.125: GEV daily parameter estimation using MLE method for SILVER bid and ask side at time resolution of 10 seconds.
Figure M.126: GEV daily CDF plots for the year 2010 using MLE method for SILVER bid and ask side at time resolution of 10 seconds.

Figure M.127: GEV KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLE method for SILVER bid and ask side at time resolution of 10 seconds.

Appendix M.2.2. Mixed L-Moments Estimation - GEV Model Results

Figure M.128: GEV daily parameter estimation using MLM method for SILVER bid and ask side at time resolution of 10 seconds.
Appendix M.3. Generalized Pareto Distribution (GPD)

In this section we present first the Maximum Likelihood Estimation results for the reparameterized like-
lihood model and then the results based on both the Pickands estimator and also the Empirical Percentile
Method estimation results for the GPD model parameters of the LOB volume profiles.
Appendix M.3.1. Maximum Likelihood Estimation - Reparameterized GPD Model Results

Figure M.131: GPD daily parameter estimation using MLE method for SILVER bid and ask side at time resolution of 10 seconds.

Figure M.132: GPD daily CDF plots for the year 2010 using MLE method for SILVER bid and ask side at time resolution of 10 seconds.

Figure M.133: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using MLE method for SILVER bid and ask side at time resolution of 10 seconds.
Appendix M.3.2. Pickands Estimator - GPD Model Results

Figure M.134: GPD daily parameter estimation using Pickands method for SILVER bid and ask side at time resolution of 10 seconds.

Figure M.135: GPD daily CDF plots for the year 2010 using Pickands method for SILVER bid and ask side at time resolution of 10 seconds.

Figure M.136: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using Pickands method for SILVER bid and ask side at time resolution of 10 seconds.

98
Appendix M.3.3. Empirical Percentile Method - GPD Model Results

Figure M.137: GPD daily parameter estimation using EPM method for SILVER bid and ask side at time resolution of 10 seconds.

Figure M.138: GPD daily CDF plots for the year 2010 using EPM method for SILVER bid and ask side at time resolution of 10 seconds.

Figure M.139: GPD KS test statistic, theoretical CDF and empirical CDF boxplots for the year 2010 using EPM method for SILVER bid and ask side at time resolution of 10 seconds.