Price jump prediction in a limit order book
Ban Zheng, Eric Moulines, Frédéric Abergel

To cite this version:

HAL Id: hal-00684716
https://hal.archives-ouvertes.fr/hal-00684716v2
Submitted on 31 May 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Price Jump Prediction in a Limit Order Book

Ban Zheng\(^1,2,3\), Eric Moulines\(^2\), Frédéric Abergel\(^3\)

\(^1\)Natixis, Equity Markets, Paris, France
\(^2\)Département TSI, Télécom ParisTech, Paris, France
\(^3\)BNP Paribas Chair of Quantitative Finance, Ecole Centrale Paris, MAS Laboratory, Chatenay-Malabry, France

Email: ban.zheng@melix.net, eric.moulines@telecom-paristech.fr, frederic.abergel@ecp.fr

Received November 4, 2012; revised April 2, 2013; accepted April 26, 2013

ABSTRACT

A limit order book provides information on available limit order prices and their volumes. Based on these quantities, we give an empirical result on the relationship between the bid-ask liquidity balance and trade sign and we show that the liquidity balance on the best bid/best ask is quite informative for predicting the future market order’s direction. Moreover, we define price jump as a sell (buy) market order arrival which is executed at a price which is smaller (larger) than the best bid (best ask) price at the moment just after the precedent market order arrival. Features are then extracted related to limit order volumes, limit order price gaps, market order information and limit order event information. Logistic regression is applied to predict the price jump from the features of a limit order book. LASSO logistic regression is introduced to help us make variable selection from which we are capable to highlight the importance of different features in predicting the future price jump. In order to get rid of the intraday data seasonality, the analysis is based on two separated datasets: morning dataset and afternoon dataset. Based on an analysis on forty largest French stocks of CAC40, we find that trade sign and market order size as well as the liquidity on the best bid (best ask) are consistently informative for predicting the incoming price jump.

Keywords: Limit Order Book; High Frequency Trading; Price Jump; Trade-Through; Logistic Regression; LASSO

1. Introduction

The determination of jumps in financial time series already has a long history as a challenging, theoretically interesting and practically important problem. Be it from the point of view of the statistician trying to separate, in spot prices, those moves corresponding to “jumps” from those who are compatible with the hypothesis of a process with continuous paths, or from the point of view of the practitioner: market maker, algorithmic trader, arbitrageur, who is in dire need of knowing the direction and the amplitude of the next price change, there is a vast, still unsatisfied interest for this question.

Several attempts have been made at theorizing the observability of the difference between processes with continuous or discontinuous paths, and the major breakthrough in that direction is probably due to [1], who introduced the concept of bi-power variation, and showed that—in a nutshell—the occurrence of jumps can be seen in the limiting behavior as the time step goes to zero of the bi-power variation: for a process with continuous paths, this quantity should converge to (a multiple of) the instantaneous variance, and the existence of a possibly different limit will be caused by the occurrence of jumps.

Since then, many authors, in particular [2] have contributed to shed a better light on this phenomenon, and one can safely say that rigorous statistical tests for identifying continuous-time, real-valued processes with discontinuous paths are now available to the academic community as well as the applied researcher.

However, it is a fact that the physics of modern, electronic, order-driven markets is not easily recast in the setting of real-valued, continuous-time processes, and it is also a fact that the time series of price, no matter how high the sampling frequency, is not anymore the most complete and accurate type of information one can get from the huge set of financial data at our hands. In fact, a relatively recent trend of studies has emerged over the past 10 years, where the limit order book became the center of interest, and the price changes are but a byproduct of the more complicated set of changes on limit orders, market orders, cancellation of orders, etc. see e.g. [3-6] for the latest developments in the econophysics of order-driven markets. This new standpoint is quite
enlightening, in that the physics of price formation becomes much more apparent, but it calls for a drastic change in the basic modeling tools: prices now live on a discrete grid with a step size given by the tick, the changes in price occur at discrete times. Furthermore, a host of important events that affect the order book rather than the price itself, events which are therefore essential in understanding the driving forces of the price changes, now become observable, and their role in the price dynamics must be taken into account when one is interested in understanding the latter.

Our point of view is slightly different: rather than concentrate on the one-dimensional price time series, we want to model the dynamics in event time of the whole order book, and focus on some specific events that can be interpreted in terms of jumps. To do so, we shall depart from the classical definitions—if any such thing exists—of a jump in a financial time series, and restrict ourselves to the more natural, more realistic and also more prone to experimental validation, concept of an inter-trade price jump and trade-through.

By definition, an inter-trade price jump is defined as an event where a market order is executed at a price which is smaller (larger) than the best limit price on the Bid (Ask) just after the precedent market order arrival. An inter-trade price jump permits a limit order submitted at the best bid (best ask) just after a market order arrival to be surely executed by the next market order arrival. A trade-through corresponds to the arrival of a new market order, the size of which is larger than the quantity available at the best limit on the Bid (for a sell order) or Ask (for a buy order) side of the order book. By nature, such an order will imply an automatic and instantaneous price change, the value of which will be exactly the difference in monetary units between the best limit price before and after transaction on the relevant side of the order book. Trade-through can be interpreted as the instantaneous price change triggered by a market order, meanwhile, inter-trade price jump is post-trade market impact. Most of researches on limit order book are based on stocks and often relates to characterizing features such as liquidity, volatility and bid-ask spread instead of making prediction, see [7-15]. Trade-through has also been the object of several recent studies in the econometrics and finance literature, see e.g. [16] (for cross-sectional relationship study) and [17].

In this work, we investigate whether the order book shape is informative for the inter-trade price jump prediction and whether trade-through contributes to this prediction. Recently, many researchers propose machine learning methods to make prediction on limit order book. [18] presents a non-parametric model for trade sign (market order initiator) inference and they show that the limit order book shape and the historical trades size are informative for the trade sign prediction. [19] applies multi kernel learning with support vector machine in predicting the EURUSD price evolution from the limit order book information. Here, logistic regression is introduced to predict the occurrence of inter-trade price jump. Variable selection by LASSO logistic regression provides us an insight into the dynamics of the limit order book and allows us to select the most informative features for predicting relevant events. We will show that some features of the limit order book have strong predictive and explanatory power, allowing one to make a sound prediction of the occurrence of inter-trade price jump knowing the state of the limit order book. Trade-through is also confirmed to be quite informative for inter-trade price jump prediction. This result in itself is interesting in that it allows one to use the full set of available information in order to do some prediction: whereas the history of the price itself is known not to be a good predictor of the next price moves—the so-called efficiency of the market is relatively hard to beat when one only uses the price information—we shall show that the limit order volumes contain more information, and the market order size contributes also to an accurate prediction of inter-trade price jump.

This paper is organized as follows. Section 2 describes the main notations in a limit order book. Section 3 gives an empirical result on the relationship between bid-ask liquidity balance and trade sign. Section 4 introduces logistic regression for inter-trade price jump prediction and lasso logistic regression for variable selection. The conclusion is in Section 6.

2. Description and Data Notation

The Euronext market adopts NSC (Nouveau Système de Cotation) for electronic trading. During continuous trading from 9h00 to 17h30, NSC matches market orders against the best limit order on the opposite side. Various order types are accepted in NSC such as limit orders (an order to be traded at a fixed price with certain amount), market orders (order execution without price constraint), stop orders (issuing limit orders or market orders when a triggered price is reached) and iceberg orders (only a part of the size is visible in the book). Limit order is posted to electronic trading system and they are placed into the book according to their prices, see Figure 1. Market order is an order to be executed at the best available price in the limit order book. The lowest price of limit sell orders is called best ask; the highest price of limit buy orders is called best bid. The gap between the best bid and the best ask is called the spread. When a market buy order with price higher/equal than the best ask price, a trade occurs and the limit order book is updated accordingly. Limit orders can also be cancelled if there have not
been executed, so the limit order book can be modified due to limit order cancellation, limit order arrival or market order arrival. In case of iceberg orders, the disclosed part has the same priority as a regular limit order while the hidden part has lower priority. The hidden part will become visible as soon as the disclosed part is executed. The case that the hidden part is consumed by a market order without being visible is quite rare. In this study, we neglect stop orders and iceberg orders which are relatively rare compared to limit order and market order events.

In a limit order book, as shown in Figure 1, only a certain number of best buy/sell limit orders are available for public. We denote the number of available bid/ask limit prices by $L$.

In this study, for simplicity, we focus on limit order arrival events, limit order cancellation events and market order arrival events, see Figure 2. The number of visible limit order levels is chosen to be five $L = 5$. Our dataset is provided by NATIXIS via Thomson Reuter’s “Reuters Tick Capture Engine” and comprises of trades and limit order activities of the 40 member stocks of index CAC40 between April 1st 2011 and April 30th 2011. In order to get rid of open hour and close hour, we extract the data from 09h05 to 17h25. Every transaction and every limit order book modification are recorded in milliseconds. The data contains information on the $L$ best quotes on both bid and ask sides. The trade data and quotes data are matched.

Denote $t$ as a time index indicating all limit order book events. $P_{i,t}^{b,j}$ and $P_{i,t}^{a,j}$ for $i = 1, \ldots, L$ define the $i^{th}$ best log bid/ask quote instantaneously after the $t^{th}$ event. We denote $S_{i,t} = P_{i+1,t}^{b,j} - P_{i,t}^{b,j}$ the spread instantaneously after the $t^{th}$ event. $G_{i,t}^{b,j} = P_{i,t}^{b,j} - P_{i+1,t}^{b,j}$, $G_{i,t}^{a,j} = P_{i,t}^{a,j} - P_{i-1,t}^{a,j}$ for $i = 1, \ldots, L - 1$ define respectively the $i^{th}$ best bid (ask) limit price gap instantaneously after the $t^{th}$ event. Besides, $V_{i,t}^{b,j}$ and $V_{i,t}^{a,j}$ for $i = 1, \ldots, L$ denote the log limit order volume on the $i^{th}$ best bid/ask quote instantaneously after the $t^{th}$ event. The volume of trade is denoted by $V_{t}^{mo}$ ($V_{t}^{mo} = 0$ when there is no trade) and the price of trade is denoted by $P_{t}^{mo}$ ($P_{t}^{mo} = 0$ when there is no trade, $P_{t}^{mo} = P_{t}^{b,j}$ when a market order touches bid side and $P_{t}^{mo} = P_{t}^{a,j}$ when a market order touches ask side).

Moreover, we introduce six dummy variables $BLO_t$, $ALO_t$, $BMO_t$, $AMO_t$, $BTT_t$, and $ATT_t$ to indicate the direction of each event: bid side or ask side, respectively for limit order event ($BLO_t$ and $ALO_t$), market order event ($BMO_t$ and $AMO_t$) and trade-through event ($BTT_t$ and $ATT_t$). The definition of variables is detailed in Table 1.

In order to capture the high-frequency dynamics in quotes and depths, we define a $K$-dimensional vector

$$R_t = \left[ G_{t}^{b,L-1}, \ldots, G_{t}^{b,1}, S_{t}, G_{t}^{a,1}, \ldots, G_{t}^{a,L-1}, V_{t}^{b,L}, \ldots, V_{t}^{b,1}, V_{t}^{a,1}, \ldots, V_{t}^{a,L} \right].$$

Modelling log prices and log volumes instead of absolute values is suggested by [20] studying the statistical properties of market impacts and trades and can be found in many other empirical studies. Price and volume changes in log is interpreted as related changes in percentage.

Another vector of variables is denoted by

$$R_t^2 = [BMO_t, AMO_t, BLO_t, ALO_t, BTT_t, ATT_t],$$

indicating the nature of the $t^{th}$ event.

Table 2 provides a descriptive statistics of the data used in this paper. It comprises limit order events, market order events and inter-trade price jump events. The

---

**Figure 1.** Limit order book description. Limit order price is discretized by tick price.

---

**Table 1. Variable definitions.**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{i,t}^{b,j}$</td>
<td>the $i^{th}$ best log bid price just after the $t^{th}$ event</td>
</tr>
<tr>
<td>$P_{i,t}^{a,j}$</td>
<td>the $i^{th}$ best log ask price just after the $t^{th}$ event</td>
</tr>
<tr>
<td>$G_{i,t}^{b,j}$</td>
<td>the $i^{th}$ bid gap price just after the $t^{th}$ event</td>
</tr>
<tr>
<td>$G_{i,t}^{a,j}$</td>
<td>the $i^{th}$ ask gap price just after the $t^{th}$ event</td>
</tr>
<tr>
<td>$S_{t}$</td>
<td>the spread just after the $t^{th}$ event</td>
</tr>
<tr>
<td>$V_{i,t}^{b,j}$</td>
<td>log volume of the $i^{th}$ best bid quote just after the $t^{th}$ event</td>
</tr>
<tr>
<td>$V_{i,t}^{a,j}$</td>
<td>log volume of the $i^{th}$ best ask quote just after the $t^{th}$ event</td>
</tr>
<tr>
<td>$BLO_t$</td>
<td>dummy variable equal to 1 if the $t^{th}$ event is a limit order event at bid side</td>
</tr>
<tr>
<td>$ALO_t$</td>
<td>dummy variable equal to 1 if the $t^{th}$ event is a limit order event at ask side</td>
</tr>
<tr>
<td>$BMO_t$</td>
<td>dummy variable equal to 1 if the $t^{th}$ event is a market order event at bid side</td>
</tr>
<tr>
<td>$AMO_t$</td>
<td>dummy variable equal to 1 if the $t^{th}$ event is a market order event at ask side</td>
</tr>
<tr>
<td>$BTT_t$</td>
<td>dummy variable equal to 1 if the $t^{th}$ event is a trade-through event at bid side</td>
</tr>
<tr>
<td>$ATT_t$</td>
<td>dummy variable equal to 1 if the $t^{th}$ event is a trade-through event at ask side</td>
</tr>
</tbody>
</table>
analysis is done on two separated datasets: morning dataset (between 09h05 and 13h15) and afternoon dataset (between 13h15 and 17h25). We observe that there are more market order events in the afternoon than in the morning. Similarly, inter-trade price jump events are slightly more frequent in the afternoon than in the morning. However, trade-through events are more frequent in the morning than in the afternoon.


Before making an analysis on price jump prediction, we try to reveal whether the limit order volume information plays a role in determining the future market order’s direction (trade sign). In order to study the conditional probability given the knowledge about bid/ask limit order liquidity, we propose a bid-ask volume ratio corresponding to depth \( i \) just before the \( k^{th} \) trade, which is defined as \( W_{q-1}(i) (i \in \{1, \ldots, L\}) \), more precisely,

\[
W_{q-1}(i) = \log \left( \frac{\sum_{j=1}^{L} \exp(V_{q-1}^{b,j})}{\sum_{j=1}^{L} \exp(V_{q-1}^{a,j})} \right) = \log \left( \frac{\sum_{j=1}^{L} \exp(V_{q-1}^{b,j})}{\sum_{j=1}^{L} \exp(V_{q-1}^{a,j})} \right) - \log \left( \frac{\sum_{j=1}^{L} \exp(V_{q-1}^{a,j})}{\sum_{j=1}^{L} \exp(V_{q-1}^{b,j})} \right),
\]

where \( t_k \) is time index of the \( k^{th} \) market order event.

For all \( x \in \mathbb{R} \), the conditional probability of a future buy market order (positive trade sign) that the next trade is triggered by a buy market order given \( V_q(1) \geq x \) is defined as,

\[
\mathbb{P}(I_{q+1}^b = 1 | W_{q-1}(i) \geq x) \quad \text{and} \quad \mathbb{P}(I_{q+1}^s = 1 | W_{q-1}(i) \geq x)
\]

where we denote the trade sign at time \( t_q \) by \( I_{q+1}^b \).

Similarly, for all \( x \in \mathbb{R} \), the conditional probability of a future sell market order (negative trade sign) that the next trade is triggered by a sell market order given \( V_q(i) \geq x \) is defined as,

\[
\mathbb{P}(I_{q+1}^s = 1 | W_{q-1}(i) \geq x)
\]

The relationship between \( \mathbb{P}(I_{q+1}^b = 1 | W_{q-1}(i) \geq x) \) and \( x \) for \( i \in \{1, \ldots, L\} \) is shown in Figures 3 and 4. We observe that the conditional probability of the next trade sign is highly correlated with the bid-ask volume ratio corresponding to depth 1. Nevertheless, the dependence between the conditional probability of the next trade sign and the bid-ask volume ratio corresponding to depth larger than 1 is much more noisy. Figure 5 shows the relationship between \( \mathbb{P}(I_{q+1}^b = 1 | W_{q-1}(i) \geq x) \) and \( x \) for all stocks of CAC40. It is worth remarking that the trade sign’s conditional probability reaches 0.80 in average when the liquidity on the best limit prices is quite unbalanced.
Table 2. Summary of limit order events, market order events and inter-trade price jump events, CAC40 stocks, April, 2011.

<table>
<thead>
<tr>
<th>Stock</th>
<th>#LO AM</th>
<th>#MO AM</th>
<th>#BidJump AM</th>
<th>#AskJump AM</th>
<th>#BidTT AM</th>
<th>#AskTT AM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCP.PA</td>
<td>40505</td>
<td>44788</td>
<td>906</td>
<td>1067</td>
<td>120</td>
<td>125</td>
</tr>
<tr>
<td>AIRP.PA</td>
<td>65775</td>
<td>83199</td>
<td>1358</td>
<td>1715</td>
<td>190</td>
<td>257</td>
</tr>
<tr>
<td>ALSO.PA</td>
<td>92410</td>
<td>102069</td>
<td>1319</td>
<td>1590</td>
<td>165</td>
<td>199</td>
</tr>
<tr>
<td>ALUA.PA</td>
<td>141048</td>
<td>110173</td>
<td>1900</td>
<td>2379</td>
<td>202</td>
<td>233</td>
</tr>
<tr>
<td>AXAF.PA</td>
<td>156049</td>
<td>131722</td>
<td>1694</td>
<td>1951</td>
<td>155</td>
<td>167</td>
</tr>
<tr>
<td>BNPP.PA</td>
<td>432091</td>
<td>236054</td>
<td>3164</td>
<td>3160</td>
<td>377</td>
<td>368</td>
</tr>
<tr>
<td>BOUY.PA</td>
<td>28864</td>
<td>36112</td>
<td>973</td>
<td>1227</td>
<td>103</td>
<td>140</td>
</tr>
<tr>
<td>CAGR.PA</td>
<td>133449</td>
<td>86078</td>
<td>1795</td>
<td>1645</td>
<td>177</td>
<td>129</td>
</tr>
<tr>
<td>CAPP.PA</td>
<td>27679</td>
<td>30846</td>
<td>993</td>
<td>1262</td>
<td>125</td>
<td>164</td>
</tr>
<tr>
<td>CARR.PA</td>
<td>111559</td>
<td>104513</td>
<td>1536</td>
<td>1799</td>
<td>186</td>
<td>244</td>
</tr>
<tr>
<td>CNTA.PA</td>
<td>25216</td>
<td>29859</td>
<td>1144</td>
<td>1228</td>
<td>142</td>
<td>147</td>
</tr>
<tr>
<td>DANO.PA</td>
<td>124929</td>
<td>106412</td>
<td>1618</td>
<td>2160</td>
<td>180</td>
<td>182</td>
</tr>
<tr>
<td>EAD.PA</td>
<td>35720</td>
<td>35618</td>
<td>772</td>
<td>983</td>
<td>98</td>
<td>102</td>
</tr>
<tr>
<td>EDF.PA</td>
<td>151715</td>
<td>75212</td>
<td>1881</td>
<td>1750</td>
<td>190</td>
<td>158</td>
</tr>
<tr>
<td>ESSL.PA</td>
<td>21678</td>
<td>31743</td>
<td>528</td>
<td>751</td>
<td>55</td>
<td>66</td>
</tr>
<tr>
<td>FTE.PA</td>
<td>78140</td>
<td>83328</td>
<td>1370</td>
<td>1710</td>
<td>82</td>
<td>85</td>
</tr>
<tr>
<td>GSZ.PA</td>
<td>145293</td>
<td>105185</td>
<td>1781</td>
<td>2052</td>
<td>190</td>
<td>217</td>
</tr>
<tr>
<td>ISPA.AS</td>
<td>165149</td>
<td>170835</td>
<td>1877</td>
<td>2663</td>
<td>205</td>
<td>294</td>
</tr>
<tr>
<td>LAFP.PA</td>
<td>116640</td>
<td>87808</td>
<td>1149</td>
<td>1464</td>
<td>167</td>
<td>213</td>
</tr>
<tr>
<td>LVMH.PA</td>
<td>80949</td>
<td>84063</td>
<td>1006</td>
<td>1181</td>
<td>74</td>
<td>63</td>
</tr>
<tr>
<td>MICP.PA</td>
<td>101712</td>
<td>65052</td>
<td>1483</td>
<td>1564</td>
<td>200</td>
<td>182</td>
</tr>
<tr>
<td>OREP.PA</td>
<td>33981</td>
<td>40395</td>
<td>1182</td>
<td>1393</td>
<td>152</td>
<td>187</td>
</tr>
<tr>
<td>PERP.PA</td>
<td>65502</td>
<td>43922</td>
<td>971</td>
<td>1154</td>
<td>133</td>
<td>124</td>
</tr>
<tr>
<td>PEUP.PA</td>
<td>51684</td>
<td>57536</td>
<td>1166</td>
<td>1258</td>
<td>137</td>
<td>135</td>
</tr>
<tr>
<td>PRT.PA</td>
<td>29682</td>
<td>51908</td>
<td>359</td>
<td>627</td>
<td>36</td>
<td>34</td>
</tr>
<tr>
<td>PUBP.PA</td>
<td>71049</td>
<td>52461</td>
<td>1093</td>
<td>1350</td>
<td>136</td>
<td>144</td>
</tr>
<tr>
<td>RENA.PA</td>
<td>136579</td>
<td>96872</td>
<td>1766</td>
<td>1843</td>
<td>242</td>
<td>213</td>
</tr>
<tr>
<td>SASY.PA</td>
<td>111349</td>
<td>94709</td>
<td>1709</td>
<td>2221</td>
<td>153</td>
<td>170</td>
</tr>
<tr>
<td>SCHN.PA</td>
<td>100690</td>
<td>82453</td>
<td>1297</td>
<td>1397</td>
<td>97</td>
<td>78</td>
</tr>
<tr>
<td>SEV.PA</td>
<td>33050</td>
<td>29122</td>
<td>619</td>
<td>659</td>
<td>72</td>
<td>57</td>
</tr>
<tr>
<td>SGES.PA</td>
<td>94111</td>
<td>65542</td>
<td>1454</td>
<td>1624</td>
<td>178</td>
<td>198</td>
</tr>
<tr>
<td>SGOB.PA</td>
<td>158051</td>
<td>148326</td>
<td>1931</td>
<td>2336</td>
<td>240</td>
<td>308</td>
</tr>
<tr>
<td>SOGN.PA</td>
<td>285430</td>
<td>172865</td>
<td>3455</td>
<td>3317</td>
<td>496</td>
<td>475</td>
</tr>
<tr>
<td>STM.PA</td>
<td>96566</td>
<td>94170</td>
<td>1148</td>
<td>1367</td>
<td>172</td>
<td>213</td>
</tr>
<tr>
<td>TCEF.PA</td>
<td>77319</td>
<td>70533</td>
<td>1276</td>
<td>1405</td>
<td>163</td>
<td>172</td>
</tr>
<tr>
<td>TOTF.PA</td>
<td>287449</td>
<td>224132</td>
<td>2388</td>
<td>3228</td>
<td>308</td>
<td>396</td>
</tr>
<tr>
<td>UNBP.PA</td>
<td>41956</td>
<td>40961</td>
<td>612</td>
<td>755</td>
<td>63</td>
<td>68</td>
</tr>
<tr>
<td>VIE.PA</td>
<td>49086</td>
<td>52755</td>
<td>936</td>
<td>1074</td>
<td>106</td>
<td>105</td>
</tr>
<tr>
<td>VIV.PA</td>
<td>70356</td>
<td>73725</td>
<td>1478</td>
<td>1822</td>
<td>130</td>
<td>151</td>
</tr>
<tr>
<td>VLLP.PA</td>
<td>45336</td>
<td>37801</td>
<td>1198</td>
<td>1471</td>
<td>167</td>
<td>208</td>
</tr>
</tbody>
</table>
4. Logistic Regression Analysis and Variable Selection by LASSO

Logistic model is largely applied in many domains including quantitative finance. [21] uses logistic regression to classify trades and finds systematic temporal patterns in interday and intraday probabilities of trade sign on the US and Canadian exchanges. [22] analyzes the intraday probability of trade sign on the Australian Stock Exchange. Their results are encouraging and they have shown the robustness of logistic regression in prediction tasks in quantitative finance. Meanwhile, they have not studied the problem of variable selection in logistic regression. The objective of our work is not to show that logistic regression can beat all other machine learning algorithms, but to show that a well recognized model like logistic regression confirms the predictability of the limit order book and find out which variables are the most informative.

4.1. Logistic Regression Analysis

The result shown in the previous section reveals that bid-ask liquidity balance provides important information on the incoming market order. In this section, we introduce the standard logistic regression to predict the inter-trade price jump occurrence and use LASSO select regularization to evidence the importance of each vari-
We denote the number of market order events by \( N \) and for each \( i \in \{1, \cdots, N\} \),
\[
X_i = [1, V_{i1}^{m}, R_{i1}, R_{i1-1}, \cdots, R_{i1-m+1}^{m}, R_{i1}^{2}, R_{i1-1}^{2}, \cdots, R_{i1-m+1}^{2}]
\]

\( (X \in \mathbb{R}^{P+1}) \), \( P = m(2L-1) + 6n \) the explanatory variables summarizing the available order book information when the \( i^{th} \) event is a market order event, \( y_i \) is a binary variable indicating whether the event is an bid/ask inter-trade price jump, \( y_i \) is defined as follows,

Bid side inter-trade price jump indicator:
\[
y_i = \begin{cases} 
1, & \text{if } P_{i+1}^{m} < P_{i+1}^{b_1} \\
0, & \text{otherwise} 
\end{cases}
\]

or

Ask side inter-trade price jump indicator:
\[
y_i = \begin{cases} 
1, & \text{if } P_{i+1}^{m} > P_{i+1}^{b_1} \\
0, & \text{otherwise} 
\end{cases}
\]

In the logistic model, the probability of the bid/ask inter-trade price jump occurrence is assumed to be given by:
\[
\log \frac{P_{\beta}(Y = 1|X)}{1 - P_{\beta}(Y = 1|X)} = \beta^T X,
\]

where \( \beta = [\beta_0, \beta_1, \cdots, \beta_p]^T \).
Observing that for $i = 1$
\[ W_{q_{i-1}}(i) = V_{q_{i-1}}^{b,1} - V_{q_{i-1}}^{a,1}, \]
we see that the linearity of the conditional probability $P_b(V = 1 | X)$ on variables $V_{q_{i-1}}^{b,1}$ and $V_{q_{i-1}}^{a,1}$ in formula (1) allows us to capture the contribution of $W_{q_{i}}(i)$ in the prediction.

The parameters $\beta$ are unknown and should be estimated from the data. We use the maximum likelihood to estimate the parameters. It is well known that the
log-likelihood function given by
\[ L(\beta) = \sum_{i=1}^{N} \left\{ \log \left(1 + e^{\beta^T x_i}\right) - y_i \beta^T x_i \right\}. \]

The likelihood function is convex and therefore can be optimized using a standard optimization method.

### 4.2. Variable Selection by LASSO

Since the number of explanatory variables \( p \) being quite large, it is of interest to perform a variable selection procedure to select the most important variables. A classical variable selection procedure when the number of regressors is large is the LASSO procedure. Instead of using a BIC penalization, the LASSO procedure adds to the likelihood the norm of the logistic coefficient, which is known to induce a sparse solution. This penalization induces an automatic variable selection effect, see [23] and [24]. The LASSO penalty is an effective device for continuous model selection, especially in problems where the number of predictors far exceeds the number of observations, see [23,25-28].

The LASSO estimate for logistic regression is defined by
\[
\hat{\beta}^{lasso}(\lambda) = \arg \min_{\beta} \sum_{i=1}^{N} \left\{ -\log \left(1 + e^{\beta^T x_i}\right) + y_i \beta^T x_i \right\} + \lambda \sum_{j=1}^{p} |\beta_j|.
\]

The constraint on \( \sum_{j=1}^{p} |\beta_j| \) makes the solutions nonlinear in the \( y_i \) and there is no closed form expression as in ridge regression. Because of the nature of constraint, making \( \lambda \) sufficiently large will cause some of the coefficients to be exactly zero. [29] gives the uniform consistency and a functional central limit theorem for the LASSO regularization path for the general linear model.

For the implementation, we use the package glmnet in R which implements the fast algorithm via coordinate descent. This algorithm select the smallest value \( \max \) for which the entire parameter vector \( \beta = 0 \) then constructs a sequences of \( K \) values of \( \lambda \) decreasing from \( \lambda_{\max} \) to \( \epsilon \lambda_{\max} \) on the log scale where \( K \) and \( \epsilon \) are some constants, see more details in [30].

### 4.3. Results

We apply the LASSO logistic regression on three data-sets: morning dataset (between 09h05 and 13h15), afternoon dataset (between 13h15 and 17h25) and allday dataset (between 09h05 and 17h25). For each dataset, one third is used to estimate the parameters and two thirds is applied to evaluate the prediction quality of the estimator. The size of datasets is given in Table 2.

In this work, we choose \( L=5, m=5 \) and \( n=5 \), then the dimension of the profile vector of a limit order book becomes \( p=1+m(2L-1)+6n=76 \). The choice of \( L, m, n \) can be arbitrary, nevertheless, following the result shown in [31] on the long memory of order flow, it makes sense to choose \( n=5 \) which represents an order flow correlation larger than 6%.

The parameter \( \lambda \) in LASSO is estimated by cross-validation, then we calculate AUC value (area under ROC curve) to measure the prediction quality. A ROC (receiver operating characteristic) curve is a graphical plot of the true positive rate vs. false positive rate. The area under the ROC curve is a good measure to measuring the model prediction quality. The AUC value is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

We show the out-of-sample AUC value in Figure 6. The stocks are sorted in alphabetic order. We see that for each prediction task, the median AUC values lie around 0.70 and they are consistently high over all datasets and all stocks of CAC40. Moreover, we remark that the mean of all mixed AUC values obtained from three datasets are not significantly different in Table 3. It confirms that the predictability of order book information is not time-variant.

In order to discover the contribution of each variable to the prediction, we add an analysis on the five firstly selected variables for each prediction task of all stocks of CAC40 (with allday dataset).

Figures 7 and 8 show how many times a variable is selected as the first (second, third, fourth, fifth) selected variable by LASSO. We denote the \( i \) events lagged log volume on the \( j \)th bid (ask) limit price by \( VB_{i,j} \) (\( VA_{i,j} \)). Similarly, \( i \) events lagged log market order volume is denoted by \( VMO_{i,j} \) and \( i \) events lagged binary variables are denoted by \( BMO_{i,j}, AMO_{i,j}, ATT_{i,j} \) etc. For the sake of simplicity, for each selection order \( i \in \{1, \ldots, 5\} \), we show the frequency distribution of the five most selected variables among 746 backtests in each figure.

We observe that \( VB_{1,0}, BMO_{b} \) and \( VMO_{b} \) are the most selected variables for predicting the future bside \textit{inter-trade price jump} and that \( VA_{1,0}, AMO_{b} \) and

<table>
<thead>
<tr>
<th>Period</th>
<th>Event Type</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morning</td>
<td>BidJump</td>
<td>0.7184</td>
<td>0.0579</td>
</tr>
<tr>
<td></td>
<td>AskJump</td>
<td>0.7165</td>
<td>0.0596</td>
</tr>
<tr>
<td>Afternoon</td>
<td>BidJump</td>
<td>0.7217</td>
<td>0.0573</td>
</tr>
<tr>
<td></td>
<td>AskJump</td>
<td>0.7170</td>
<td>0.0555</td>
</tr>
<tr>
<td>All day</td>
<td>BidJump</td>
<td>0.7223</td>
<td>0.0482</td>
</tr>
<tr>
<td></td>
<td>AskJump</td>
<td>0.7323</td>
<td>0.0496</td>
</tr>
</tbody>
</table>

Table 3. AUC values obtained from all datasets.

Copyright © 2013 SciRes.
Figure 6. AUC value, price jump prediction, CAC40, April, 2011.

$VMO_0$ are the most selected variables for predicting the future askside inter-trade price jump. In contrast, trade-through is less informative and contributes few to the price jump prediction. It implies that the market order is sensitive to the liquidity on the best limit price. As soon as the liquidity on the best limit price becomes significantly low, the next market order may touch it immediately. The information provided by $BMO_0(AMO_0)$ and
Figure 7. Variable selection for BidJump prediction, CAC40, April, 2011. From left to right, from top to bottom, each figure shows how many times a variable is selected as the $k^{th}$ selected variable by LASSO, $k = \{1, \cdots, 5\}$. 
Figure 8. Variable selection for AskJump prediction, CAC40, April, 2011. From left to right, from top to bottom, each figure shows how many times a variable is selected as the $k^{th}$ selected variable by LASSO, $k = \{1, \cdots, 5\}$. 

Copyright © 2013 SciRes.
\( VMO \) recalls the phenomena of long memory of order flow, see [31]. When a trader tries to buy or sell a large quantity of assets, he may split it into small pieces and execute them by market order successively. Consequently, precedent market order direction contributes to predict the next market order event.

5. Conclusion

In this paper, we provide an empirical result on the relationship between bid-ask limit order liquidity balance and trade sign and an analysis on the prediction of the \textit{inter-trade price jump} occurrence by logistic regression. We show that limit order liquidity balance on best bid/best ask is informative to predict the next market order’s direction. We then use limit order volumes, limit order price gaps and market order size to construct the features of the limit order book for the prediction of \textit{inter-trade price jump} occurrence. LASSO logistic regression is introduced to help us identify the most informative limit order book features for the prediction. Numerical analysis is done on two separated datasets: morning dataset and afternoon dataset. LASSO logistic regression gets very good prediction results in terms of AUC value. The AUC value is consistently high on both datasets and all stocks whatever the liquidity is. This good prediction quality implies that the limit order book profile is quite informative for predicting the incoming market order event. The variable selection by LASSO logistic regression shows that several variables are quite informative for \textit{inter-trade price jump} prediction. The trade sign and market order size and the liquidity on the best limit prices are the most informative variables. Nevertheless, the aggressiveness of market order, measured by \textit{trade-through}, has less important impact than we had expected. These results confirm that the limit order book is quite sensitive to the liquidity on the best limit prices and there is a long memory of order flow like what is shown by other authors. This paper is merely a first attempt to discover the information hidden in the limit order book and further studies will be needed to understand better the full dynamics of the limit order book.

6. Acknowledgements

We are grateful to François Roueff, Michel Crouhy, Adil Reghai, Nicolas Huth, Julien Puvilland, Marc Souaille and Mohamed Lakhdar, Fabrizio Pomponio for helpful discussions. This research is supported by NATIXIS quantitative research department.

REFERENCES


