Forecasting Prices in the Presence of Hidden Liquidity

Marco Avellaneda[*], Josh Reed[†]& Sasha Stoikov [‡]

December 14, 2010

Abstract

Bid and ask sizes at the top of the order book provide information on short-term price moves. Drawing from classical descriptions of the order book in terms of queues and order-arrival rates (Smith et al (2003)), we consider a diffusion model for the evolution of the best bid/ask queues. We compute the probability that the next price move is upward, conditional on the best bid/ask sizes, the *hidden liquidity* of the market and the correlation between changes in the bid/ask sizes. The model can be useful, among other things, to rank trading venues in terms of the "information content" of their quotes and to estimate the hidden liquidity in a market based on high-frequency data. We illustrate the approach with an empirical study of a few liquid stocks using quotes from various exchanges.

[*]Courant Institute, New York University and Finance Concepts LLC
[†]Stern School of Business, New York University
[‡]Cornell Financial Engineering Manhattan, corresponding author, email: sfs33@cornell.edu

Contents

1 Introduction **3**

2 Modeling Level I quotes **4**

2.1 Hidden liquidity 5

2.2 The discrete Poisson model 6

3 Diffusion approximation **8**

3.1 Probability of an upward move 8

3.2 Boundary conditions 9

3.3 Solution 10

4 Data analysis **11**

4.1 Data description 11

4.2 Estimation procedure 13

4.3 Results 13

5 Conclusions **14**

1 Introduction

The term "order book" (OB) is generally used to describe the bid and ask prices and sizes in continuous-auction exchanges, such as NYSE-ARCA, BATS or NASDAQ. A distinction is often made between Level I quotes, i.e. the best bid/ask prices and sizes, and Level II quotes, which consist of all prices and sizes available in the order book. In either case, the OB provides information on market depth, allowing traders to estimate the impact of their trades. A question of obvious interest, given the high degree of transparency of OB data, is whether the order book provides any information on short-term price moves.

Order book dynamics have been studied by many authors in the econophysics literature (see for instance Smith et al [5]), who generally focus on estimating and simulating the arrivals of limit, market and cancel orders, in order to model the highly complex dynamics of the market. Order books models are also of central interest in the market microstructure literature; they are used to study, for instance, the impact of different types of agents and trading rules on the market in aggregate. In a 1995 paper, Hasbrouck [4] used econometric methods to analyze the initial stages of the US stock market "fragmenting" into regional exchanges, and at the time found "that the preponderance of the price discovery takes place in the New York Stock Exchange (NYSE) (a median 92.7 percent information share)". In recent years, with the emergence of competing electronic trading venues (ECN's and Dark Pools) for the same asset and algorithmic trading, questions related to the quality, speed and transparency of information on various exchanges have become ever more relevant for regulators and practitioners alike.

We propose here a modeling approach that allows one to measure and compare the information content of order books and to generate short term forecasts for price moves. Our approach is inspired by Markov-type models for the order book, first proposed by Smith, Farmer, Gillemot and Krishnamurthy (SFGK) and more recently studied in Cont, Stoikov and Talreja (CST). These models are high-dimensional Markov processes with a state-space consisting of vectors (bid price, bid size) and (ask price, ask size), and of Poisson-arrival rates for market, limit and cancellation orders. They are often referred picturesquely as "zero-intelligence models", because orders arrive randomly, rather than being submitted by rational traders with a budget, utility objective, memory, etc. Needless to say, a full description of order books as a Markov process gives rise to a highly complex system and the solution of the model (in any sense) is often problematic and of questionable value in practice.[1] For this

[1] In fact, one might argue that "zero-intelligence" may not characterize the fashion in which continuous auctions are conducted. Traders, often aided by sophisticated computer algorithms, position their orders to take advantage of situations observed in the order book as well as to fill large block orders on behalf of customers. Rules such as first-in-first-out, and the possibility of capturing rebates for posting limit orders (adding liquidity) result in

reason, we choose to simplify such models by considering instead a reduced, diffusion-type dynamics for bid and ask sizes and focusing on the top of the book instead of on the entire OB. Of course, such methods could be generalized, in principle, to incorporate second-best bids and asks and even more complex descriptions at the expense of simplicity.

We ask a simple, fundamental, question about the OB. Do Level II quotes, or even Level I quotes, give information about price direction? In other words, can we forecast the direction of price movements based on bid and ask sizes? The degree to which this can be done given the OB could be called the *information content*. For example, if the sizes of queues do not provide information, then, if ΔP denotes the next price move, then

$$Prob.\{\Delta P > 0 \,|\, OB\} = Prob.\{\Delta P < 0 \,|\, OB\} = 0.5,$$

where the probabilities are conditional on observing the OB. If the OB is "informative", we expect that

$$Prob.\{\Delta P > 0 \,|\, OB\} = p(OB),$$

i.e. that the order book provides a forecast of next price move in the form of a conditional probability. The information contained in the OB, if any, should tell us to what extent $p(OB)$ differs from 0.5 based on the observation of limit orders in the book and on the statistics of the queue sizes as they vary in time.

Thus, our goal is to create diffusion models, inspired by SFGK or CST, that can be used to forecast the direction of stock-price moves based on measurable statistical quantities. In contrast to CST, we explicitly model bid and ask quotes with some *hidden liquidity*, i.e. sizes that are not shown in the OB, but which may influence the probability of an upward move in the price. The idea of estimating hidden liquidity is not new to the trading literature (see Burghadt et al (2006), who estimate it's magnitude by comparing sweep-to-fill prices to VWAP prices).

2 Modeling Level I quotes

In the Markov model of CST, the OB has two distinguished queues representing the sizes at best bid and the best ask levels, which are separated by the minimum tick size. Market, limit and cancellation orders arrive at both queues according to Poisson processes. One of the following two events must then happen first:

1. The ask queue is depleted and the best ask price goes up by one tick and the price

markets in which there is a significant amount of strategizing *conditionally* on the state of the OB.

"moves up".

2. The bid queue is depleted and the best bid price goes down by one tick and the price "moves down".

The dynamics leading to a price change may thus be viewed as a "race to the bottom": the queue that hits zero first causes the price to move in that direction.

As it turns out, the predictions of such models are not consistent with market observations. If they were, this would imply that if the best ask size becomes much smaller than the best bid size, the probability that the next price move is upward should approach 100%. However, empirical analysis (see Section 4) shows that this probability does not increase to unity as the ask size goes to zero.

2.1 Hidden liquidity

We hypothesize that this happens for two reasons: first, markets are fragmented; liquidity is typically posted on various exchanges. In the U.S. stock markets, for example, Reg NMS requires that all market orders be routed to the venue with the best price. Moreover, limit orders that could be immediately executed at their limit price on another market need to be rerouted to those venues. Thus, one needs to consider the possibility that once the best ask on an exchange is depleted, the price will not necessarily go up, since an ask order at that price may still be available on another market and a new bid cannot arrive until that price is cleared on all markets. The second reason is the existence of trading algorithms that split large orders into smaller ones that replenish the best quotes as soon as they are depleted ("iceberg orders"). In the sequel, we will model this by assuming that there is a fixed hidden liquidity (size) behind the best bid and ask quotes. This hidden liquidity may correspond to iceberg orders or orders present on another exchange. Quotes on other exchanges, although not technically hidden from anybody, may be subject to latencies and therefore only available to some traders with the fastest data feeds. The main adjustable parameter in our model, the *hidden liquidity*, will be an important indicator of the information content of the OB.

In summary, the main new idea in our interpretation of the OB models is that we do not immediately assume that a *true* change in price occurs when either of the queues first hits zero. Rather, we take the following view. We postulate that a price transition takes place whenever the first of two events happens:

1. The size for the best ask price goes to zero and the hidden liquidity at that price is depleted. Intuitively, we assume that the price has only moved if there is "support" at a new bid level. This can only happen if all ask orders on all exchanges are cleared at

that price and iceberg orders are exhausted.

2. Alternatively, the size for the best bid price goes to zero and the hidden liquidity at that price is depleted.

2.2 The discrete Poisson model

Adopting the language of queuing theory, we refer to the number of shares offered at the lowest ask price as the *ask queue*. Similarly, the number of shares bid at the highest bid price is called the *bid queue*. Following CST (2010), we view these queues as following a continuous time Markov chain (CTMC) where time is continuous and share quantities are discrete, consistently with a minimum order size. [2] We adopt the following notation

$$\begin{aligned} h &= \text{minimum order size} \\ \lambda_a &= \text{arrival rate of limit orders at the ask} \\ \lambda_b &= \text{arrival rate of limit orders at the bid} \\ \mu_a &= \text{arrival rate of (buy) market orders at the ask or cancellations at the ask} \\ \mu_b &= \text{arrival rate of (sell) market orders at the bid or cancellations at the bid} \end{aligned} \tag{2.1}$$

The model for the top of the order book is a continuous-time discrete space process in which the evolution of the queues follows a Markov process in which a state is (X, Y), where X = bid queue size and Y=ask queue size. Each state can transition into four neighboring states by increasing or decreasing the queue sizes. The transition rates are given by

$$\begin{aligned} \lambda_{0,1} &= \lambda_a \\ \lambda_{0,-1} &= \mu_a \\ \lambda_{-1,0} &= \mu_b \\ \lambda_{1,0} &= \lambda_b. \end{aligned} \tag{2.2}$$

Empirically, we know that the queue sizes are negatively correlated. Therefore, it is convenient to incorporate correlation between the bid and the ask queues in our model as well.

[2] Note that in CST, market orders and cancellations are modeled separately resulting in queues sizes that revert around an equilibrium level. Such microstructural distinctions are not essential at the macroscopic level, i.e. when we fit the model to transactions data.

To do this, we introduce additional diagonal transitions.[3]

We set

$$\lambda_{-1,+1} = \lambda_{+1,-1} = \eta > 0. \tag{2.3}$$

With these conventions, we have

$$\begin{aligned} E\left[X_{t+\Delta t} - X_t | X_t, Y_t\right] &= h\left(\lambda_{1,0} - \lambda_{-1,0}\right)\Delta t + o(\Delta t) \\ &= h\left(\lambda_b - \mu_b\right)\Delta t + o(\Delta t) \end{aligned}$$

and, similarly,

$$\begin{aligned} E\left[Y_{t+\Delta t} - Y_t | X_t, Y_t\right] &= h\left(\lambda_a - \mu_a\right)\Delta t + o(\Delta t) \\ E\left[(X_{t+\Delta t} - X_t)^2 | X_t, Y_t\right] &= h^2\left(\lambda_b + \mu_b + 2\eta\right)\Delta t + o(\Delta t) \\ E\left[(Y_{t+\Delta t} - Y_t)^2 | X_t, Y_t\right] &= h^2\left(\lambda_a + \mu_a + 2\eta\right)\Delta t + o(\Delta t) \\ E\left[(X_{t+\Delta t} - X_t)(Y_{t+\Delta t} - Y_t) | X_t, Y_t\right] &= h^2\left(2\eta\right)\Delta t + o(\Delta t). \end{aligned}$$

It follows that the drifts and the variances of the queue sizes are given by

$$\begin{aligned} m_X &= h\left(\lambda_b - \mu_b\right) \\ m_Y &= h\left(\lambda_a - \mu_a\right) \\ \sigma_X^2 &= h^2\left(\lambda_b + \mu_b + 2\eta\right) \\ \sigma_Y^2 &= h^2\left(\lambda_a + \mu_a + 2\eta\right). \end{aligned} \tag{2.4}$$

If we assume, for simplicity that there is symmetry between bid and offer sizes, i.e. that $\lambda_a = \lambda_b = \lambda$, $\mu_a = \mu_b = \mu$, then

$$\begin{aligned} m_X = m_Y &= h\left(\lambda - \mu\right) \\ \sigma_X^2 = \sigma_Y^2 &= h^2\left(\lambda + \mu + 2\eta\right) \end{aligned} \tag{2.5}$$

[3] An alternative approach would be to keep the 4-point template and make the transition rates state-dependent. We chose a simple "diagonal transition" model instead. The latter can be viewed as describing transitions observed after a two time-units instead of one, for example. Such microstructural distinctions are not essential at the macroscopic level, i.e. when we fit the model to transaction data.

In particular, the correlation between the bid and the ask queues is

$$\rho = \frac{-2\eta}{\lambda + \mu + 2\eta}. \tag{2.6}$$

A further simplification can be made if we assume that the drifts vanish, which is accomplished by setting

$$\lambda = \mu.$$

This gives

$$\begin{aligned} \sigma_X^2 = \sigma_Y^2 &= 2h^2 (\lambda + \eta) \\ \rho &= \frac{-\eta}{\lambda + \eta} \end{aligned} \tag{2.7}$$

3 Diffusion approximation

3.1 Probability of an upward move

Let $< X >$ and $< Y >$ denote, respectively, the average (or median) size of the queues X_t, Y_t. We assume that the average queue sizes are much larger than the typical quantity of shares traded and that the frequency of orders per unit time is high, *i.e.* $< X >=< Y >\gg h$ and $\lambda, \eta \gg 1$.

We define the coarse-grained variables

$$x = X/ < X >, \quad y = Y/ < Y >,$$

which measure the queue sizes "macroscopically", and set

$$\sigma^2 = \frac{2h^2 (\lambda + \eta)}{< X >^2}, \tag{3.1}$$

Under these assumptions, by the functional central limit theorem for Poisson processes [1], the process (x_t, y_t) can be approximated by the diffusion

$$\begin{aligned} dx_t &= \sigma dW_t^{(1)} \\ dy_t &= \sigma dW_t^{(2)} \\ E\left(dW^{(1)} dW^{(2)}\right) &= \rho dt, \end{aligned} \tag{3.2}$$

where σ is defined in (3.1), ρ is defined in (2.7), and $W^{(1)}, W^{(2)}$ are standard Brownian

motions. [4]

We consider the function $u(x, y)$ representing the probability that the next price move is up, given that we observe the (standardized) bid/ask sizes (x, y). From diffusion theory, this function satisfies the differential equation

$$\sigma^2 \left(u_{xx} + 2\rho u_{xy} + u_{yy}\right) = 0, \quad x > 0,\ y > 0, \tag{3.3}$$

or, simply,

$$u_{xx} + 2\rho u_{xy} + u_{yy} = 0 \text{ for } \quad x > 0,\ y > 0. \tag{3.4}$$

3.2 Boundary conditions

The boundary conditions satisfied by $u(x, y)$ will depend on what assumptions are made about how the price changes once one of the queues is depleted. This is a key point to formulate a realistic model for forecasting price moves, based on a coarse-grained description of the OB. If we assume, naively, that the order-book represents fully the liquidity in the market at a particular price level, then the mid-price will move up once the ask queue is depleted *i.e.* when $y_t = 0$ for the first time, since no more sellers are present at that level. In this case, the probability that the price will increase corresponds to the probability that the diffusion (3.2) exits the quadrant $\{(x, y); x > 0, y > 0\}$ through the x-axis. The corresponding boundary conditions for $u(x, y)$ are therefore

$$\begin{aligned} u(0, y) &= 0, \quad \text{for} \quad y > 0, \\ u(x, 0) &= 1, \quad \text{for} \quad x > 0. \end{aligned} \tag{3.5}$$

However, we know empirically that an upward price move might not take place when the ask queue is depleted, due to additional liquidity at that level, which we call *hidden liquidity*. This hidden liquidity can be attributed to either iceberg orders or by virtue of a Reg-NMS-type mechanism in which there are other markets that still post liquidity on the ask-side at the same level and which must be honored before the mid price can move up.

A simple way to model this is to assume that there is an additional amount of liquidity, denoted by H, representing the fraction of average book size ($< X >$ or $< Y >$) which is

[4] Notice that these processes are pure diffusions without drift. Drift considerations are important to maintain queue sizes finite and, for instance, to model the fact that bid and ask sizes are typically mean-reverting. However, drifts are less important to describe order books in the vicinity of a price transition, when one or both queues are small. See Cont *et al.* (CST) for discussions about drifts and the constraints that they impose on the Poisson model for the OB.

"hidden" or absent from the book. A true price transition takes place if the hidden liquidity is exhausted. In other words, we observe queues of size x or y but the "true" size of the queues are $x + H$ and $y + H$. Thus, if we denote by $p(x, y; H)$ the probability of an upward price move conditional on the observed queue sizes (x, y) and the hidden liquidity parameter H, we have

$$p(x, y; H) = u(x + H, y + H), \tag{3.6}$$

where $u(x, y)$ satisfies the diffusion equation on the first quadrant of the (x, y) plane with boundary conditions (3.3).

3.3 Solution

Theorem 3.1. *The function*

$$u(x, y) = \frac{1}{2}\left(1 - \frac{Arctan\left(\sqrt{\frac{1+\rho}{1-\rho}}\frac{y-x}{y+x}\right)}{Arctan\left(\sqrt{\frac{1+\rho}{1-\rho}}\right)}\right). \tag{3.7}$$

satisfies equation (3.4). *Furthermore, we have* $u(x, 0) = 1$ *and* $u(0, y) = 0$.

The proof of the above result is in the Appendix. Together with (3.6), equation (3.7) provides a closed form formula for the probability that the next price move is upward, given the bid and ask sizes (x,y) and the parameters ρ and H.

Remarks. 1. If we set $\rho = 0$ the above expression simplifies to

$$u(x, y) = \frac{2}{\pi} Arctan\left(\frac{x}{y}\right). \tag{3.8}$$

2. As ρ approaches -1, the numerator and denominator in (3.7) both tend to zero. The limit as $\rho \to -1$ is

$$u(x, y) = \frac{x}{x + y}. \tag{3.9}$$

3. If we consider the sector $y < x$, i.e. the sector for which the ask queue is smaller than the bid queue and therefore that the price is then we notice that, in this region, $u(x, y)$ is an increasing function of ρ. In fact, setting $\xi = \frac{y-x}{y+x}$ and $\alpha = \sqrt{\frac{1+\rho}{1-\rho}}$,

$$\frac{\partial u}{\partial \rho} = -\frac{1 + \alpha^2}{1 + \alpha^2\xi^2}\frac{1}{(1 - \rho)^2}\frac{1}{2\xi}.$$

This is a positive quantity since ξ is negative in the sector. Therefore,*the assumption* $\rho = -1$ *will underestimate the probability of an up-tick if the "true correlation was higher than* -1.

4 Data analysis

In this section, we study the information content of the best quotes for the tickers QQQQ, XLF, JPM, and AAPL, over the first five trading days in 2010 (i.e. Jan 4-8). All four tickers are traded on various exchanges, and this allows us to compare the information content of these venues. In other words we will be computing the probability:

$$Prob.\{\Delta P > 0 \,|\, OB\} = p(OB),$$

discussed in the introduction, for ΔP defined to be the next midprice move, and for OB defined to be the pair of bid and ask sizes (x, y).

In our data analysis, we focus on the hidden liquidity for the perfectly negatively correlated queues model, i.e.

$$p(x, y; H) = \frac{x + H}{x + y + 2H} \tag{4.1}$$

which we estimate by minimizing square errors with respect to the empirical probabilities.

In practice, when performing our data analysis, we find it easier to bucket the data in deciles of queue sizes, rather than normalizing by the average queue size, as we did in Section 3. The implied hidden liquidity parameter we compute in the sequel can therefore be interpreted as a fraction of the maximum observed queue size.

4.1 Data description

The data comes from the WRDS database, more specifically the consolidated quotes of the NYSE-TAQ data set. Each row has a timestamp (between the hours of 10:00 and 16:00, rounded to the nearest second), a bid price, an ask price, a bid size, an ask size and an exchange flag, indicating if the quote was on NASDAQ (T), NYSE-ARCA (P) or BATS (Z), see Table 1 for a sample of the data. There are other regional exchanges, but for the purpose of this study, we focus on these venues as they have significantly more than one quote per second.

In table 2, we present some summary statistics for the tickers QQQQ, XLF, JPM and AAPL, across the three exchanges. The tickers QQQQ, XLF and JPM are ideal candidates, because their bid-ask spread is almost always one tick (or one cent) wide, much like our stochastic model. We also pick AAPL, whose spread most often trades at 3 cents (or three ticks wide), due to AAPL's relatively high stock price. Though our model does not strictly consider spreads greater than one, we use it to fit our model, conditional on the spread, i.e. $OB = (x, y, s)$ where s is the spread in cents.

symbol	date	time	bid	ask	bsize	asize	exchange
QQQQ	2010-01-04	09:30:23	46.32	46.33	258	242	T
QQQQ	2010-01-04	09:30:23	46.32	46.33	260	242	T
QQQQ	2010-01-04	09:30:23	46.32	46.33	264	242	T
QQQQ	2010-01-04	09:30:24	46.32	46.33	210	271	P
QQQQ	2010-01-04	09:30:24	46.32	46.33	210	271	P
QQQQ	2010-01-04	09:30:24	46.32	46.33	161	271	P

Table 1: A sample of the raw data

Ticker	Exchange	num quotes	quotes/sec	avg(spread)	avg(bsize+asize)	avg(price)
XLF	NASDAQ	0.7M	7	0.010	8797	15.02
XLF	NYSE	0.4M	4	0.010	10463	15.01
XLF	BATS	0.4M	4	0.011	7505	14.99
QQQQ	NASDAQ	2.7M	25	0.010	1455	46.30
QQQQ	NYSE	4.0M	36	0.011	1152	46.27
QQQQ	BATS	1.6M	15	0.011	1055	46.28
JPM	NASDAQ	1.2M	11	0.011	87	43.81
JPM	NYSE	0.7M	6	0.012	47	43.77
JPM	BATS	0.6M	5	0.014	39	43.82
AAPL	NASDAQ	1.3M	13	0.034	9.1	212.50
AAPL	NYSE	0.4M	4	0.046	5.7	212.66
AAPL	BATS	0.6M	6	0.054	4.5	212.43

Table 2: Summary statistics

4.2 Estimation procedure

1. We split the data set into three subsets, one for each exchange. Items 2-6 are repeated separately for each exchange and each ticker.
2. We remove zero and negative spreads.
3. We "bucket" the bid and ask sizes, by taking deciles of the bid and ask size and normalizing queue sizes so that (i, j) represents the ith decile of the bid size and the jth decile of the ask size respectively.
4. For each bucket (i, j), we compute the empirical probability that the price goes up u_{ij}. This is done by looking forward to the next mid price change and computing the empirical percentage of occurrences of (i, j) that ended up going up, before going down.
5. We count the number of occurrences of the (i, j) bucket, and denote this distribution d_{ij}.
6. We minimize least squares for the negatively correlated queues model, i.e.

$$\min_H \sum_{i,j} \left[\left(u_{ij} - \frac{i+H}{i+j+2H} \right)^2 d_{ij} \right] \tag{4.2}$$

   and obtain an implied hidden liquidity H for each exchange.

4.3 Results

We first illustrate the predictions of our model for the ticker XLF on the Nasdaq exchange (T). We report the empirical probabilities of an up move, given the bid and ask sizes in table 3, as well as the model probabilities, given by equation (4.1) with H estimated with the procedure described above. Notice that even for very large bid sizes and small ask sizes (say the 90th percentile of sizes at the bid and the 10th percentile of sizes at the ask) the empirical probability of the mid price moving upward is high (0.85) but not arbitrary close to one. The same is true of our model, which assumes there is a hidden liquidity H behind both quotes. We interpret H as a measure of the information content of the bid and ask sizes: the smaller H is, the more size matters. The larger the H, the closer all probabilities will be to 0.5, even for drastic size imbalances.

In table 4, we display the hidden liquidity H for the four tickers and three exchanges. These results indicate that size is most important for

- XLF on NASDAQ,
- QQQQ on NYSE-ARCA and for

decile	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
sizes	< 1250	< 1958	< 2753	< 3841	< 4835	< 5438	< 5820	< 6216	< 6742
0.1	0.50	0.38	0.25	0.25	0.32	0.26	0.23	0.23	0.15
0.2	0.61	0.50	0.47	0.41	0.36	0.40	0.38	0.27	0.20
0.3	0.75	0.53	0.50	0.43	0.39	0.37	0.43	0.39	0.28
0.4	0.74	0.58	0.57	0.50	0.42	0.42	0.47	0.46	0.37
0.5	0.68	0.64	0.61	0.58	0.50	0.51	0.48	0.49	0.41
0.6	0.74	0.60	0.63	0.58	0.49	0.50	0.50	0.50	0.49
0.7	0.78	0.62	0.57	0.53	0.52	0.50	0.50	0.60	0.53
0.8	0.77	0.73	0.61	0.54	0.51	0.50	0.40	0.50	0.42
0.9	0.85	0.79	0.72	0.63	0.60	0.51	0.47	0.57	0.50
decile	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
sizes	$= 1250$	$= 1958$	$= 2753$	$= 3841$	$= 4835$	$= 5438$	$= 5820$	$= 6216$	$= 6742$
0.1	0.50	0.42	0.36	0.31	0.28	0.25	0.23	0.21	0.19
0.2	0.58	0.50	0.44	0.39	0.35	0.32	0.29	0.27	0.25
0.3	0.64	0.56	0.50	0.45	0.41	0.37	0.35	0.32	0.30
0.4	0.69	0.61	0.55	0.50	0.46	0.42	0.39	0.37	0.34
0.5	0.72	0.65	0.59	0.54	0.50	0.46	0.43	0.41	0.38
0.6	0.75	0.68	0.63	0.58	0.54	0.50	0.47	0.44	0.42
0.7	0.77	0.71	0.65	0.61	0.57	0.53	0.50	0.47	0.45
0.8	0.79	0.73	0.68	0.63	0.59	0.56	0.53	0.50	0.47
0.9	0.81	0.75	0.70	0.66	0.62	0.58	0.55	0.53	0.50

Table 3: Empirical vs. Model probabilities for the probability of an upward move (XLF), on Nasdaq (T). Rows represent bid size percentiles (i), columns represent ask size percentiles (j). The model is given by $p(i,j) = \frac{i+H}{i+j+H}$ with $H = 0.15$

- JPM on BATS.

Finally we calculate H for AAPL, for different values of the bid-ask spread ($s = 1$, 2 and 3 cents). We find that sizes of

- AAPL are more informative on NASDAQ, and that they matter most when the spread is small.

Modeling stocks with larger spreads may require more sophisticated models of the order book, possibly including Level II information. Since a majority of US equities trade at average spreads of several cents, we consider this avenue worthy of future research.

5 Conclusions

Based on a diffusion model of the liquidity at the top of the order book, we proposed closed-form solutions for the probability of a price uptick conditional on Level-I quotes. The probability is a function of the bid size, the ask size and an adjustable parameter, H, the *hidden liquidity*. The advantage of this simple model is that it can be fitted to high-frequency

Ticker	NASDAQ	NYSE	BATS
XLF	0.15	0.17	0.17
QQQQ	0.21	0.04	0.18
JPM	0.17	0.17	0.11
AAPL $s = 1$	0.16	0.90	0.65
AAPL $s = 2$	0.31	0.60	0.64
AAPL $s = 3$	0.31	0.69	0.63

Table 4: Implied hidden liquidity across tickers and exchanges

data and produces an *implied* hidden liquidity parameter, obtained by fitting tick data (from WRDS) to the proposed formulas. The result is that we can classify different markets in terms of their hidden liquidity or, equivalently, how informative the Level I quotes of a stock are in terms of forecasting the next price move (up or down). If the hidden size is small (compared to the typical size shown in the market under consideration), we say that the best quotes are informative. Statistical analysis for different stocks shows the following results:

- for XLF (SPDR Financial ETF), NASDAQ has the least hidden liquidity;
- for QQQQ (Powershares Nasdaq-100 Tracker), NYSE-ARCA has the least hidden liquidity;
- for JPM (J.P. Morgan & Co.) BATS has the least hidden liquidity and
- for AAPL (Apple Inc.) NASDAQ has the least hidden liquidity.

We used only 5 days of data for these calculations and a study of the stability of our hidden liquidity parameter over longer periods remains to be done. Nevertheless, the approach presented here seems to provide a way of comparing trading venues, in terms of their information content and hidden liquidity, and hence on the possibility of forecasting price changes from their orderbook data.

Appendix

Solution of the PDE for general ρ

Proposition 1 Let $\Omega(X,Y)$ be a harmonic function. Let us set

$$v(\zeta,\eta)=\Omega(\frac{\zeta}{\sigma_1},\frac{\eta}{\sigma_2}).$$

Then,

$$\sigma_1^2 v_{\zeta\zeta}+\sigma_2^2 v_{\eta\eta}=0. \tag{5.1}$$

Proof: By the chain rule, $\sigma_1^2 v_{\zeta\zeta}=\sigma_1^2\frac{\Omega_{XX}}{\sigma_1^2}=\Omega_{XX}$. The same holds for the η-derivative. Add and use harmonicity of Ω.

Proposition 2 Let Ω be a harmonic function. Then

$$u(x,y)=\Omega(\frac{y+x}{\sqrt{2}\sqrt{1+\rho}},\frac{y-x}{\sqrt{2}\sqrt{1-\rho}}) \tag{5.2}$$

satisfies

$$u_{xx}+2\rho u_{xy}+u_{yy}=0. \tag{5.3}$$

Proof: Let $\sigma_1=\sqrt{1+\rho},\sigma_2=\sqrt{1-\rho}$ and set $\zeta=\frac{y+x}{\sqrt{2}},\eta=\frac{y-x}{\sqrt{2}}$. Clearly, by Proposition 1, $v(\zeta,\eta)\equiv\Omega(\frac{\zeta}{\sigma_1},\frac{\eta}{\sigma_2})$ satisfies

$$\sigma_1^2 v_{\zeta\zeta}+\sigma_2^2 v_{\eta\eta}=0.$$

Since $u(x,y)=v(\frac{y+x}{\sqrt{2}},\frac{y-x}{\sqrt{2}})$, we have, after differentiating twice the function u

$$\begin{aligned} u_{xx} &= \frac{1}{2}v_{\zeta,\zeta}+\frac{1}{2}v_{\eta,\eta}-v_{\zeta\eta} \\ u_{yy} &= \frac{1}{2}v_{\zeta,\zeta}+\frac{1}{2}v_{\eta,\eta}+v_{\zeta\eta} \\ u_{xy} &= \frac{1}{2}v_{\zeta,\zeta}-\frac{1}{2}v_{\eta,\eta}. \end{aligned} \tag{5.4}$$

Adding the first two terms and then adding the third one multiplied by 2ρ gives

$$
\begin{aligned}
u_{xx} + 2\rho u_{xy} + u_{yy} &= \frac{1}{2}v_{\zeta,\zeta} + \frac{1}{2}v_{\eta,\eta} - v_{\zeta\eta} + \\
&\quad 2\rho\left(\frac{1}{2}v_{\zeta,\zeta} - \frac{1}{2}v_{\eta,\eta}\right) + \\
&\quad \frac{1}{2}v_{\zeta,\zeta} + \frac{1}{2}v_{\eta,\eta} + v_{\zeta\eta} \\
&= v_{\zeta,\zeta} + \rho\left(v_{\zeta,\zeta} - v_{\eta,\eta}\right) + v_{\eta,\eta} \\
&= (1+\rho)v_{\zeta,\zeta} + (1-\rho)v_{\eta,\eta} \\
&= \sigma_1^2 v_{\zeta\zeta} + \sigma_2^2 v_{\eta\eta} \\
&= 0. \qquad (5.5)
\end{aligned}
$$

Theorem 3.1 The function

$$u(x,y) = \frac{1}{2}\left(1 - \frac{Arctan\left(\sqrt{\frac{1+\rho}{1-\rho}}\frac{y-x}{y+x}\right)}{Arctan\left(\sqrt{\frac{1+\rho}{1-\rho}}\right)}\right). \qquad (5.6)$$

satisfies equation (3). Furthermore, we have $u(x,0) = 1$ and $u(0,y) = 0$.

Proof: Use $\Omega(X,Y) = Arctan(Y/X)$ and apply Proposition 2, using

$$X = \frac{y+x}{\sqrt{2}\sqrt{1+\rho}}\ ;\quad Y = \frac{y-x}{\sqrt{2}\sqrt{1-\rho}}.$$

References

[1] P. Billingsley, Convergence of Probability Measures, John Wiley and Sons, 1999, New York.

[2] G. Burghadt, J. Hanweck, and L. Lei (2006) Measuring Market Impact and Liquidity, The Journal of Trading, Fall 2006, Vol. 1, No. 4, pp. 70-84

[3] R. Cont, S. Stoikov, R. Talreja (2010) *A Stochastic Model for Order Book Dynamics*, Operations Research, Vol. 58, No. 3, May-June 2010, pp. 549-563.

[4] J. Hasbrouck (1995), One security, many markets: determining the contributions to price discovery, Journal of Finance, Vol 1, No. 4 , pp 1175-1199

[5] E. Smith, J. D. Farmer, L. Gillemot, and S. Krishnamurthy, (2003), Statistical Theory of the Continuous Double Auction, Quantitative Finance, Vol. 3, pp. 481-514.