Fordham University
Department of Economics
Discussion Paper Series

Regime Identification in Limit Order Books

Rossen Trendafilov
Fordham University, Department of Economics
Erick W Rengifo
Fordham University, Department of Economics

Discussion Paper No: 2012-04
December 2012

Department of Economics
Fordham University
441 E Fordham Rd, Dealy Hall
Bronx, NY 10458
(718) 817-4048

Regime Identification in Limit Order Books *

Rossen Trendafilov† Erick W. Rengifo‡

September 14, 2012

Abstract

This article develops and implements a new methodology for identifying intra-day information regimes in limit order books. Based on Lehmann (2008), in an information regime all the information is trade related and arrives via order flow and, the fundamental value that underlines the prices does not change, it is simply translated by the size of the executed market order and the backfilling adjustment. During an information regime the best quotes and the underlying values follow a path defined by the limit order book. A change of information regime within a given day is shown to alter the provision of liquidity to the market with consequences for asset prices, trading behavior, and optimal trading strategies. By applying wavelet theory we have developed a methodology that allowed us to clearly identify information regimes. Our results show that information regimes have an impact on price formation and price discovery, including dynamic issues such as the process by which prices come to capture information over time. The discovery and identification of information regimes essentially uncovers the mechanism by which latent demands are translated into realized prices and volumes. These results empirically support Lehmann's theoretical model.

JEL Classification codes: G10, G12, C58

Keywords: Market microstructure, Information regimes, Limit order books, Wavelets, Wavelet multi-resolution analysis

*The authors would like to thank Duncan James, Jerome Lahaye and Hrishikesh D. Vinod. The usual disclaimers apply.

†Corresponding Author. Department of Economics at Fordham University New York. 441 East Fordham Road, Dealy Hall, Office E542, Bronx, NY 10458, USA. Phone: +1 (718) 817 4061, fax: +1 (718) 817 3518, e-mail: trendafilov@fordham.edu.

‡Department of Economics and the Center for International Policy Studies (CIPS) at Fordham University New York. 441 East Fordham Road, Dealy Hall, Office E513, Bronx, NY 10458, USA. Phone: +1 (718) 817 4061, fax: +1 (718) 817 3518, e-mail: rengifomina@fordham.edu.

1 Introduction

Nowadays most of the equity and derivatives markets fall in two categories. They are either pure electronic limit order markets or allow for customer limit orders along with on-exchange market making. Moreover, most of these modern financial markets provide their users with full price-quantity schedules, which can be seen as supply and demand curves. The worldwide spread of limit order markets created the need for economic and statistical models for these institutions.

This article develops and implements a new methodology for identifying intra-day changes in the limit order book information regimes. With the sole exception of Lehmann (2008), which is cited and used in this study, previous theoretical analysis do not account for the existence of information regimes, and no previous empirical work addresses the issue of intra-day changes in information regimes in the limit order book.

During an information regime the best quotes and the underlying values follow a path defined by the limit order book. A change of information regime within a given day is shown to alter the provision of liquidity to the market with consequences for asset prices, trading behavior, and optimal trading strategies. Information regimes impact price formation and price discovery, including dynamic issues such as the process by which prices come to capture information over time. The discovery and identification of information regimes essentially uncovers the mechanism by which latent demands are translated into realized prices and volumes.

An underlying, cornerstone paper analyzing an idealized electronic open limit order book was written by Glosten (1994). In it he established that prior to the arrival of a buy market order the marginal ask price schedule or the price of the last ask limit order executed should be equal to the upper-tail conditional expectation of the full information value. Analogously the price of the last bid limit order executed should be equal to the lower-tail conditional expectation.

Building on that, Lehmann (2008) defines information regimes under which the limit order book should operate. According to him, in the case that all the information is trade related and arrives via order flow, the fundamental value that underlines the price does not change, it is simply translated by the size of the executed market order and the backfilling adjustment. The backfilling adjustment is the process of replenishing some of the limit orders against which the market order was executed. In other words, moving up and down the book at a point of time, in an information regimen, is the same as trading over the time span of the regime for the purpose of price determination. During an information regime the best quotes follow a pattern prescribed by one and the same

limit order book. Thus given the size of the incoming market order or the depth of the book, the future best quotes can be estimated.

The best quotes process has three components: movements along the same limit order book within a regime; movements from one regime to another due to either a change in the underlying value or a change in the mapping between the price schedule and the underlying value. Simply put, in the price process there are movements along the curve (while in the same regime) and movements from one curve to another (moving from one regime to another). The current market microstructure models do not account for these two distinct intra-day phenomena and in that respect are incomplete.

The limit order book data is very detailed since the whole supply and demand curves are available. As helpful as it seems, it also poses a considerable challenge for the market participant. When the trader forms his expectations for the future value of the asset and chooses the limit or market order quantities and prices he is going to post, he has to condition his decision on everything that can affect the future unobservable value and price. This includes the full description of the limit order book plus the past trading histories and future expectations about the order flows. Thus the high dimensionality of the data poses a tremendous challenge for theoretical modeling and empirical estimations for trade applications. Further, besides the high dimensionality of the data, the book updates very frequently. The limit order book is a snapshot at a point of time. There is a limit order book at every second, and although it may not be different for every second it does change quite rapidly. As a result, there is a large number of limit order books even for a short period of time. Under this, the identification of information regimes reduces the noise that is inherent in intra-day limit order book data.

In order to identify different information regimes we propose the use of wavelet multi-resolution analysis. In the methodology part, we will provide the reasoning and demonstrate why such techniques are applicable and suitable. The rest of the study is organized as follows: the following section introduces the theoretical model developed by Lehmann (2008); Section 3 briefly introduces wavelet theory that we develop in detail in the technical appendix; Section 4 describes the data set used in this paper and, Section 5 presents our methodology followed for information regimes identification. Section 6 describes the results and the last section presents the summary and future research ideas.

2 Lehmann's Theoretical Model

Following Glosten (1994), the possibility of information motivated trades imply that the schedule of offers generally moves in upward direction. It costs more per share to execute

a large order versus small order. The agents are assumed to be risk neutral and come from a large population, and as such they are price takers. Together these two assumptions imply that the equilibrium is characterized by zero expected profit condition.

Glosten's study presents the case that in environment with discrete prices, the submitted bids and offers are related to, respectively, lower tail and upper tail conditional expectations. This is due to the discriminatory nature of the limit order book and adverse selection.

The agents who submit limit orders know neither the size of incoming market orders nor when they will occur. This has two implications: First, the limit orders bear the risk of not being executed. Second, the limit orders are exposed to the winner's curse problem of being adversely picked off if they are bypassed by the security value before the agent has a chance to cancel. There is the possibility of regret if the market order cleared the limit order and continued through the book, because the limit order seller (buyer) obviously could have executed at a better price. The limit order trader cannot prevent this from happening. However, the prices he quotes will reflect all of the previous risks he faces. For that reason, in the limit order book the asset is priced in such way that the expected full information value of the asset is conditional on the incoming order of a size just sufficient or larger to trigger execution. This is described as upper (lower) tail conditional expectations.

Limit order traders cannot condition on the quantity of the next market order (Q) when they place their orders. They know that the limit order to sell at price $P(q)$ is hit when the total trade size is at least as large as the cumulated depth of the book $q(P)$ up to that price.

Let's consider the case of the offer side of the limit order book. Denote with $P(q)$ the price schedule that gives the price of the last limit order executed, where q is the cumulated volume of the book up to that price. If m is the marginal valuation of incoming order, then $\underline{m}(q)$ is the value of m for which the last limit order executed is the one at $P(q)$, meaning the market order is just as big enough to cover that last limit order. In order for the agent not to regret his price, when his order is the last one executed, he prices at:

$$P(q) \quad \geq \quad E[X|m = \underline{m}(q)] \tag{2.1}$$

Where X is the full information value. However, since his order may not be the last one, he needs to take this into account and his price has to be at least as large as $E[X|m \geq \underline{m}(q)]$. That means that any agent submitting market order with valuation $m > \underline{m}(q)$ will purchase more than q ($Q > q$). The competitive equilibrium condition results in:

$$P(q) \quad = \quad E[X|m \geq \underline{m}(q)] \tag{2.2}$$

Based on Glosten's model, Bruce Lehmann extends the analysis to explain the dynamics of the limit order book through time - how the limit order book evolves. Lehmann (2008) begins with providing an analog for the fundamental theorem of asset pricing in order driven markets. The theorem conveys that in absence of arbitrage the asset prices follow a martingale process. The hypothesis of arbitrage free market, is a source of significant restrictions on asset prices. The complication, in the case of order driven markets, is the absence short sales, which is needed in the definition of arbitrage opportunities of first and second type. The absence of arbitrage opportunities insures the existence of positive state price vector and the whole arbitrage free pricing theory can be applied. Limit orders cannot be sold short (shorted). However the payoff of a zero net investment portfolio has a second meaning which is the payoff from a marginal change in an existing portfolio that is long in all of the assets. The analogue in a order driven market involves the ability to cancel and replace limit orders freely. Frictions and taxes, however are not a major issue if potential pre-tax arbitrage opportunities also represent after-tax arbitrage opportunities.

Lehmann (2008) places three assumptions (assuming that buy market orders are positive and sell market orders are negative):

- Assumption 1: Let $V_t(q)$ denote the asset value if a market order of size q arrives at time t. $V_t(q)$ is both strictly increasing in q and common knowledge among market participants.

- Assumption 2: $\text{sgn}_q[P_t(q) - V_t(q)] > 0, \forall q \in \mathbb{Q}_t$ where sgn_q is the sign of its argument. And $\mathbb{Q}_t \subseteq \mathbb{R}$, which is countable if there is lot size.

- Assumption 3: A market order can only arrive and be executed against the book after all limit order traders are satisfied with their order placements.

The last assumption can be plausible if market orders arrive according to a continuous time jump process, giving limit order traders time to refresh the book and, if the market agents who determine the marginal behavior of the book are active and perfectly competitive limit order trades.

The price priority and the strictly increasing $V_t(q)$ imply that $P_t(q + \text{sgn}_q dq) - P_t(q)$ and $P_t(q + \text{sgn}_q dq) - V_t(q)$ have the same sign and:

$$P_t(q + \text{sgn}_q dq) - P_t(q) \quad = \quad \lambda_t(q)[P_t(q + \text{sgn}_q dq) - V_t(q)] \tag{2.3}$$

with $\lambda_t(q) > 0$. See Figure (1) for graphical representation of this equation. In that figure we assume an unobservable underlying value to be linear.

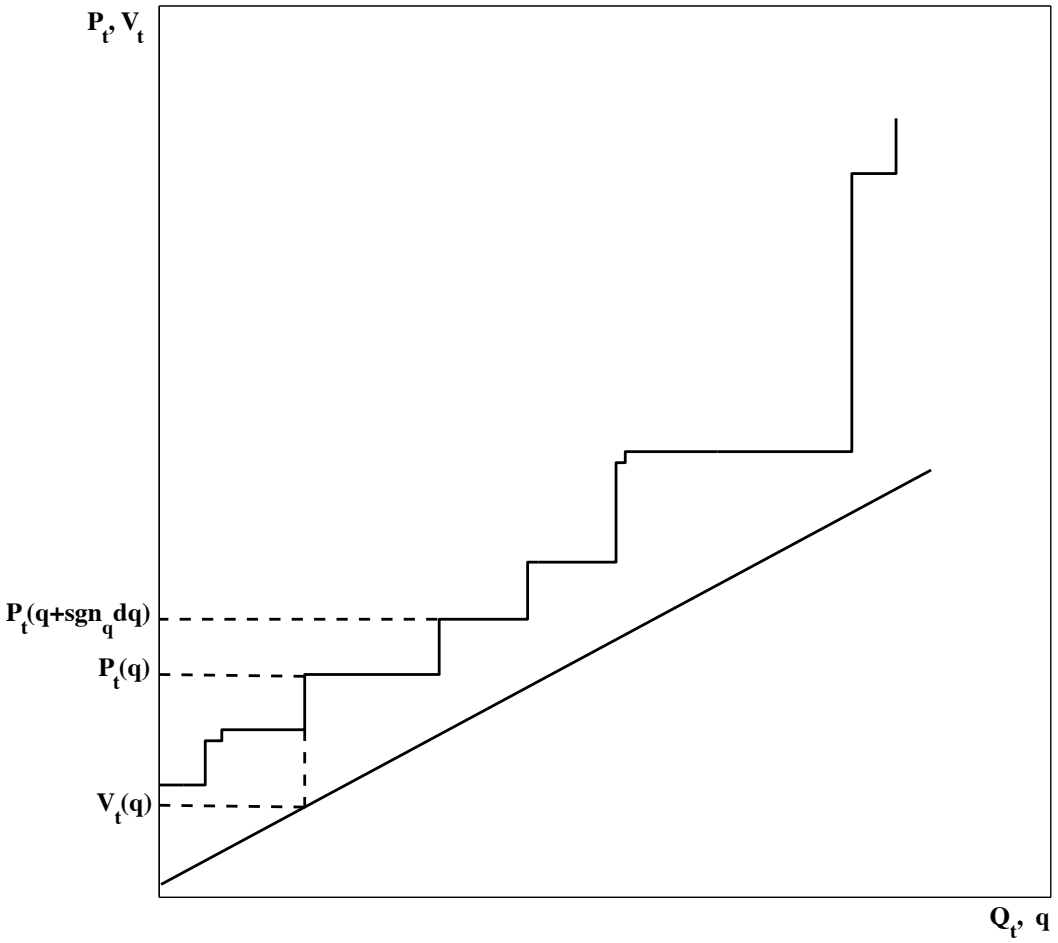

Figure 1: **Limit Order Book vs Asset Value**

This figure shows the offer (ask) side of the limit order book and an underlying asset value (assumed to be linear). The price priority and the strictly increasing $V_t(q)$ imply that $P_t(q + \text{sgn}_q dq) - P_t(q)$ and $P_t(q+\text{sgn}_q dq) - V_t(q)$ have the same sign. Moreover, $P_t(q+\text{sgn}_q dq) - P_t(q) = \lambda_t(q)[P_t(q+\text{sgn}_q dq) - V_t(q)]$ with $0 < \lambda_t(q) < 1$.

Rearranging equation 2.3, Lehmann (2008) derives the following pricing rule:

$$P_t(q) \quad = \quad \lambda_t(q)V_t(q) + [1 - \lambda_t(q)]P_t(q + \text{sign}_q dq) \tag{2.4}$$

Equation 2.3 implies that $\lambda_t(q) < 1$. It further insures that there are no arbitrage opportunities if and only if the limit order prices satisfy the pricing rule represented in equation 2.4. Thus, Equation 2.4 provides the means to obtain the price at a given tear

of the book if the underlying value at that tear, the $\lambda_t(q)$ and the price at $q + \text{sign}_q dq$ tier of the book are known.

$\lambda_t(q)$ is supported by the risk neutral probabilities $\psi_t = Pr(Q_t = q|\mathbb{T}_{t-1})$. Where $\mathbb{T}_{t-1}$ is the information set at time $t-1$, q is the cumulative volume of the book and Q_t is the size of the market order. From there $\lambda_t(q)$, represents the conditional probabilities that the market order $Q_t = q$ given that $Q_t \geq q$:

$$\begin{aligned} \lambda_t(q) &= Pr[Q_t = q|\text{sgn}_q \times Q_t \geq |q|, \mathbb{T}_{t-1}] \\ &= \frac{\psi_t(q)}{\int\limits_{\text{sgn}_q u \geq |u|} \psi_t(u)du} \end{aligned} \tag{2.5}$$

From there Lehmann, derives the same upper-tail valuation as in Glosten but with risk neutral probabilities replacing the actual ones in the Golsten's paper[1] .

If the three assumptions mentioned before hold, Lehmann's first proposition states that, there is positive pricing rule supported by a set of unique state prices $\psi_t(q) > 0, \forall q \in \mathbb{Q}_t$ if and only if there are no arbitrage opportunities. Contained in the proposition is the implication that any upward sloping marginal price schedule can be rationalized as being arbitrage free in the sense of the proposition.

The value of the asset right after the execution of a market order of size Q_{t-1} is $V_{t-1}(Q_{t-1})$. The asset value before the arrival of the next market order is approximately equal to the midquote:

$$P_t(0) = V_{t-1}(Q_{t-1}) + v_t(0);\ E_\psi[v_t(0)|Q_{t-1}, \mathbb{T}_{t-2}] = 0 \tag{2.6}$$

Where $v_t(0)$ is the risk neutral martingale increment. This equation along with equation 2.4 forms the following relation:

$$\begin{aligned} P_t(0) &= P_{t-1}(Q_{t-1} + \text{sgn}_{Q_{t-1}} dq) \\ &\quad - \frac{P_{t-1}(Q_{t-1} + \text{sgn}_{Q_{t-1}} dq) - P_{t-1}(Q_{t-1})}{\lambda_{t-1}(Q_{t-1})} + v_t(0) \end{aligned} \tag{2.7}$$

Bruce Lehmann (2008)'s second proposition is that under the three assumptions that he places (see above), the limit order book has no holes in $\mathbb{Q}_t$ if there are no arbitrage opportunities. Moreover, the marginal price schedule is continuous in q at any $q \neq 0$ if $V_t(q)$ is continuous and if there are no indivisibilities in market order sizes.[2]

One scenario of interest that Lehmann discusses is when both the order flow dependent asset values and state prices depend only on the cumulative signed order flow. In that

[1] For the derivation see Lehmann (2008) 2008 page 11

[2] This second proposition does not hold empirically, because holes do exist in reality in the limit order book and there is minimum lot size.

case, when a market buy order of size Q_{t-1} arrives at $t-1$, if it does not exhaust the depth on the offer side, it will walk up the book until it reaches $P_{t-1}(Q_{t-1})$ and the asset value after the trade is $V_{t-1}(Q_{t-1})$. The highest unexecuted offer at $P_{t-1}(Q_{t-1})$ will be the new best offer. The original best bid will be Q_{t-1} shares away from the new best bid, because it will take a market order of that size to reach it. Order state prices and values stay unchanged when no additional information arrives at the market, besides the execution of the market order. Thus:

$$\begin{aligned} V_t(q) &= V_{t-1}(q + \mathrm{sgn}_q Q_{t-1}) \\ \psi_t(q) &= \psi_{t-1}(q + \mathrm{sgn}_q Q_{t-1}) \end{aligned} \tag{2.8}$$

This means that the state prices and values are simply translated by the size of the market order. The reason for that is the risk that a market order of (Q_{t-1}) shares at time $t-1$ will be followed by one for q shares is the same as that a market order with size $(q + Q_{t-1})$ will arrive at $t-1$ under the restriction that there is no new information besides the market order. The assumption that all the information is related to trade is frequent in dynamic market microstructure models.

The limit order traders will backfill the portion of the book that was cleared by the market order with bid prices at which they will be willing to buy up to (Q_{t-1}) shares. The backfilled best bids are weighted average of the prior best bid and offer, and prior offer at $q + Q_{t-1}$ shares.

Another way to look at (Q_{t-1}) is that it could be considered as net order flow from some earlier time to time $t-1$, with the assumption that the only information arrived with the market orders. Based on that Lehmann defines an information regime or epoch as a period during which it is common knowledge that asset values and state prices in different order flow states satisfy the following:

$$\begin{aligned} V_{t-1}(q+q') &\equiv E_\psi[\tilde{V}|Q_{t-1} = q + q', \mathbb{I}_{t-2}] \\ &= E_\psi[\tilde{V}|Q_t = q', Q_{t-1} = q, \mathbb{I}_{t-2}] \\ &\equiv V_t(q'|Q_{t-1} = q) \end{aligned} \tag{2.9}$$

$$\begin{aligned} \psi_{t-1}(q+q') &\equiv E_\psi[1_{Q_{t-1}=q+q'}|\mathbb{I}_{t-2}] \\ &= E_\psi[1_{Q_t=q'}|Q_{t-1} = q, \mathbb{I}_{t-2}] \\ &\equiv \psi_t(q'|Q_{t-1} = q) \end{aligned} \tag{2.10}$$

where $\mathbb{I}$ is the information set; $\mathbb{I}_{t-2} = \{Q_{t-2}, \ldots, Q_2, Q_1, \mathbb{I}_0\}$ and $\{q, q', q+q'\} \in \mathbb{Q}_t$. Equations 2.9 and 2.10 tell us that there is no change in the fundamental value and the

risk neutral probabilities after execution of market order $Q_{t-1} = q$, when trade happens in the same information regime. In this paper we concentrate our efforts to empirically test Equation 2.9.

During an information regime it can be imagined that there is just a single market order in the amount if cumulative signed volume that walks up and down the time one limit order book. The movement up and down the book is the same as trading for the purposes of determining prices when all information arrives trough the market orders.

It is plausible to assume that the agents act as if they can cancel and replace limit orders up to their preferences before the arrival of the next market order, because it is easy to create algorithmic trading strategy over an information regime. There is also no risk for limit order to be picked off in a change of the information regime, as long as this change is common knowledge.

3 Wavelet Theory

The wavelet, as the name may suggest, is a localized wave form. It is a wave-like oscillation with an amplitude that starts at zero, increases, and then decreases back to zero. It can be pictured as a brief oscillation like the one that might be seen on a heart monitor. Figure (3) graphically represents several wavelets. The wavelet function $\varphi(t)$ [3] must satisfies certain mathematical criteria which are listed in the technical appendix.

In the case of non stationary data, which may contain transient phenomena (i.e. aperiodic), the wavelet basis functions are precisely localized in time and frequency. The wavelet function separates the data into different frequency components, and then study each component with a resolution matched to its scale. The wavelet transform provides efficient and complete representation of the signal.[4] The wavelet is manipulated trough a process of translation and dilation (or scaling) in order to better fit the signal.

There are continuous and discrete wavelet transforms. In this paper we apply the discrete wavelet transform because it is fast to compute and, because the size of the transformed data is the same as the size of the original data.[5] These are important considerations since we are working with intra-daily data and, as it is well known, the size of these data set are rather large.

[3] In the literature the wavelet function is usually denoted by ψ. However, this notation conflicts with the one used in Lehmann's model for state prices. For that reason we use φ to indicate the wavelet function.

[4] The term signal is used to indicate a function that conveys information about the behavior of some phenomenon. In this article the signal is the limit order book.

[5] In the case of the continuous wavelet transformation the size of the transformed data is considerably larger than the original data.

Having in mind that one goal in this paper is to identify information regimes or epochs as defined theoretically by Lehmann (2008), in this section we briefly introduce the discrete wavelet transform, concentrating our explanation on a particular family of discrete wavelets known as Daubechies Wavelet Family. We provide a detailed explanation of wavelet theory in the technical appendix.

3.1 Discrete Wavelet Transform

Instead of working with the whole wavelet function, one can just use a handful of coefficients describing the wavelet, which greatly simplifies the calculations. These coefficients are selected by a process of subsampling. Any continuous wavelet function that is subsampled is called discrete wavelet transform (DWT). If the original series can be reconstructed from the discrete families of the wavelet functions created by subsampling of the continuous wavelet transform (CWT), then the DWT constitutes a full representation of the time series. Certain conditions must be met met by the wavelets in order for them to provide stable and complete representation and reconstruction of the time series. These conditions are provided by Frame theory.[6]

One way to sample the wavelet parameters a and b is by logarithmic discretization of the scale a and connect this to the size of the steps taken between b locations. To achieve this one moves in discrete steps to each location b which are proportional to the scale a. Thus:

$$\varphi(t)_{m,n} = \frac{1}{\sqrt{a_0^m}}\varphi(\frac{t - nb_0a_0^m}{a_0^m}) \tag{3.1}$$

where $\varphi(t)_{m,n}$ is the wavelet function. A natural way to choose the discrete wavelet parameters a and b is 2 an 1 respectively. This is known as dyadic grid arrangement and is the simplest and most efficient discretization for practical purposes. Thus:

$$\varphi(t)_{m,n} = \frac{1}{\sqrt{2^m}}\varphi(\frac{t - n2^m}{2^m}) \tag{3.2}$$

where m and n control the wavelet position and scale. In our study we use the dyadic grid, which imposes the condition that the length of the data series has to be a two to a power number (2^k). Using the above equation the discrete wavelet transform (DWT) can be written in the following way, where the $T_{n,m}$ are the wavelet detail coefficients and $x(t)$ is the signal:

[6]For more details on that topic please refer to the technical appendix.

$$T_{n,m} = \int_{-\infty}^{\infty} x(t)\varphi_{n,m}(t)dt \tag{3.3}$$

The detail wavelet coefficients can be tiled or indexed as shown by Figure (2). Our methodology is based on the analysis of the distribution of the detail wavelet coefficients. We provide a detailed description of the way we use this distribution in the methodology section. As usual, for further mathematical explanation we refer the reader to the technical appendix.

level index 4	$T_{4.0}$	$T_{4.1}$	$T_{4.2}$												$T_{4.15}$
level index 3	$T_{3.0}$		$T_{3.1}$		$T_{3.2}$								$T_{3.7}$		
level index 2	$T_{2.0}$				$T_{2.1}$				$T_{2.2}$				$T_{2.3}$		
level index 1	$T_{1.0}$								$T_{1.1}$						
level index 0	$T_{0.0}$														
level index -1	signal mean component														

Figure 2: **Wavelet Tiling**

Figure shows the relation of the discrete wavelet transform detail coefficients to the time frequency plane.

3.1.1 Daubechies Wavelet Family

In this study, we use Daubechies family of wavelets. This family of wavelets, as denoted by Fugal (2009), are robust, fast and adaptable. They are in wide use for identifying signals with both time and frequency characteristics. They are non-symmetric and are especially well suited to non-symmetrical transients.

The Daubechies family is excellent for the high number of vanishing moments it can accommodate. The number of vanishing moments corresponds to the ability of the wavelet to correlate with polynomials of different degrees. Thus wavelets with certain number of vanishing moments are more suitable to identify a polynomial of certain degree.[7] We conducted the study using wavelets with 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 vanishing moments as well as the summed up results of the whole family.

One of the reasons to use this family of wavelets is that the limit order book is a step function, where it is not known beforehand the number of steps of the overlapping portion that is to be compared. In the methodology section we provide argument for using only

[7]For more extensive treatment of vanishing moment and examples see Fugal (2009) pages 175-183.

one specific member: Daubechies 6. Figure (3) shows the different wavelets used in the study.

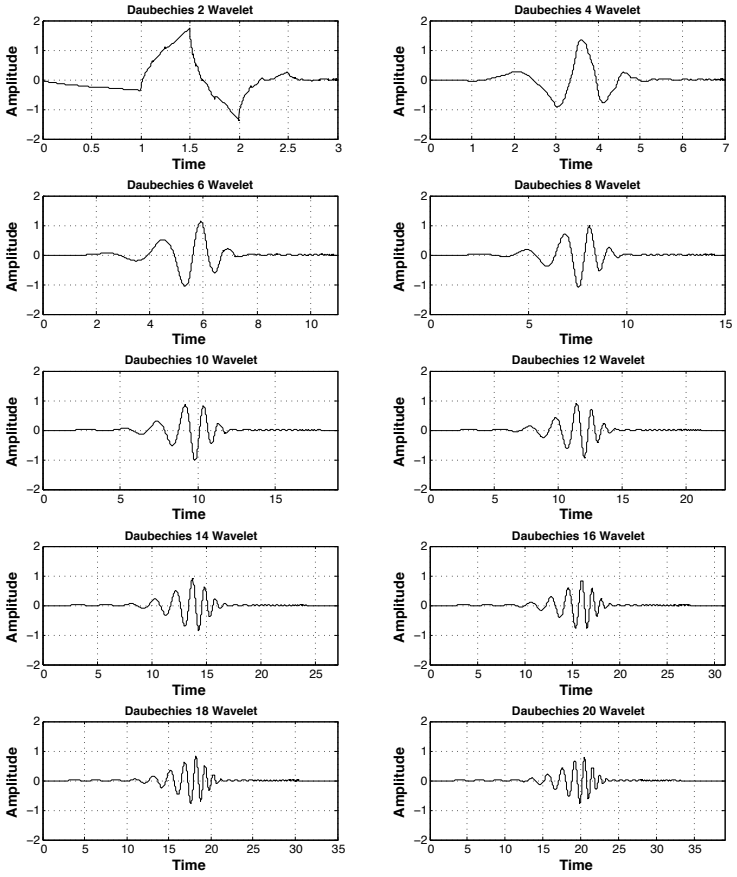

Figure 3: **Daubechies Wavelets**

This figure shows Daubechies family of wavelets with 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 vanishing moments. The number of vanishing moments pertains to the ability of the wavelet to correlate with polynomials of different degrees.

4 Limit Order Book Data

The dataset used in the study is similar to the data used by Grammig, Heinen, and Rengifo (2004). We follow some of their description.

The data contains complete information about Xetra market events, including all entries, cancelations, revisions, expirations, partial-fills and full-fills of market and limit orders that occurred between August 1, 1999 and September 13 1999 (30 trading days). Market events were extracted for DCX (Daimler Chrisler).

Xetra trading hours at Frankfurt Stock Exchange (FSE) runs from 8.30 a.m to 5.00 p.m. CET. The trading day begins and ends with call auctions and is interrupted by another call auction which is conducted at 12.00 p.m. CET. The regular, continuous trading process is organized as a double auction mechanism with automatic matching of orders based on price and time priority. The following features additional to the ones mentioned before should be noted.

- Market orders exceeding the volume at the best quote are allowed to "walk up the book". The market orders are guaranteed immediate full execution, at the cost of incurring a higher price impact on the trades.

- Before 2002, and during the time interval from which the data is taken, only round lot order sizes could be filled during continuous trading hours. A Xetra round lot was defined as a multiple of 100 shares. Execution of odd-lot parts of an order - this is an integer valued fraction of one hundred shares - was possible only during call auctions.

- Hidden limit orders (or iceberg orders) were not allowed during the period from which the data is taken.

- Xetra trading is completely anonymous, i.e. the Xetra order book does not reveal the identity of the traders submitting market or limit orders.

For the purpose of the study we use the data from 8.40 a.m to 4.50 p.m. We cut 10 minutes form the beginning and the end of the trading day in order to exclude any possible extreme values that my occur due to the opening and closing auctions.

Since the lot is equal to 100 shares we insert additional columns for the volumes, such that the cumulative volume is the same but each price volume combination has always a volume of 100. For example, if we have the best offer at \$30 for 200 shares and the second best offer at \$30.5 for 300 shares. After the transformation we have price of \$30 and 100 shares then again price of \$30 and 100 shares. For the second best offer a price of \$30.5 for a hundred shares, and this is repeated 3 times. The reason to do this is to have equal increments in volume for each price in order to facilitate the calculation of the wavelet transform.

The data, in terms of complete limit order books, is presented each time that there is a change due to new limit orders, market orders or cancelations. This means that our data is in *limit order book time*. To switch to clock time where it is needed, we also have created additional file, where we have added missing limit order books for each second equal to the previous book until a change occurs.

An additional issue to consider is that sometimes there are more than one record for each book at a point of time, due to separate posting of limit orders. In that case we kept the last record, which shows the limit order book where all orders have been taken into consideration.

The market order data is in a separate file and it is in market order time, which means that there is not market order in each second. In order to switch to limit order book time or clock time we have inserted zeros at the points of time with no market order.

5 Methodology

This section has two parts. In the first one we explain the methodology followed for information regime detection and, in the second one we demonstrate how to choose the best fitting wavelet and how to choose the appropriate lag. In this paper we use the word lag to express the distance between books that we compare. For example, lag 1 implies that we are comparing the first book with the second, the second with the third and so on. A lag of 10 implies that we compare the fist book with the eleventh, the second with the twelfth, and so on.

The study is conducted in *limit order book time*. To understand this, note that even though in *clock time* there is limit order book in each second, the limit order book does not change in each of these seconds. In this sense there are many consecutive books which are identical. By looking only at the book when it changes (*limit order book time*) we avoid this inflation with identical books and speeding up calculations. However, the results are the same in clock time and we preserve the time stamp of each limit order book for other type of analysis.

In this section we present the methodology together with examples using our data set to better explain it.

5.1 Detection of Regimes

The detection of the information regimes is performed in three steps: alignment of the limit order books, frequency decomposition by discrete wavelet transform and, comparison of the distributions of the detail wavelet coefficients by frequency via two-sample Kolmogorov-Smirnov test. We explain each of them in the next sections.

5.1.1 Limit Order Book Alignment

As seen in Figure (4), when we plot the data for each limit order book we end up with a lot of (apparently) different curves. In this figure we plotted all the different limit order bid schedules for one day. As one can see it is hard to identify any intra-daily regimes at which the exchange may be operating.

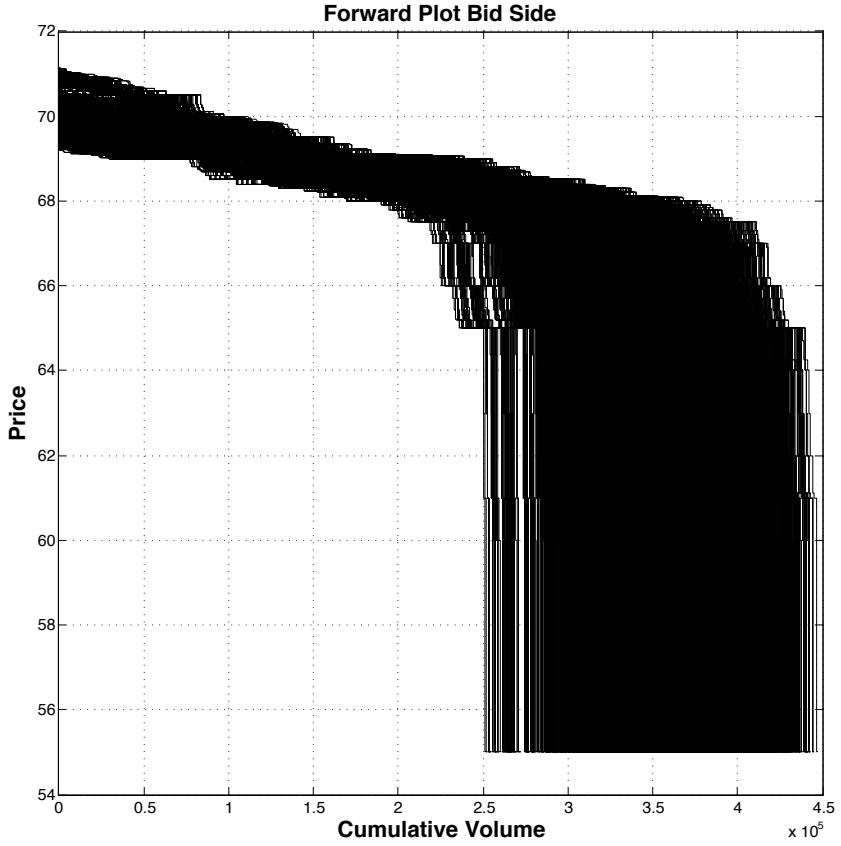

Figure 4: **Bid Limit Order Books Forward Plotted**

This figure shows the bid side of one day limit order books forward plotted (7408 books). The books are plotted starting from the best quote and proceeding to the back end of the book.

In an information regime, there is one fundamental value that does not change and, where the market orders and the best quote moves up and down over the same fundamental values and price schedules. An empirical fact is that most of the activity in the limit order book happens close to the best quote of the book where the new orders and cancelations are submitted, and where the book interacts with the market orders. Of course, there are orders and cancelations at higher tiers but they are infrequent.

One important facts to note is that the orders that are at the end of the book stay unchanged for a considerable amount of time. In a sense they are staled, not canceled

by the traders who posted them (since it does not cost anything to keep that orders). Moreover, the system also does not automatically cancel them for a period of one year.

All this means that, at least for the period of one trading day (or big portions of the day) the backward part of the book is fixed and it is the tip portion(close to the best quotes) that changes. This is important because limit order books change from one to another and have different lengths. In order to identify different regimes, the books need to be aligned in such way that:

$$V_t(q) \quad = \quad V_{t-1}(q + Q_{t-1}) \tag{5.1}$$

is within the same regime. Recall that q is the cumulated volume of the book up to some tier, $V_t(q)$ is the value of the asset at q and Q_t is the size of the market order. According to thsi equation, much of the limit order books in the particular regime should be overlapping. The parts that are not overlapping are the ones that create the difference in the depth (length) of the book for consecutive books.

The fact that the back portion and especially the very top (on the ask side) or the bottom (on the bid side) does not change, gives the opportunity to align the books from common point of reference - the end of the limit order book. Thus, instead of arranging at the books from best quote to end (forward), we proceed in the reverse direction (backwards), arranging and plotting them form very back end of the book and proceeding towards the best quote. In this case the very back of the book has cumulative volume 0 and the best quote has the maximum cumulative volume, which is exactly the opposite from the forward (usual) case where one starts from the best quote and proceeds to the back end. Figure (5) depicts the backward aligning and plotting. Both Figures (4) and (5) plot of the same limit order books, with the only difference from which side the plot starts. In Figure (5) one can visually see that the different limit order books look like they are clustered in more distinct bunches.

5.1.2 Frequency Decomposition by Discrete Wavelet Transform

Once the books are aligned we proceed to identify which of them belong to the same regime. To do so, we need to compare the shapes of consecutive books and identify when they significantly different from each other. In order to do that, we select the parts of the two books that are completely overlapping and that are close to the best quote of the book. Since the back part of the book is relatively fixed, what one needs to compare are the parts closer to the best quote.

To compare the overlapping parts, we apply fast discrete wavelet transform and to

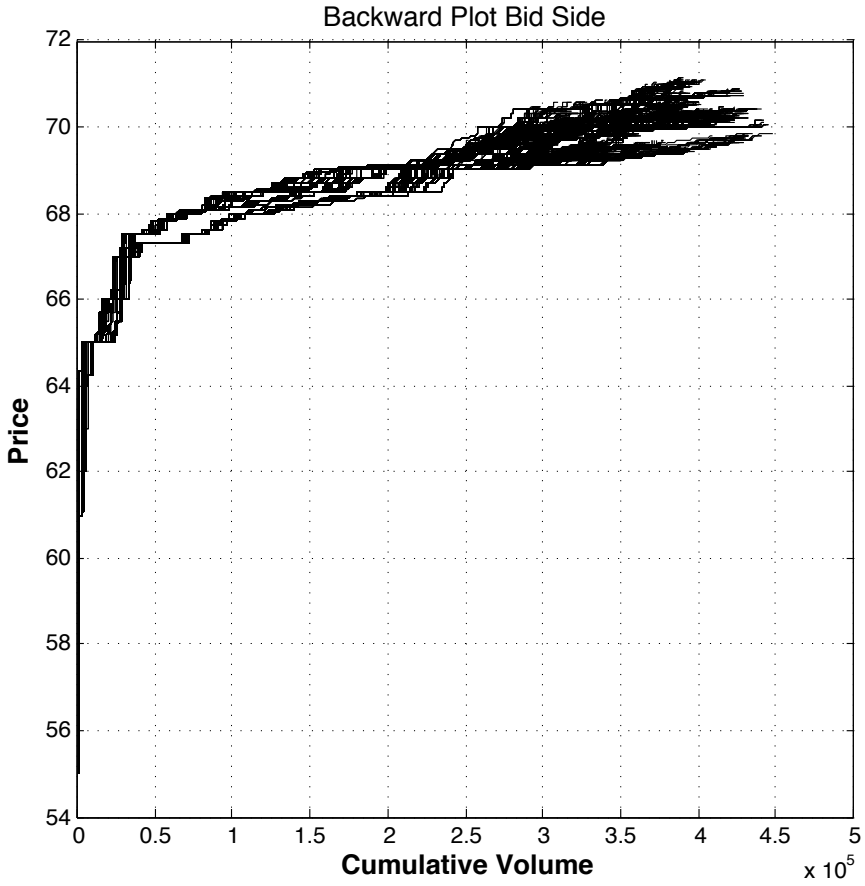

Figure 5: **Bid Limit Order Books Backward Plotted**

This figure shows the bid side of a one day limit order books backward plotted (7408 books). The books are plotted starting from the very end of the book and proceeding towards the best quote. The back portion and especially the very end does not change for a considerably period of time. This gives the opportunity to align the books from common point of reference - the end of the limit order book. Both this figure and Figure (4), plot the same limit order books, with the only difference from which side the plot starts.

decompose each of them by frequency. Since the wavelets operate on dyadic grid the length of the overlapping parts has to be a power of 2. We have found that the greatest length that accommodates all the data is 1024 (2^{10}) 100-share units of the length of the book (the minimum lot size is 100). That translates in to $1024 * 100 = 102,400$ shares of cumulative volume depth. In the data set, the shortest book has cumulative volume of $196,400$, which is larger than $102,400$ but smaller than $2048 * 100 = 204,800$ (where 2048 is equal 2^{11}). Thus the greatest length that accommodates all the data is 1024. Recall that the limit order book is in price-volume dimension, so we compare 1024 100-share units of the length of the book.

Once the parts of the limit order book are decomposed by frequency and due to the downsampling in the discrete wavelets in the highest frequency there are 512 detail wavelet

coefficients (2^9), in the next frequency there are 256 (2^8) and, so on (see Figure (2)). In our case, the detailed coefficients are organized in volume-frequency plane.[8] Volume is on the horizontal axis and consists of 100-share units, meaning that the distance between each point on the axis is 100 shares. Frequency is on the vertical axis with highest frequencies on top and lowest on the bottom.

On the top are the 512 highest frequency detail coefficients that span the shortest volume. Below are the 256 coefficients which are in the second highest frequency and span second shortest volume. Continuing all the way on the bottom are four coefficients at lowest frequency that was used for the decomposition. This corresponds to the level index 2 in Figure (2). Obviously the decomposition is not taken all the way when there is only one coefficient (level index 0 in Figure (2)), because it is not needed for the purpose of comparing the wavelet coefficient distributions. Graphical presentation of the actual detail coefficients and their tilling is presented in Figure (18) in the appendix.

5.1.3 Two-Sample Kolmogorov-Smirnov Test

The next step is to compare the distributions of the detail coefficients by frequency for two consecutive limit order books. We compare the distribution of the coefficients in the highest frequency, then the second highest and so on. To do so, we use two-sample Kolmogorov-Smirnov test to compare the distributions.

The null hypothesis is that the two samples are from the same continuous distribution. The alternative hypothesis is that they are from different continuous distributions. The result is 1 if the test rejects the null hypothesis at the 5% significance level and 0 otherwise.

That procedure examines the books frequency by frequency. Since we compare just the distributions of the detail coefficients, the positioning of the volume series of the wavelets is lost. However, in this particular case this is not a serious problem because the books are increasing in volume, the back parts are fixed and we use the closest 1024 100-share units to the best quote.

If the coefficients come from the same distribution, the implications is that certain frequencies of both books carry the same amount of the variability of the signal. Further, this will suggest that the books have same or very similar shape and thus, in our context, this will imply that they are in the same regime.

It is a question of choice for the researcher to decide to compare books one right after the other, or to skip several and to compare, for example, the first to eleventh, then the second to the twelfth and so on. Otherwise said, the question is how big the lag (distance)

[8]Recall that for time series the detailed coefficients are organized by time.

between the two books should be. The closer the books (in time sense) the more alike they are and some regime switches might not be captured. This is possible since even though the difference between the books is very small, the books could still drift away. In order to avoid this potential problem one may want to skip several books and compare not consecutive books but books separated n-lags, where $n \geq 2$.

On the other hand, the further the books are in time from each other, the bigger the differences between them. This can cause additional noise and one can identify switches at points where actual regimes do not change. Moreover the opportunity of practical application of these regimes diminishes because by the time one regime is identified it can be almost over. In the following section we provide several guidelines for selecting a certain lag (distance between books). For the following procedure and graphs we have selected 16 lags and Daubechies 6 wavelet (we explain the reasons later in the next sections).

Next, we explain in detail and using an example the steps to apply the above procedure for all books, frequencies and Daubechies 6 wavelet.

1. Obtain the detail coefficients by applying fast discrete wavelet transform.

2. Compare the distributions of the wavelet coefficients frequency by frequency for each two books that are 16 observations away from each other (16 lags).

3. Assign 1 if the results of the Kolmogorov-Smirnov test indicate different distributions, 0 otherwise. In this case there is a vector of 0s and 1s for each frequency.

4. The result is a 0 or 1 for each frequency for each pair of books that are compared.

5. Lastly, we impose a minimum size of at least 31 (more than 30) books for a regime. Even though there are a number of shorter regimes we required to have at least 30 or more observations for statistical purposes.

The outcome of the first four steps of the procedure can be seen in Figure (6). In the figure the procedure has been done for both the bid and the ask side and the results are combined.[9] The "o" indicate the results for the ask side and the "+" indicate the results on the bid side. The frequencies are arranged from high to low, from top to bottom of the graph. The regimes (instances where the books are overlapping) are represented by the areas that are clear. These are the areas where the Kolmogorov-Smirnov test results are zero meaning that the wavelet coefficients could come from the same distribution.

[9]The results for the bid and the ask side need to be combined in order to insure that we capture asset value changes in either side.

Due to the high dimensionality of the data, the figure shows only the four highest frequencies and 1200 observations. It can be seen that at certain points at the highest frequency (top figure) there are clusters of points, which would indicate the existence of transition periods (periods between regimes) and, at the same time indicate where the regimes change (at the beginning and end of these transition periods). One can also observe that there are clusters in the lower frequencies, which in combination with the highest one should help to better identify the points where the regimes switch.

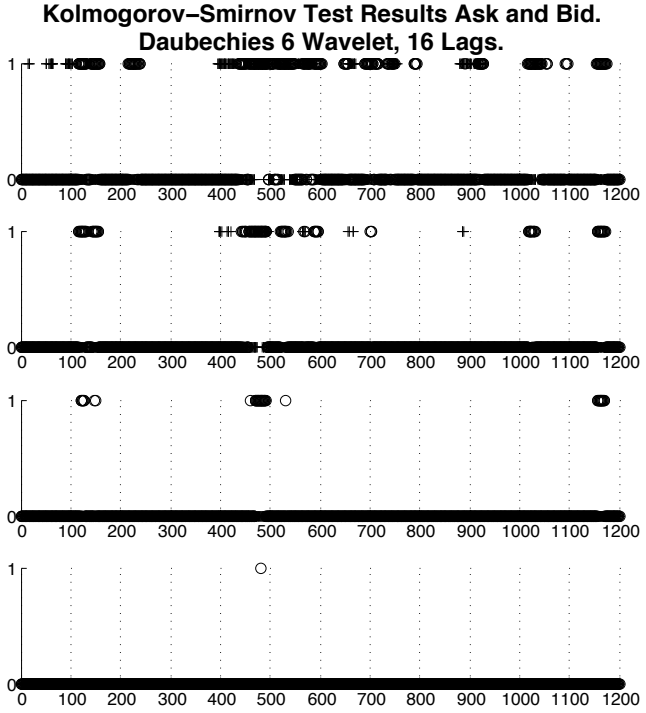

Figure 6: **Two-Sample Kolmogorov-Smirnov Test Results**

This figure shows the summed results for comparison of the distributions of the Daubechies 6 wavelet coefficients for both bid and ask by frequency. The frequencies are arranged from high to low, from top to bottom of the graph. The 'o' depict the results for the ask side and the '+' depict the results for the bid side. Due to the high dimensionality of the data, the figure shows only the four highest frequencies and 1200 observations. The regimes are the clear areas where zeros are obtained in the Kolmogorov-Smirnov test showing that the wavelet coefficients come from the same distribution.

In order to identify the information regimes, we add up the results across the different frequencies. In this study we use only the three highest frequencies. One reason is that since there is downsampling in the process of the wavelet decomposition, the lower frequencies contain a fewer coefficients and the samples that are compared are smaller (in the case of the lowest frequencies the coefficients are insufficient). Another reason is that, since the overlapping parts that are compared are close to the best quotes of the book,

that lower frequency portion of the book is relatively flatter and thus, there should not be much variability at these frequencies.

Based on the fifth step of the procedure, any regime shorter than 31 books is eliminated. The results can be seen on Figure (7), where the shaded area represent the regimes and the white areas the transition periods. The upper part of the figure presents the way the regimes and the transitions are distributed trough a given day in limit order book time. There are 62 regimes during that day (day 1, August 2, 1999). To have a better macroscopic view, the lower panel presents only the first 10 regimes in that day.

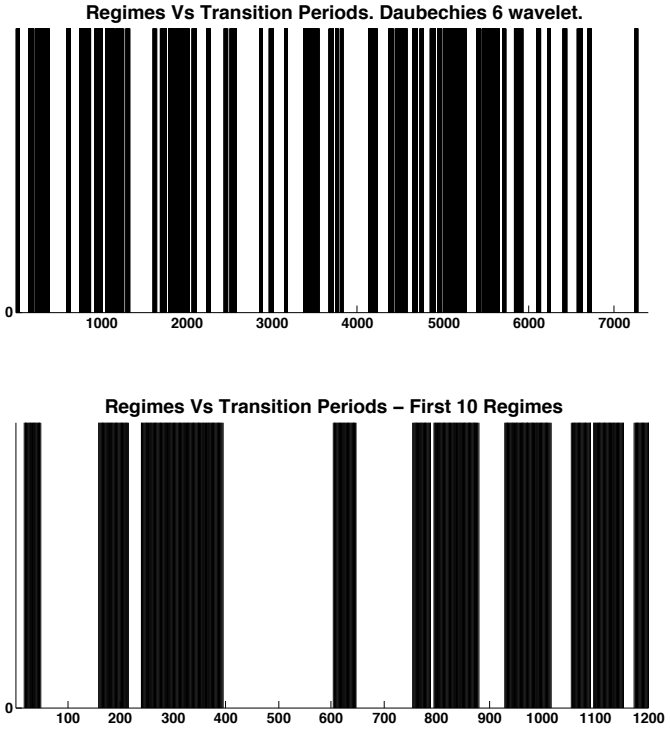

Figure 7: **Regimes Versus Transitions**

This figure shows the regimes as shaded areas and the white areas are the transitions. The top graph shows the results for entire day (day 1, August 2, 1999). There are 62 regimes during that day. The bottom graph shows the first 10 regimes and transitions.

Finally, Figure (8) presents two consecutive regimes, identified by the above procedure. The new regime (regime 2) is plotted bellow and the old regime (regime 1) is plotted above. These two regimes correspond to the fourth and fifth shaded areas in the bottom plot of Figure (7).[10] In Figure (19) in the appendix, these two regimes are plotted against the all the books for the day.

It can be clearly seen on Figure (8) that a change of information regime within a

[10]Similar results are available upon request.

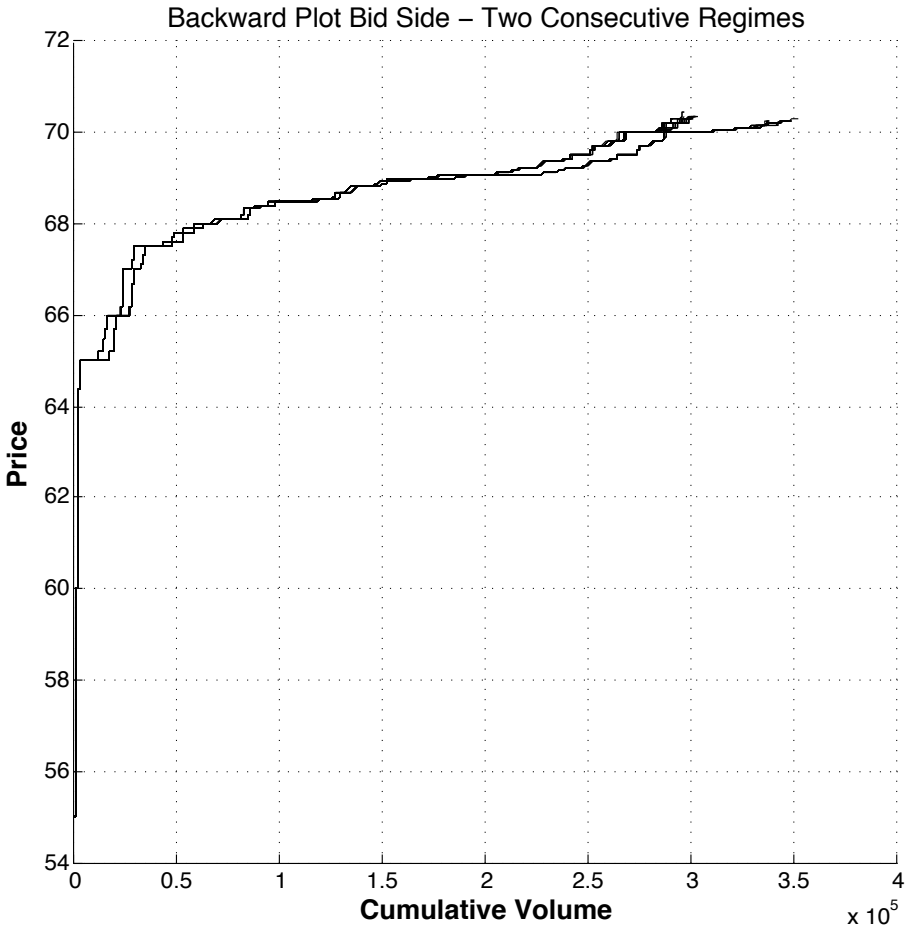

Figure 8: **Two Consecutive Regimes**

The new regime is plotted bellow: books 755 - 788, regime size = 33, from sec 33354 to 33515 sec., duration: 161 sec. The old regime is plotted above: books 604 - 647, regime size = 43, from sec 32916 to 33056 sec., duration: 140 sec. Transition size = 106 books, from sec 33060 to 33350 sec., duration: 290 sec.

given day alters the provision of liquidity to the market. The two regimes provide two different price impact functions for the incoming market orders and thus impact the asset prices and the behavior of the market participants differently. For example, a straight forward strategy can be formulated in this case. Let's assume that a trader wants to buy a large number of shares and is setting his strategy to optimally execute them in order to minimize the price impact of his trades. If this trader knows the regime in which he is (say the one bellow, regime 2), he can easily forecast the cumulative volume in this regime. With this information he could decide to enter a buy market order (immediate execution) to buy a large number of shares to fulfill his requirements. This should a be a correct strategy since the cumulative volume observed close to the best quotes is large and by doing so, he would lower the price impact of his trade. This is not the case if the

regime above (regime 1) is considered.

5.2 Parametrization

In the following two subsections we discuss the choice of specific wavelet and the subsequently the choice of lags that best matches the signal. Recall that in this paper we use the word lag to express the distance between books that we compare. For example, lag 1 implies that we are comparing the first book with the second, the second with the third and so on. A lag of 10 implies that we compare the fist book with the eleventh, the second with the twelfth, and so on.

5.2.1 Choice of Wavelet

In order to find the wavelet that best matches the signal we count the regimes of any size discovered by a given Daubechies wavelet member for all the lags from 1 to 50. Next, we determine which wavelet identifies the largest number of regimes per day. Looking for the maximum number of regimes can be argumented by the fact that the more regimes are identified, the more of the changes in the limit order books are also successfully identified. The reasoning for this is that the wavelet (or wavelets) that repeatedly identifies the largest number of regimes for any lag and any day would be the one that is best fitted for the limit order book data, basically because it is the most sensitive to the changes in the book. In a very persistent manner we have identified this to be Daubechies 6.

The results are summarized on the top graph of Figure (9). On the horizontal axis are the days and their corresponding lags. There are 30 days with 50 lags for each day. So the first 50 observations correspond to lags from 1 to 50 for day 1. Recall that lag 1 implies that we are comparing two consecutive books; lag two that we compare the first with the third, the second with the forth, and so on. On the vertical axis are the different Daubechies members, where the data point represents the member with the maximum number of regimes for the given day and lag. It is clear that Daubechies 6 is the best choice since that wavelet outperforms the rest in 99.47 % of the cases.

The second graph of the same figure shows the number of regimes of any size for Daubechies 6 wavelet versus different days and lags. On the horizontal axis are the days and the lags in the same fashion as above. On the vertical axis are the number of regimes (of any size) for a given lag and for a given day.

We can see that for any day with the increase of the lags the number of regimes of any size also increases. This makes sense, because with the increase of the lag between two books, there are more changes in the books and their shapes become more and more

different. Note that in this case the number of regimes increases but their size (in number of limit order books in the regime) decreases.

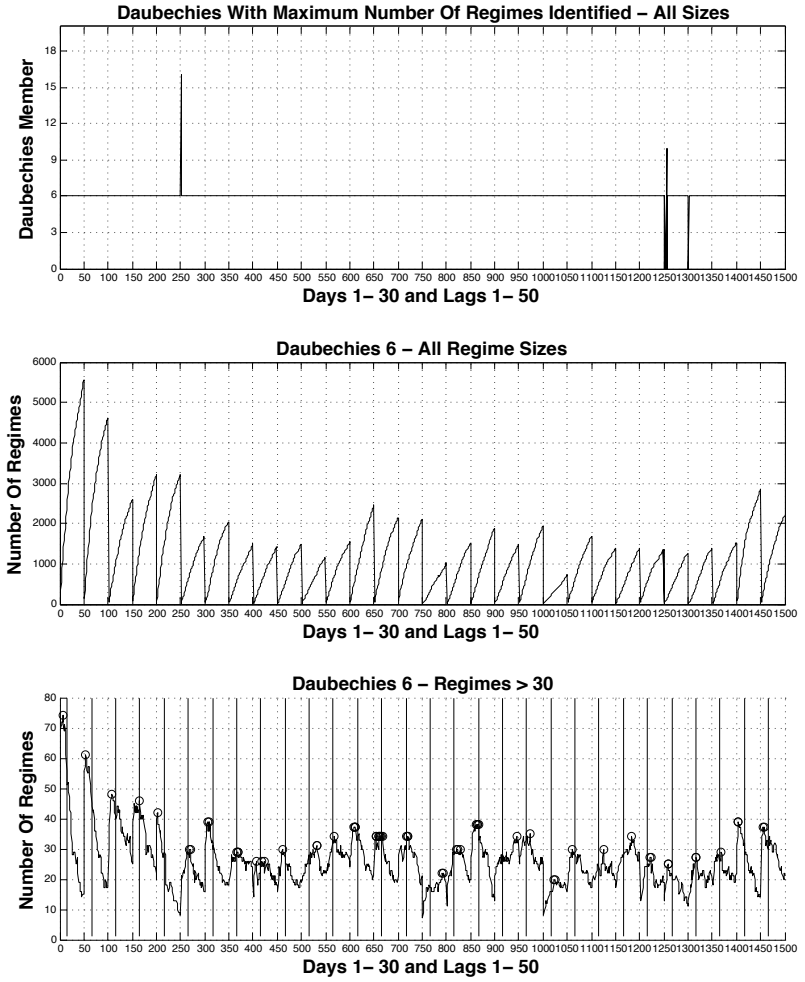

Figure 9: **Daubechies Wavelets Vs Number Of Regimes**

The top figure shows Daubechies wavelets with maximum number of regimes identified versus different days and lags. On the horizontal axis are the days and their corresponding lags. There are 30 days with 50 lags for each day. So the first 50 observations correspond to lags from 1 to 50 for day 1. Recall that lag 1 implies that we are comparing two consecutive books; lag two that we compare the first with the third, the second with the forth, and so on. On the vertical axis are the different Daubechies members, where the data point represents the member with maximum number of regimes for the given day and lag. The middle figure shows the number of regimes of any size for Daubechies 6 wavelet versus different days and lags. The last figure shows the number of regimes with size > 30 for Daubechies 6 wavelet versus different days and lags. In this last figure, the solid vertical lines represent every 16th lag for each day.

The last graph on the bottom of Figure (9) depicts number of regimes with size $>$ 30 for Daubechies 6 wavelet versus different days and lags. Here the picture is different. What is observed in all of the days is that the number of regimes initially increase with the increase number of lags, then reaches a maximum and finally decreases. This makes intuitive sense, because when the number of lags is small and the books that are compared

are close to each other, the changes are very small. In such cases the wavelet may not be able to pick up gradual changes in the limit order book, resulting in fewer but larger regimes. However, as the lag between the books increases even the gradual changes are picked up and the larger regimes are broken down in several smaller ones. Thus, at a given lag there is a maximum. Continuing to increase the lags breaks the regimes more and more and the number of the regimes of any size increase, however their size decreases and the number of regimes with size > 30 falls down. The smallest lag with maximum number of regimes > 30 is 2 in day five (data point 202), the largest lag with maximum number regimes is 45 in day 19 (data point 945). The average lag with maximum regimes (with minimum size of 31) is 16.13 with standard deviation of 11.22 lags. The solid vertical lines in the graph represent every 16th lag for each day.

5.2.2 Choice of Lags

In order to choose the optimal number of lags there are two approaches that we have explored. The first is to find the number of lags that produces the largest number of regimes with size > 30. This is discussed above and on average 16 lags result in the largest number of regimes > 30.

The second approach is to measure the maximum vertical distance between two books for each 100 shares (minimum lot size) and then to sum up for the first 1024 (100 share) lots of the overlapping portion between the two books. We denote this as the width of the information regime. After we measure the width we look for the lag that gives the minimum width. By finding the minimum width we can ensure that the books in a given regime have the same or very similar shape and that any significant change is adequately captured and correctly identified. The width measured in dollars and the average can be obtained by dividing the resulting number by 1024. However, it has to be noted that most of the changes occur in the region of the best quotes and thus most of the sum is due to this part of the book.

The results are summarized by Figure (10). The top graph presents the average width for the regimes on the ask side versus the different days and lags. The second graph shows the same for the bid side. The third graph (from the top down) depicts the number of regimes, for both the ask and the bid, with size > 30 for Daubechies 6 wavelet versus different days and lags. The last graph shows both the ask and the bid widths versus days and lags.

From the top two graphs it can be seen that in general the average regime width decreases with the increase of the lags and that the curves are U shaped. On average the

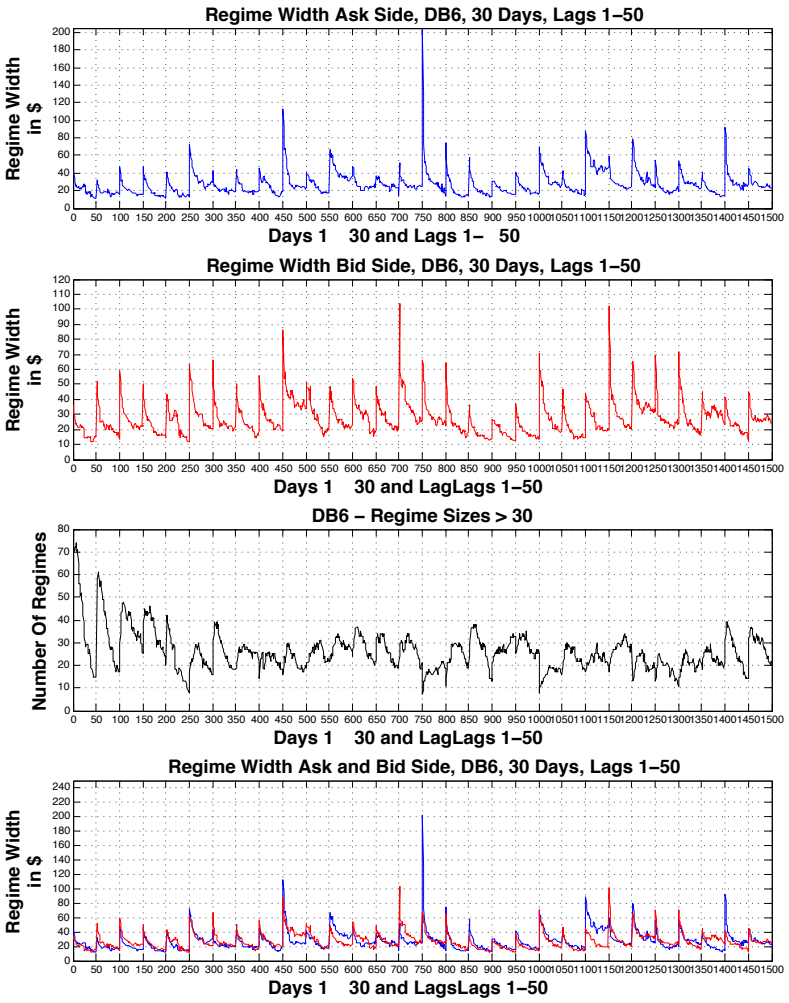

Figure 10: **Average Regime Width Vs Number Of Lags**

The top graph presents the average width for the regimes on the ask side versus the different days and lags. The second graph from the top shows the width of the regimes on the bid side versus the different days and lags. The third graph from the top depicts number of regimes, for both the ask and the bid, with size > 30 for Daubechies 6 wavelet versus different days and lags. The graph on the bottom shows both the ask and the bid width versus days and lags.

minimum regime width for the ask side is at 41 lags with standard deviation of 7.86. For the bid side the average is 44.5 with standard deviation of 5.94 lags. The explanation of the shape is similar to above. When the number of lags is small and the books that are compared are close to each other, the changes are very small. Again, in such case the wavelet may not be able to pick up gradual change and the resulting regimes have larger width. However, as the lag between the books increase even the more gradual changes are picked up and the larger regimes are broken down into several smaller ones which have smaller width and at a given number of lags the width is minimized. As the number

of lags further increases the number of regimes with size > 30 decreases. At the same time the regimes with size > 30 that are left have larger width, because with more lags the changes become more and more pronounced meaning that the shape of the books are more different.

When we look at the third plot from top down and compared to the previous two, what we can see is that the lags that lead to the maximum number of regimes and the lags that produce the minimum width are different. This is also represented by the mean values: the average number of lags with maximum regimes is 16 versus the average number of lags with the minimum regime width for the ask side which is 41 lags and 44 lags for the bid side respectively. What this means is that there is a trade-off between the two approaches. The researcher has to compromise between number of regimes and average width of the regimes.

We noted earlier that if the books are far apart in time from each other this can cause additional noise and potentially identify regime switches at points that actual the regime does not change. The empirical application of these regimes diminishes, because by the time one regime is identified it can be over since the regime sizes decrease with the increase of the lag. With that said using too many lags may not be reasonable.

6 Empirical Results

Using the procedure described in the methodology section for identifying regimes, we present two sets of results depending on the number of lags chosen. The first one is for 16 lags based on maximum number of regimes criteria and, the second is for 43 lags $((41 + 44)/2)$ based on minimized average regime width.

6.1 Results with 16 Lags

When we use 16 lags, the results are presented in Tables (1) and (2). In this case we identify on average 30.13 regimes per day. Some days have more activity and more data observations. For example, the first day has almost twice the data observations than other days, that is why it has the most regimes per day - 51. The minimum number of regimes is 16 in the 21^{st} day with standard deviation of 7.74.

The regimes are built in limit order book time and have different amount of books in them. The minimum is set at 31 books per regime, the maximum size of regime is 763 books in day 24. On average the regimes size is 94.24 books with standard deviation of 25.19.

In clock time, the regimes have different durations (in seconds). For example, if there are are two regimes with the same number of books, their duration in clock time not need be the same. It depends on the frequency at which event occur. The shortest duration for a regime is 61 seconds in day 2, the longest is 7415 seconds (approximately 2 hours) in day 25. On average, the regime duration is 751.78 seconds with standard deviation of 283.61.

The distribution of market orders throughout the day and throughout the regime is not uniform. There are some regimes that do not have any market orders and trades. The maximum number of market orders is 54 in day 9, the average is 5.10 market orders per regime, with standard deviation of 1.32. All these results are summarized in Table (1).

		Number Of Books In Regime				Number Of Seconds In Regime				Number Of Market Orders In Regime			
Day	Number Of Regimes Per Day	Min	Max	Average	StD	Min	Max	Average	StD	Min	Max	Average	StD
1	51	32	172	60.22	33.06	84	989	255.67	187.81	0	21	4.76	3.84
2	45	32	216	66.18	38.74	61	1197	306.89	215.50	0	16	4.58	4.04
3	40	31	275	81.50	57.30	138	1861	492.18	393.04	0	25	5.48	5.89
4	45	31	181	71.14	37.59	69	1404	398.42	314.90	0	14	4.02	3.49
5	29	31	164	62.70	38.47	94	2049	453.55	440.60	0	21	4.34	5.18
6	28	34	296	83.14	57.59	135	3201	738.04	671.09	0	28	5.86	6.33
7	32	31	230	72.70	46.79	130	2676	595.78	553.83	0	21	5.63	5.09
8	28	32	340	94.81	62.96	182	2382	800.75	568.39	0	16	4.50	3.28
9	24	34	582	112.13	116.90	116	5213	927.92	1135.07	0	54	8.17	11.50
10	27	31	414	117.69	116.81	70	4578	912.19	1268.69	0	30	7.44	8.68
11	26	31	513	111.42	102.75	89	3355	932.46	859.17	0	25	5.77	7.27
12	32	31	484	100.56	92.53	101	4727	745.72	957.86	0	28	4.97	6.16
13	33	32	262	82.12	52.60	72	3040	616.09	597.50	0	21	4.52	3.87
14	33	32	280	83.22	67.09	112	4274	676.58	908.82	0	17	3.33	3.84
15	31	32	381	83.97	64.68	147	2928	648.90	579.08	0	22	4.55	4.62
16	17	43	317	121.53	81.96	203	4720	1340.24	1139.31	1	17	6.76	4.40
17	28	31	288	94.89	62.43	98	2302	815.11	611.51	0	21	4.64	4.74
18	38	31	231	64.78	40.62	70	1624	486.13	335.10	0	19	4.39	3.75
19	25	38	609	148.83	121.97	90	3466	925.60	787.05	0	30	6.44	6.33
20	34	31	505	103.09	96.22	129	3723	665.56	751.81	0	22	3.82	4.73
21	16	31	435	148.81	114.19	251	4700	1625.00	1442.83	0	20	7.06	6.05
22	28	32	383	96.64	76.44	133	4017	769.36	845.60	0	16	4.96	4.44
23	24	31	363	107.83	89.57	140	2900	904.21	749.48	0	22	6.38	5.99
24	27	31	763	138.52	168.65	139	6669	945.85	1323.66	0	35	6.67	9.03
25	25	31	715	124.83	141.32	95	7415	975.96	1481.62	0	27	5.84	6.79
26	21	35	255	76.05	46.09	215	2983	904.00	627.14	1	10	4.14	2.57
27	27	31	260	85.08	59.77	114	2795	785.59	790.85	0	11	3.81	3.32
28	27	33	331	87.27	81.13	142	3351	798.56	791.60	0	15	3.07	3.62
29	32	33	116	57.42	22.01	71	1514	454.53	370.93	0	9	2.81	2.53
30	31	31	279	88.19	65.49	82	3160	656.52	682.61	0	17	4.39	4.29
Average	30.13	32.33	354.67	94.24	75.12	119.07	3307.10	751.78	746.08	0.07	21.67	5.10	5.19
StD	7.74	2.55	160.92	25.19	35.03	46.39	1522.63	283.61	347.70	0.25	8.67	1.32	2.01
Min	16.00	31.00	116.00	57.42	22.01	61.00	989.00	255.67	187.81	0.00	9.00	2.81	2.53
Max	51.00	43.00	763.00	148.83	168.65	251.00	7415.00	1625.00	1481.62	1.00	54.00	8.17	11.50

Table 1: **Regimes Summary** This table presents the summarized results for number of books, seconds and market orders per regime for Daubechies 6 wavelet and 16 lags.

The distribution of the regimes, the transitions and their sizes is not uniform. To get an idea of how they are spanned during the day, observe the blue lines in Figure (11). The number of transition periods is the same as the number of regimes minus one. Note that we have not set a minimum size for the transition periods as we did for the regimes. For the transition periods, the minimum number of books is 1, the maximum is 638 in day 5. On average there are 40.63 books in transition period with standard deviation of 16.59. The average number of books per regime is obviously higher than the transition periods due to the imposed restriction on minimum size.

In clock time transitions have different durations. The shortest duration for a transition period is 1 second in multiple days, the longest is 3267 seconds (approximately 55 minutes) in day 5. On average, the transition duration is 257.05 seconds (more or less 4 minutes) with standard deviation of 86.65.

The distribution of market orders throughout the transition periods is also not uniform. Number of transitions do not have any market orders and trades. The maximum number of market orders in transition periods is 55 in day 1, the average is 2.08 market orders per transition, with standard deviation of 1.35. The results are presented in Table (2).

		Number Of Books In Transition				Number Of Seconds In Transition				Number Of Market Orders In Transition			
Day	Number Of Transitions Per Day	Min	Max	Average	StD	Min	Max	Average	StD	Min	Max	Average	StD
1	50	1	508	84.76	97.53	1	1798	309.92	369.13	0	55	6.66	9.46
2	44	1	584	77.86	109.86	1	1929	325.25	388.04	0	34	4.73	6.50
3	39	2	219	38.26	48.79	2	1365	194.44	265.62	0	9	1.97	2.73
4	44	1	219	48.50	54.23	1	846	221.75	218.43	0	13	2.32	3.21
5	28	1	638	75.71	128.77	1	3267	469.29	688.73	0	42	5.89	8.69
6	27	1	226	40.74	50.32	1	1497	294.63	400.97	0	15	2.52	3.85
7	31	1	160	44.45	44.09	1	1024	283.84	270.12	0	8	2.35	2.33
8	27	1	111	27.70	27.02	1	828	201.41	207.58	0	5	0.81	1.30
9	23	1	207	38.48	48.20	1	1832	272.39	394.19	0	7	1.78	2.00
10	26	1	189	37.96	44.13	1	989	163.92	209.47	0	10	1.65	2.30
11	25	1	93	28.04	27.12	1	920	188.52	222.18	0	5	1.36	1.47
12	31	1	197	31.39	43.55	1	917	154.19	199.82	0	14	1.39	2.73
13	32	4	287	44.97	59.63	29	1792	256.72	334.14	0	18	2.16	3.50
14	32	1	438	39.00	76.07	1	1864	187.81	325.39	0	20	1.47	3.51
15	30	1	258	43.80	65.25	1	1547	281.53	408.59	0	8	1.23	1.99
16	16	2	116	37.38	32.20	2	1510	341.38	384.50	0	6	1.75	1.81
17	27	1	144	25.70	32.22	1	974	206.78	257.26	0	10	1.63	2.27
18	37	1	152	32.41	36.29	1	1384	261.89	338.64	0	10	1.59	2.14
19	24	1	250	33.50	50.32	1	2513	241.38	504.44	0	11	1.58	2.54
20	33	2	190	32.00	35.51	8	943	179.18	204.39	0	6	1.39	1.73
21	15	1	57	16.93	13.94	1	416	160.73	135.47	0	3	0.60	0.91
22	27	1	119	36.59	36.22	1	1192	241.67	284.99	0	7	1.67	1.92
23	23	1	176	31.57	41.06	1	1833	285.65	409.09	0	8	1.26	2.07
24	26	1	98	23.54	19.58	1	354	124.62	93.70	0	7	1.27	1.76
25	24	1	167	32.92	45.83	1	1451	192.88	301.56	0	15	1.96	3.32
26	20	1	164	31.45	42.28	1	2890	454.15	768.19	0	15	2.30	3.50
27	26	1	117	32.73	32.32	1	1244	248.12	301.39	0	7	1.88	2.29
28	26	2	155	31.92	36.31	2	1405	257.19	345.92	0	5	1.19	1.67
29	31	1	381	75.45	99.06	1	2094	452.94	553.95	0	17	2.52	3.94
30	30	1	221	43.23	44.59	1	1108	257.50	280.46	0	11	1.53	2.32
Average	29.13	1.23	228.03	40.63	50.74	2.27	1457.53	257.05	335.55	0.00	13.37	2.08	2.99
StD	7.74	0.63	143.63	16.59	26.76	5.21	658.34	86.65	147.31	0.00	11.45	1.35	1.97
Min	15.00	1.00	57.00	16.93	13.94	1.00	354.00	124.62	93.70	0.00	3.00	0.60	0.91
Max	50.00	4.00	638.00	84.76	128.77	29.00	3267.00	469.29	768.19	0.00	55.00	6.66	9.46

Table 2: **Transitions Summary**. This table shows the results for number of books, seconds and market orders per *transition period* for Daubechies 6 wavelet and 16 lags.

Finally, in Table (3) we summarize some basic statistics about the market orders in different days, without distinction of regimes or transition periods. The maximum number of market orders per day is 597 in day one, the minimum is 117 in day 28, the average number of market orders per day is 224.80 with standard deviation of 96.58. The maximum total volume is 488,000 shares in day one, the minimum is 78,000 shares in day 16, the average volume is 253,353.33 shares per day with standard deviation of 84,396.84. The minimum size of market order is 100 and is set by the exchange, the maximum is 37,000 in day 22. The average market order size for the 30 day period is 1164.85 with standard deviation of 266.90.

Day	Number Of MO	Total Volume	Min MO	Max MO	Average MO	StD
1	597	488000	100	12500	817.42	1415.97
2	429	337100	100	10500	785.78	1302.30
3	326	284400	100	8300	872.39	1407.36
4	296	341900	100	14600	1155.07	1996.62
5	326	302500	100	15500	927.91	1783.68
6	241	255400	100	6000	1059.75	1416.51
7	262	373000	100	10800	1423.66	1579.70
8	160	163200	100	10000	1020.00	1659.69
9	247	264200	100	10000	1069.64	1577.44
10	246	259800	100	10600	1056.10	1711.53
11	185	212600	100	8000	1149.19	1636.34
12	209	299200	100	10000	1431.58	1711.94
13	222	335200	100	10000	1509.91	2004.70
14	164	274000	100	30000	1670.73	2962.01
15	186	311600	100	20000	1675.27	3420.01
16	146	78000	100	5000	534.25	840.56
17	178	233900	100	10000	1314.04	1647.55
18	237	292200	100	8800	1232.91	1444.16
19	200	196800	100	12400	984.00	1499.63
20	185	217800	100	10000	1177.30	1925.97
21	125	134800	100	5500	1078.40	1319.25
22	199	290800	100	37000	1461.31	2924.14
23	186	225700	100	9000	1213.44	1455.90
24	217	330600	100	10000	1523.50	1474.08
25	197	167900	100	15000	852.28	1525.62
26	136	143500	100	5000	1055.15	1280.73
27	168	227200	100	37000	1352.38	3093.12
28	117	137500	100	7600	1175.21	1388.56
29	170	208200	100	16000	1224.71	1851.49
30	187	213600	100	15000	1142.25	2036.12
Average	224.80	253353.33	100.00	13003.33	1164.85	1776.42
StD	96.58	84396.84	0.00	8177.79	266.90	588.28
Min	117.00	78000.00	100.00	5000.00	534.25	840.56
Max	597.00	488000.00	100.00	37000.00	1675.27	3420.01

Table 3: **Market Orders (MO) Summary** The table presents summarized results for number of market orders, total volume and average volume per day without distinction of regimes or transition periods.

6.2 Results with 43 Lags

The results for this case are presented in Tables (4) and (5). In the case we identify on average 21.3 regimes per day. The average number of regimes per day, as expected, is lower than the case of 16 lags. The minimum number of regimes is 12 in the 5^{th} day with standard deviation of 4.94. the maximum size of regime is 685 books in day 25. On average the regimes size is 73.58 books with standard deviation of 19.39.

In clock time, the shortest duration for a regime is 50 seconds in day 12, the longest is 7119 seconds (2 hours approximately) in day 25. On average, the regime duration is 602.15 seconds with standard deviation of 197.87. The maximum number of market orders is 34 in day 9, the average is 4.12 market orders per regime, with standard deviation of 1.25. All these results are summarized in Table (4).

The number of transition periods is the same as the number of regimes minus one. The minimum number of books is 1, the maximum is 1779 in day 2. On average there are 114.41 books in transition period with standard deviation of 72.92. The average number of books per transition period is obviously higher than the case of 16 lags, because on average the regimes are smaller and less as total number. In clock time, the shortest duration for a transition period is 1 seconds in multiple days, the longest is 8537 seconds

		Number Of Books In Regime				Number Of Seconds In Regime				Number Of Market Orders In Regime			
Day	Number Of Regimes Per Day	Min	Max	Average	StD	Min	Max	Average	StD	Min	Max	Average	StD
1	16	31	97	49.73	17.30	98	481	213.75	93.76	1	10	3.06	2.43
2	17	32	155	67.94	33.65	55	785	317.18	193.69	0	15	5.88	4.55
3	32	31	212	56.13	35.55	51	1709	329.72	344.11	0	23	3.97	4.46
4	28	31	107	55.04	18.78	100	899	331.43	196.72	1	10	3.21	2.44
5	12	31	132	50.58	27.22	75	1833	416.25	488.11	0	20	4.25	5.59
6	19	32	232	66.79	45.50	138	2488	580.21	517.44	0	21	5.00	4.28
7	18	31	171	60.83	32.61	124	1815	582.56	360.77	1	18	5.06	4.32
8	26	31	141	67.28	33.71	142	2072	569.04	469.35	0	8	2.96	2.32
9	20	31	394	89.53	80.85	190	3175	749.05	708.18	0	34	6.50	7.37
10	18	35	339	120.18	89.14	88	4999	1015.44	1231.29	0	32	7.78	8.39
11	21	32	282	88.67	57.67	122	1859	781.95	497.33	0	15	4.67	5.02
12	29	31	246	83.00	62.27	50	2578	617.69	716.94	0	19	4.28	4.59
13	23	31	152	56.00	31.28	101	1570	412.61	378.33	1	8	3.22	2.19
14	20	31	283	76.74	67.54	131	3619	709.70	912.69	0	20	3.20	4.31
15	23	32	223	64.76	44.18	137	1802	504.43	379.56	0	16	3.13	3.76
16	21	31	200	72.57	44.61	203	1992	797.43	566.72	0	14	3.81	3.22
17	21	31	122	69.40	27.76	86	1281	578.57	353.25	0	9	3.67	2.37
18	18	31	202	60.11	42.71	116	1526	445.78	337.86	0	18	3.78	4.24
19	32	34	483	87.34	87.46	111	2916	562.38	612.04	0	24	3.56	5.10
20	24	31	273	88.33	59.46	190	2232	661.21	542.96	0	14	3.63	3.77
21	18	34	295	99.83	62.63	252	3278	1038.39	766.94	0	15	4.83	3.59
22	21	32	408	86.62	87.53	103	3915	738.19	947.10	0	20	4.71	4.76
23	22	35	135	72.00	31.17	210	2475	641.50	500.61	2	10	4.82	2.17
24	28	31	319	102.27	65.75	58	2737	667.04	602.50	0	21	4.68	4.51
25	19	31	685	119.28	149.84	170	7119	948.84	1566.70	0	27	5.89	7.41
26	13	31	136	55.31	28.67	184	1289	670.92	427.93	1	7	3.38	2.22
27	20	36	128	63.45	30.46	156	1669	624.00	429.83	0	8	3.00	2.32
28	21	31	152	60.62	32.71	137	1882	540.24	475.56	0	6	2.19	2.14
29	16	31	80	46.19	14.01	147	1197	445.69	282.53	0	6	2.38	1.86
30	23	31	263	70.86	51.95	140	3149	573.35	642.83	0	18	3.09	4.01
Average	21.30	31.80	234.90	73.58	49.80	128.83	2344.70	602.15	551.45	0.23	16.20	4.12	3.99
StD	4.94	1.45	131.01	19.39	28.56	50.51	1337.74	197.87	305.46	0.50	7.36	1.25	1.67
Min	12.00	31.00	80.00	46.19	14.01	50.00	481.00	213.75	93.76	0.00	6.00	2.19	1.86
Max	32.00	36.00	685.00	120.18	149.84	252.00	7119.00	1038.39	1566.70	2.00	34.00	7.78	8.39

Table 4: **Regimes Summary** This table presents summarized results for number of books, seconds and market orders per regime Daubechies 6 wavelet and 43 lags.

(2.37 hours) in day 2. On average, the transition duration is 771.71 seconds (13 minutes approximately) with standard deviation of 374.66. Results are presented in Table (5).

In Figure (11) we can see the distribution of regimes and transition periods for 5 different days for both 16 and 43 lags. The blue graphs are for 16 lags and the red for 43 lags. It is apparent from the figure that the regimes are basically at the same positions for both lags. However, for 43 lags the regimes are shorter. In fact some of them are shorter than 31 books and, that is why they now are considered transition periods and that is also why there are regimes pictured for the 16 lags without corresponding regimes with 43 lags, but not the reverse. From this figure we can clearly see that the regimes identified using 43 lags is simply a sub sample of the ones found using 16 lags.

For the rest of the days the figures are very similar and available upon request.

6.3 Limit Order Book Correlations

Once the regimes are determined by the above procedure we could further examine the following correlations: correlation between the cumulative volume on the bid and the ask side, and the correlation between the best quotes and the cumulative volume. We define high correlation of being $> |0.5|$ and low correlation as $< |0.5|$. If the correlation is high

		Number Of Books In Transition				Number Of Seconds In Transition				Number Of Market Orders In Transition			
Day	Number Of Transitions Per Day	Min	Max	Average	StD	Min	Max	Average	StD	Min	Max	Average	StD
1	15	4	1327	325.13	369.90	38	4581	1414.53	1508.40	1	117	24.20	31.80
2	16	2	1779	313.94	463.97	6	8537	1372.50	2144.08	0	90	19.13	25.05
3	31	1	308	78.55	78.18	1	2802	445.26	523.58	0	18	4.71	4.10
4	27	1	915	136.22	202.42	1	3263	686.85	835.30	0	44	6.93	10.06
5	11	5	1102	272.45	331.07	21	7610	1803.82	2295.40	0	67	19.91	24.19
6	18	2	247	95.00	66.87	3	2467	868.83	690.26	0	22	6.44	5.50
7	17	1	1198	142.76	280.28	1	7630	1016.24	1776.46	0	69	9.06	15.89
8	25	2	163	62.00	49.00	4	1778	519.16	483.89	0	9	2.76	2.65
9	19	2	562	90.21	133.59	4	4204	675.47	978.54	0	26	5.37	6.71
10	17	1	580	111.47	147.78	1	1962	578.47	600.20	0	27	5.76	8.43
11	20	2	444	76.70	99.77	4	3142	578.35	702.96	0	22	3.90	4.87
12	28	1	240	65.25	61.19	1	1383	359.82	355.43	0	16	2.54	3.45
13	22	1	411	105.55	124.15	1	2966	653.32	840.61	0	30	5.73	7.81
14	19	1	1121	122.89	251.14	1	4848	726.84	1096.24	0	42	4.68	10.04
15	22	1	410	102.32	112.16	1	2945	727.00	718.44	0	19	4.59	5.11
16	20	2	199	51.70	59.59	2	1589	529.35	513.77	0	11	2.90	3.32
17	20	1	353	83.35	96.47	1	2664	741.35	833.29	0	15	4.20	4.93
18	17	6	527	143.12	136.60	22	5564	1159.41	1430.45	1	37	8.76	8.44
19	31	1	179	50.16	50.16	1	1820	344.00	418.94	0	15	2.58	3.39
20	23	1	213	80.17	59.89	1	1558	459.78	358.61	0	12	2.96	2.82
21	17	1	119	47.82	42.36	1	2155	545.18	580.32	0	5	2.00	1.97
22	20	1	253	71.60	70.38	1	2128	478.00	539.23	0	11	3.15	3.20
23	21	1	532	67.71	114.36	1	4250	555.90	920.54	0	26	3.00	5.87
24	27	1	302	59.52	69.72	1	1618	370.11	392.42	0	13	2.96	3.35
25	18	1	629	79.22	147.02	1	2466	537.56	784.83	0	26	4.11	6.29
26	12	7	386	108.00	109.38	92	6823	1489.25	2005.80	0	31	6.50	8.76
27	19	1	399	96.68	108.77	1	2394	808.63	755.29	0	18	4.89	5.63
28	20	1	551	78.65	126.21	1	5925	750.45	1345.54	0	19	2.70	4.64
29	15	30	941	210.07	240.95	369	3408	1310.13	918.34	0	36	8.13	8.68
30	22	2	447	104.09	106.70	2	2234	645.82	585.16	0	27	4.91	7.26
Average	20.30	2.80	561.23	114.41	143.67	19.50	3557.13	771.71	931.08	0.07	30.67	6.32	8.14
StD	4.94	5.37	403.03	72.92	103.81	68.41	2022.93	374.66	540.71	0.25	25.08	5.39	7.11
Min	11.00	1.00	119.00	47.82	42.36	1.00	1383.00	344.00	355.43	0.00	5.00	2.00	1.97
Max	31.00	30.00	1779.00	325.13	463.97	369.00	8537.00	1803.82	2295.40	1.00	117.00	24.20	31.80

Table 5: **Transitions Summary** The table presents summarized results for number of books, seconds and market orders per *transition period* Daubechies 6 wavelet and 43 lags.

and negative coupled with high (positive on bid side and respectively negative on the ask side) correlation between best quotes and cumulative volume, and a small spread, it can be argued that one side is reacting to the action of the other and all the information is related through market orders.

Based on the different correlations, we propose that regimes can classified in at least six categories:

- High negative correlation between cumulative volumes coupled with:
  - high correlations between best quotes and cumulative volumes.
  - low or opposite sign correlations between best quotes and cumulative volumes.
- Low negative correlation between cumulative volumes coupled with:
  - high correlations between best quotes and cumulative volumes.
  - low or opposite sign correlations between best quotes and cumulative volumes.
- Positive correlation between cumulative volumes coupled with:

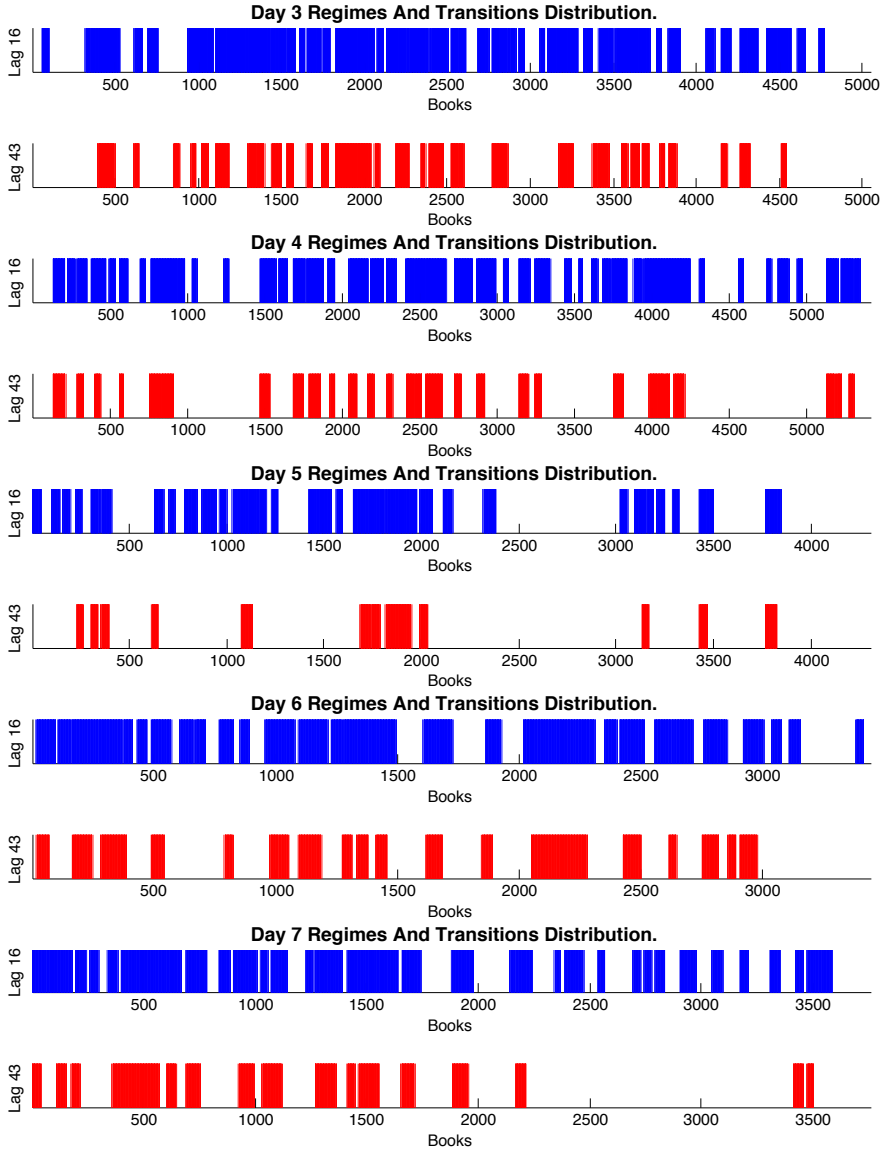

Figure 11: **Regimes Versus Transitions for 2 Different Lags**

The figure shows the distribution of regimes and transitions for five different days (day 3 to day 7). The graphs in blue are for lag of 16 and the graph right below them are for the same day for lag 43.

– high correlations between best quotes and cumulative volumes.

– low or opposite sign correlations between best quotes and cumulative volumes.

The number of regimes that fall in each of the six categories per day is presented in Table (6).

We have to note that the table represents only the cases that fall in one of the categories previously described. For example, in the category high negative correlation between

Day	High Negative Correlation Between Cumulative Volumes		Low Negative Correlation Between Cumulative Volumes		Positive Correlation Between Cumulative Volumes	
	High Correlations Between Best Quotes And Cumulative Volumes	Low or Opposite Sign Correlations Between Best Quotes And Cumulative Volumes	High Correlations Between Best Quotes And Cumulative Volumes	Low or Opposite Sign Correlations Between Best Quotes And Cumulative Volumes	High Correlations Between Best Quotes And Cumulative Volumes	Low or Opposite Sign Correlations Between Best Quotes And Cumulative Volumes
1	14	1	8	1	2	3
2	19	1	10	1	0	1
3	9	0	9	1	2	0
4	13	0	12	2	2	3
5	10	1	2	3	1	3
6	12	1	5	1	3	0
7	12	0	10	1	1	0
8	5	0	10	0	1	2
9	7	0	6	3	2	3
10	7	1	2	1	4	1
11	8	1	3	1	2	2
12	7	0	6	0	1	2
13	6	0	4	4	4	0
14	7	0	8	1	0	2
15	9	0	9	2	1	0
16	7	0	1	1	0	1
17	6	1	5	0	2	2
18	6	0	4	3	1	2
19	8	1	8	0	0	2
20	14	0	6	1	1	2
21	4	0	2	1	0	0
22	9	0	3	2	3	0
23	3	1	6	1	0	0
24	9	0	5	0	0	2
25	7	0	3	1	0	0
26	1	0	7	1	1	2
27	3	1	6	1	1	0
28	5	0	7	2	1	1
29	10	0	1	3	0	1
30	9	0	4	3	4	1
Sum	**246.00**	**10.00**	**172.00**	**42.00**	**40.00**	**38.00**
Mean	**8.20**	**0.33**	**5.73**	**1.40**	**1.33**	**1.27**
Min	**1.00**	**0.00**	**1.00**	**0.00**	**0.00**	**0.00**
Max	**19.00**	**1.00**	**12.00**	**4.00**	**4.00**	**3.00**
StD	**3.77**	**0.48**	**2.96**	**1.07**	**1.27**	**1.08**
% Of All Regimes	**27.21%**	**1.11%**	**19.03%**	**4.65%**	**4.42%**	**4.20%**
% Of Classified Regimes	**44.89%**	**1.82%**	**31.39%**	**7.66%**	**7.30%**	**6.93%**

Table 6: **Regime Correlation Categories Summary** This table presents The number of regimes that fall in each of the six correlation categories per day. It can be observed that the category with highest number of regimes is the one with high correlations between best quotes and cumulative volumes.

cumulative volumes coupled with low or opposite sign correlations between best quotes and cumulative volumes, it means that on both bid and ask side there are low or opposite sign correlations between best quotes and cumulative volumes. It does not include the mixed case where on one side there is low or opposite sign correlations between best quotes and cumulative volumes and on the other there is high correlations between best quotes and cumulative volumes. Detailed examination and classification of the regimes according to the behavior of the different variables and their correlations is a subject of a different future study.

However, there is a specific case which is of a particular interest: the case when the correlation between the best quotes and the cumulative volume is high (close to -1 on the ask side and 1 on the bid side) and the spread is small. It is important to note that this case appears 27.21% of the times (or 44.89% if we just consider the six categories mentioned before). In this case the relationship between the mid quotes (as approximation to the value) and the cumulative volume can be examined under the theoretical model of Lehmann (2008). According to Lehmann (2008), if the values are strictly increasing in quantities, this will be confirmation that it is in fact an information regime. This is equivalent to empirically test:

$$V_t(q) \quad = \quad V_{t-1}(q + Q_{t-1})$$

Since the equation has no error term the assumption is that coefficient of correlation should be 1 or -1. However, in reality this will not hold, because not all of Lehmann's assumptions hold (for example, holes in the books), the value is approximated by the mid quote, there is noise in the quotes and, the spreads are not always small. Nonetheless, if the coefficient is close to 1 or -1, this would suggests that indeed the cluster of books belong to an information regime. With that said we do not mean that the rest of the cases do not support Lehmann's theory. However they are not appropriate for that type of test because the mid quotes may not be a good approximation for the underlying value.

An additional point. As it can be seen from the summary results presented in Table (1), there are some regimes with very few market orders or even with none. In such cases instead of market orders we use cancelations in the best quotes. The impact on the book would be pretty much the same for market orders and cancelations as long as the correlations between the the cumulative volumes and between the cumulative volume and the best quotes are high. That would insure that the cancelations are at the best quote region.

In the case with 16 lags, there are 246 (out of 904 regimes detected for the 30 days, which makes 27.21 % of all cases) regimes that can be characterized with high correlations

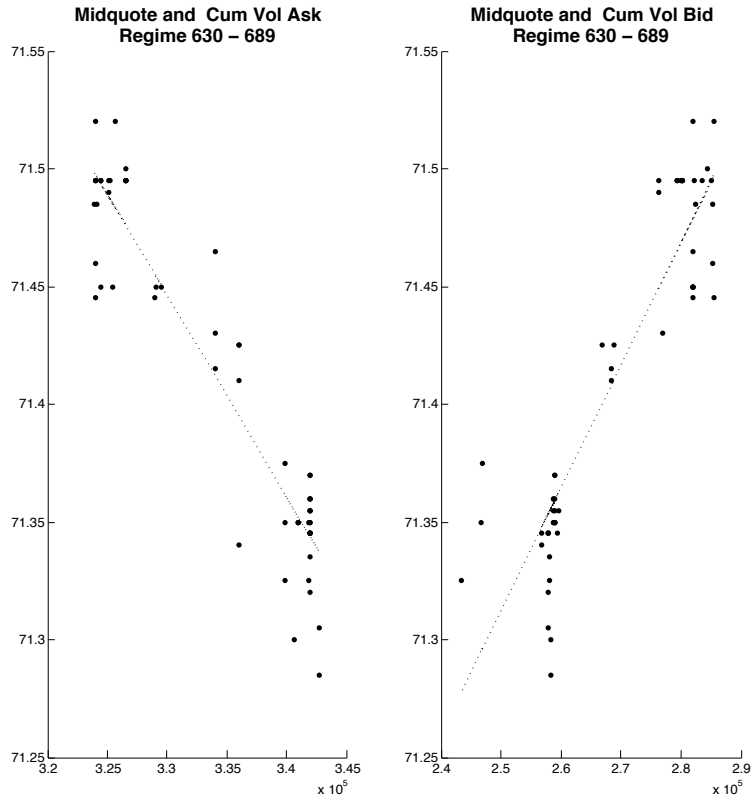

Figure 12: **Correlation Between Value and Depth**

This figure shows the relation between midquote and cumulative volume for regime 630 - 689, day 15 - August 20, 1999. On the right side - midquote vs cumulative ask volume, also shown is fitted linear trend. On the left side - midquote vs cumulative bid volume and fitted linear trend.

as described above. The example that we present is from day 15 - August 20, and it is the 5^{th} regime from book 630 to book 689. Figure (12) shows the results along with a trend line. The actual coefficient of correlation between the ask cumulative volume and the midquote is -0.9425 and the coefficient of correlation between the bid cumulative volume and the midquote is 0.9137.

Figure (13) shows the histograms for the correlations between the value (approximated by the mid quote) and the cumulative volume. From these histograms it is apparent that the distributions of the correlations are centered around the high correlation numbers (in absolute values). One can see that 98.37% of the correlations on the ask side are $> |0.5|$ (78.46 % $> |0.7|$) and 98.78% of the correlations on the bid side are $> |0.5|$ (86.59% $> |0.7|$). The actual coefficients measured are very close to what Lehmann's theory predicts. Based on these results, we provide some empirical validation that supports Lehmann's theory.

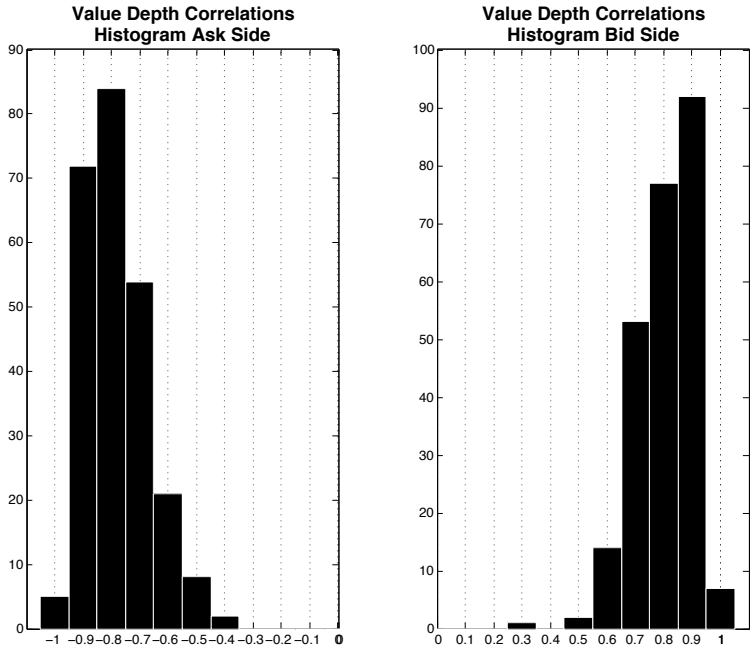

Figure 13: **Correlation Histogram**

The left figure shows the histogram for the correlations between the value and the depth on the ask side. The right figure shows the histogram for the correlations between the value (approximated by the midquote) and the depth on the bid side.

7 Conclusion

This article develops and implements a new methodology for identifying intra-day information regimes in limit order books. With the sole exception of Lehmann (2008), which is cited and used in this study, previous theoretical analysis do not account for the existence of information regimes and, to the best of our knowledge, no previous empirical work has addressed these topic. Our empirical findings are important since they imply that current market microstructure theory needs to be reconciled with the existence of these intra-daily information regimes.

We use wavelet theory and we have developed a methodology that allowed us to clearly identify information regimes. Our results show that information regimes have an impact on price formation and price discovery, including dynamic issues such as the process by which prices come to capture information over time. The discovery and identification of information regimes essentially uncovers the mechanism by which latent demands are translated into realized prices and volumes.

Our results shows that the best wavelet to analyze the limit order book data (at least for the data at hand) is the Daubechies 6 wavelet. Moreover, we show two ways to

determine the lag used in regime identification.[11] The first proposed way is by finding the wavelet that identifies the maximum number of regimes and, the second one is achieved by minimizing the regime width by changing the number of lags used. It is important to note that regimes identified with number of lags larger than the one producing the maximum number of regimes (16 in our case), are simply a sub-sample of the later.

Finally, our results empirically support Lehmann's theoretical model when the correlation between the best quotes and the cumulative volume is high (close to -1 on the ask side and 1 on the bid side) and the spread is small. In this case we were able to show that 98.37% of the correlations on the ask side are $> |0.5|$ (78.46 % $> |0.7|$) and 98.78% of the correlations on the bid side are $> |0.5|$ (86.59% $> |0.7|$). The actual coefficients are very close to what Lehmann's theory predicts (-1 or 1).

The identification of information regimes opens the door to numerous exciting research opportunities. Once the time series are divided into regimes and each one is examined separately, the outcomes of many well established studies might be revisited.

8 Technical Appendix

In this appendix we present the mathematical tools used to identify the information regimes theoretically proposed in Lehmann (2008). We describe in detail the continuous wavelet transform (CWT) an the discrete wavelet transform (DWT).

8.1 Wavelet Theory

8.1.1 Continuous Wavelet Transform

The wavelet, as the name may suggest, is localized wave form and the wavelet function $\varphi(t)$ satisfies certain mathematical criteria. In the case of non stationary data, which may contain transient phenomena (i.e. aperiodic), the wavelet basis functions are precisely localized in time and frequency. The wavelet transform provides efficient and complete representation of the signal. The wavelet is manipulated trough process of translation and dilation (or scaling). For the description of both continuous and discrete wavelet transform we follow the discussion in Los (2003) and Addison (2002) . The graphs illustrating wavelets are from Addison (2002).

[11] Recall that in this paper we use the word lag to express the distance between books that we compare. For example, lag 1 implies that we are comparing the first book with the second, the second with the third and so on. A lag of 10 implies that we compare the fist book with the eleventh, the second with the twelfth, and so on.

There are many different wavelets that on can choose from. The best one for a given application depends on the nature of the signal and on the requirements of the study. Later we discus the choice of the specific wavelet. Figure (3) shows 10 of the members of the Daubechies wavelet family.

As mentioned in order to be classified as a wavelet, a function must satisfy certain criteria:

- Wavelet must have finite energy. Where the energy of a function is the second uncentered moment:

$$E = E[\varphi(t)^2] = \int_{-\infty}^{\infty} |\varphi(t)|^2 dt < \infty \tag{8.1}$$

- if $\hat{\varphi}(f)$ is the Fourier transform [12]of $\varphi(t)$:

$$\hat{\varphi}(f) = \int_{-\infty}^{\infty} \varphi(t) e^{-jwt} dt \tag{8.2}$$

where $w = 2\pi f = 2\pi/T$. Then the following condition must hold:

[12]A periodic variable can be represented by two trigonometric forms and one complex exponential form of the Fourier series in the following way:

$$\begin{aligned} x(t) &= \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nw_0 t + b_n \sin nw_0 t) \\ &= C_0 + \sum_{n=1}^{\infty} C_n \cos(nw_0 t + \theta_n) \\ &= \sum_{n=-\infty}^{\infty} c_n e^{jnw_0 t} \end{aligned}$$

The second equation is called harmonics form. a_n and b_n are the Cartesian coefficients, C_n are the polar coefficient and c_n are the exponential coefficients. Together they are referred as Fourier resonance coefficients. $w_n = nw_0$ is the n-th harmonic of the periodic variable, n is the wave number. $w_0 = 2\pi f_0 = 2\pi/T$ is the fundamental angular frequency and $f_0 = 1/T$ is the fundamental frequency, where T is the period of the variable. j is the imaginary unit $\sqrt{-1}$

The coefficients C_n are called harmonic amplitudes and the angles θ_n are called phase angle. C_n scale the amplitude of the sinusoidal waves and θ_n shift the position of the sinusoidal waves. In this way the sinusoidal bases $e^{jnw_0 t}$ is scaled and shifted in order to achieve analysis of the time series. For the proof and the computation of the coefficients see Los (2003) page 147-148. The Fourier transform of time series $x(t)$, denoted by $\mathcal{F}$ is defined by the inner product (or correlation):

$$F(w) = \mathcal{F}[x(t)] = \int_{-\infty}^{\infty} x(t) e^{jwt} dt$$

$$C_g = \int_0^\infty \frac{|\hat{\varphi}(f)|^2}{f} df < \infty \tag{8.3}$$

This implies that the wavelet has no zero frequency component.

- The wavelet function has zero average:

$$E[\varphi(t)] = \int_{-\infty}^{\infty} \varphi(t) dt = 0 \tag{8.4}$$

- A condition that must hold for complex wavelets is that the Fourier transform must both be real and vanish for negative frequencies.

Once a wavelet function is chosen, it is translated by a limited time interval b and scaled or dilated by a scale parameter a as follows:

$$\varphi(t) = \frac{1}{\sqrt{a}} \varphi(\frac{t-b}{a}) \tag{8.5}$$

As it can be seen on Figure (14) (figure 2.3 in Addison 2002) the contraction and the stretching of the wavelet is governed by the dilation parameter a and the movement of the wavelet along the time axis is controlled by the location (translation) parameter b.

Continuous wavelet transform (CWT) or continuous WT (wavelet transform)of $x(t)$ at position b and scale a is an inner product by convolution (or correlation [13]) of the time series x(t) with a wavelet function.

$$\begin{aligned} T(a,b) &= \int_{-\infty}^{\infty} x(t) \varphi^*(t) dt \\ &= \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{a}} \varphi^*(\frac{t-b}{a}) dt \end{aligned} \tag{8.6}$$

Where $\frac{1}{\sqrt{a}}$ is a weighting function and φ^* is the complex conjugate of the wavelet function.

When the form of the wavelet function matches the form of the signal of interest and they are both in phase, the convolution produces large resonance coefficients. If they are out of phase, the result is large negative coefficient ($T(a,b)$). If the form of the wavelet

[13] See Los 2003 for the equivalence between time convolution and covariance.

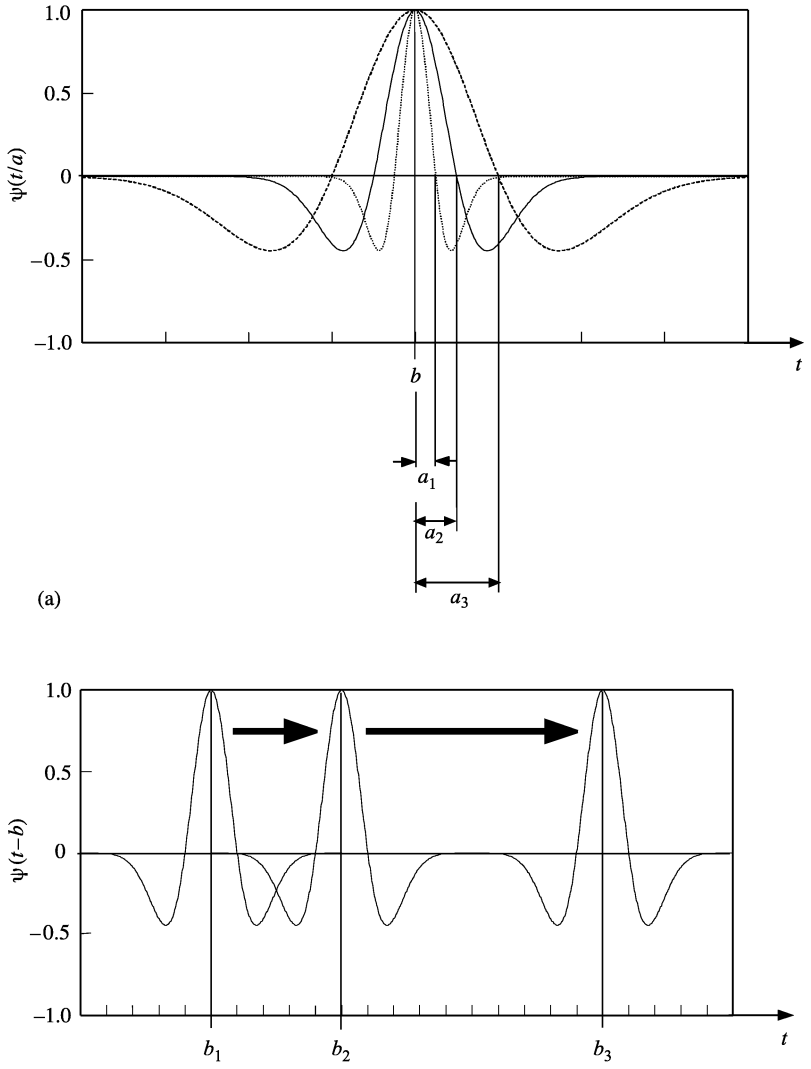

Figure 14: **Dilation and Translation**

Figure shows: (a) Dilation - contraction and the stretching of the wavelet, (b) Translation - movement of the wavelet along the time axis

does not match the shape of the signal, then the coefficients are getting close to zero. They can be again positive and negative, depending on if the wavelet and the signal are in phase. Next Figure (15) (figure 2.5 in Addison 2002) demonstrates the idea of the wavelet transform. In (a) it depicts a wavelet of scale a and position b over a signal. In the segments where the wavelet and the signal are both positive or negative, result into a positive contribution to the integral - interval A and B. In the intervals where the signal and the wavelet have opposite signs the result is negative contribution to the integral.

Figure (15) (b) a wavelet with a fixed dilation is positioned at four places on the signal. At b_1 the positive and negative parts of both the wavelet and the signal are reasonably coinciding, which would result in relatively big value of the wavelet coefficients. At b_2

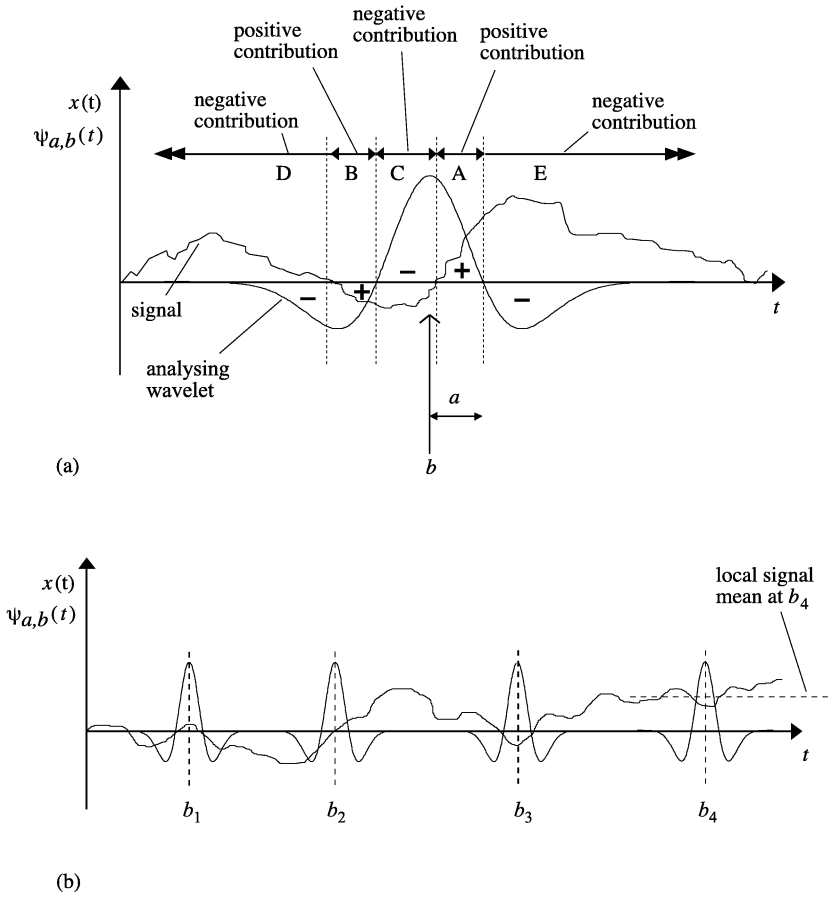

Figure 15: **Wavelet convolution with a signal**

Figure shows: (a) Wavelet with specific position and dilation over the signal - intervals A and B the wavelet and the signal are both positive or negative, which results into a positive contribution to the integral. In the rest intervals the signal and the wavelet have opposite signs and the result is negative contribution to the integral. (b) Wavelet with specific dilation at four locations on the signal - At b_1 the positive and negative parts of the wavelet and the signal are reasonably coinciding, which results in big value of the wavelet coefficients. At b_2 the positive and negative contributions cancel out and the wavelet coefficient will be close to zero. At b_3 and b_4 the wavelet and the signal are out of phase, which result in large negative wavelet coefficient.

the positive and negative contributions cancel out and as a result the wavelet resonance coefficient will be value relatively close to zero. At b_3 and b_4 the wavelet and the signal are out of phase, which will result in relatively large negative wavelet coefficient.

It can be seen that the local features of the series are highlighted by the wavelet. The mean component contributes equal positive and negative values as result it is not brought out, but rather disregarded. Through such process the wavelet picks out the features that coincide at various scales. By moving the wavelet along the signal the relevant coinciding features of the signal at the given scale are identified. This process is repeated for various scales, until all the features that coincide with the wavelet form are investigated.

The CWT has the following useful properties:

- It is linear: $T(a,b)[\gamma_1 x_1(t) + \gamma_2 x_2(t)] = \gamma_1 T(a,b)[x_1(t)] + \gamma_2 T(a,b)[x_2(t)]$.
- It is invariant under translation $T(a,b) = T(a, b - b_0)$.
- It is invariant under dilation $T(a,b) = (1/k)T(ka, kb)$, with $k = 1/\sqrt{a}$.
- It is localized in time and frequency.

A few words must be said about the relationship of frequency and scale. In the time-frequency dimension, the wavelet function can be represented by a Heisenberg box centered by certain scale (a) and position (b). When the scale a varies, the width and the hight of the box changes but its area remains the same.

The wavelet can be localize itself in time for short durations but for high frequencies. Localization in time is associated with spreading of the frequency distribution. On the reverse the more localized as frequency the wavelet is, the more spread there is in time. This is illustrated by Figure (16) (figure 2.29 in Addison 2002). It can be seen on the figure that as the wavelet contracts in time it is made of higher frequencies with a wider spread [14].

8.1.2 Discrete Wavelet Transform

Subsampling of the CWT constructs full representation of time series (or any other signal) if such series can be reconstructed from the discrete families of the wavelet functions. The value of such process is that instead of working with the whole wavelet function, one can just use a handful of coefficients describing the wavelet, which greatly simplifies the calculations. Certain conditions must be met met by the wavelets in order for them to provide stable and complete representation and reconstruction of the time series. These conditions are provided by the Frame theory.

Frame is a family of vectors which can represent any time series with finite risk by sequence of its inner products with the vectors of the family. Wavelet frames are constructed by discretely sampling the time and scale parameters of the of the CWT. A frame can be constructed by a family of wavelet functions such that the energy of the resulting wavelet coefficients lies in within a certain bounded range of the energy of the original signal. If the bounds of the frame are the same, the frame is said to be tight. Tight frames have simple reconstruction formulas. If the bounds are equal teach other but bigger than 1 the

[14] See Los (2003) and Addison (2002) for further discussion

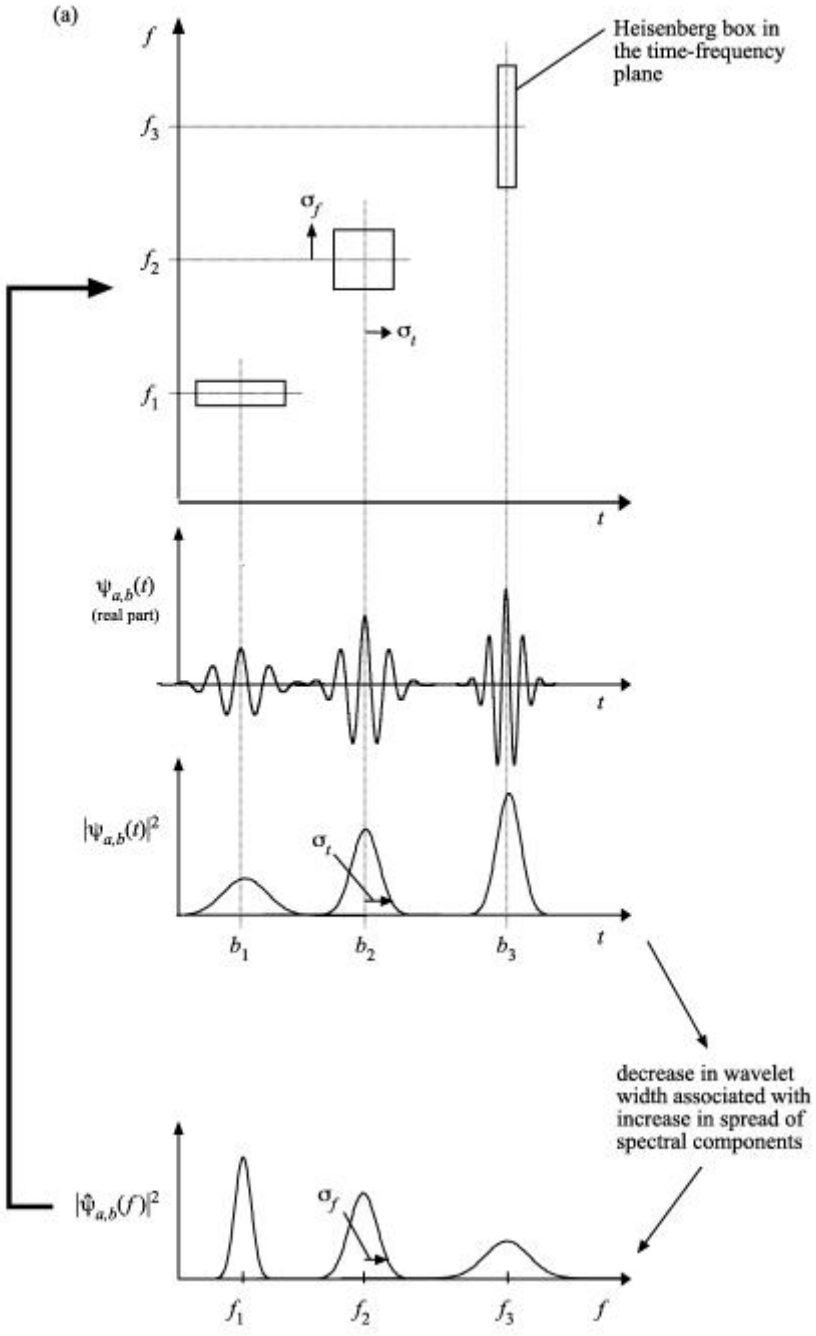

Figure 16: **Frequency and scale relationship**

Figure shows in (a) Heisenberg boxes in time-frequency plane for wavelet at various scales. Heisenberg box is centered by certain scale (a) and position (b). When the scale a varies, the width and the hight of the box changes but its area remains the same. The wavelet can be localized in time for short durations but high frequencies. On the reverse the more localized as frequency the wavelet is, the more spread there is in time.

frame is considered redundant. If the bounds are equal to each other and equal to 1, than the avert family denied by this frame forms orthonomal basis.

One way to sample the parameters a and b is by logarithmic discretization of the scale a and connect this to the size of the steps taken between b locations. To achieve this one moves in discrete steps to each location b which are proportional to the scale. Thus:

$$\varphi(t)_{m,n} = \frac{1}{\sqrt{a_0^m}}\varphi(\frac{t - nb_0a_0^m}{a_0^m}) \tag{8.7}$$

Where m and n control the wavelet position and scale. A natural way to choose the discrete wavelet parameters a and b is 2 an 1 respectively. This is know as dyadic grid arrangement and is the simplest and most efficient discretization for practical purposes. If we substitute in the wavelet equation shown above we get:

$$\varphi(t)_{m,n} = \frac{1}{\sqrt{2^m}}\varphi(\frac{t - n2^m}{2^m}) \tag{8.8}$$

Using that equation the discrete wavelet transform (DWT) can be written as :

$$T_{n,m} = \int_{-\infty}^{\infty} x(t)\varphi_{n,m}(t)dt \tag{8.9}$$

The $T_{n,m}$ are known as wavelet or detail coefficients. By choosing an orthonomal or orthogonal wavelet basis $\varphi_{n,m}$ [15]. The original signal can be reconstructed in terms of the wavelet (detail) coefficients $T_{n,m}$ by using inverse wavelet transform :

$$x(t) = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} T_{n,m}\varphi_{n,m}(t) \tag{8.10}$$

Mallat showed that one can completely decompose time series $x(t)$ in terms of approximations provided by scaling functions and details provided by the wavelet coefficients. The approximations are high scale, low frequency components, while the details are low scale high frequency components. The process can be iterated, with successive approximations being decomposed, so the time series is broken down in many low resolution components. The decomposition continues until individual details consist of a single observation.

The scaling function associated with the smoothing has the same form as the wavelet given by:

$$\phi(t)_{m,n} = 2^{-m/2}\phi(2^{-m/2}t - n) \tag{8.11}$$

with the property of:

[15] Los (2003) page 207 proposes orthogonal wavelet basis while Addison (2002) page 68 proposes orthonormal wavelet basis.

level index 4	$T_{4.0}$	$T_{4.1}$	$T_{4.2}$												$T_{4.15}$
level index 3	$T_{3.0}$		$T_{3.1}$		$T_{3.2}$										$T_{3.7}$
level index 2	$T_{2.0}$				$T_{2.1}$				$T_{2.2}$				$T_{2.3}$		
level index 1	$T_{1.0}$								$T_{1.1}$						
level index 0	$T_{0.0}$														
level index -1	signal mean component														

Figure 17: **Wavelet Tiling**

The figure shows the relation of the discrete wavelet transform detail coefficients to the time frequency plane.

$$\int_{-\infty}^{\infty} \phi(t)_{0,0} dt \quad = \quad 1 \tag{8.12}$$

where $\phi_{0,0}(t) = \phi(t)$ is the scaling function.

The scaling function can be convoluted with the time series to produce the scaling coefficients:

$$S_{m,n} \quad = \quad \int_{-\infty}^{\infty} x(t)\phi_{m,n}(t)dt \tag{8.13}$$

In this case the signal can be represented by:

$$x(t) \quad = \quad \sum_{n=-\infty}^{\infty} S_{m_0,n}\phi_{m_0,n}(t) + \sum_{m=-\infty}^{m_0} \sum_{n=-\infty}^{\infty} T_{n,m}\varphi_{n,m}(t) \tag{8.14}$$

Based on that we use a multi resolution decomposition algorithm which decomposes the input signal into scale and detail coefficients for each different scale. As stated above, each level (starting from the highest moving to the lowest) scaling coefficients are decomposed by the wavelet into a new set of detail and scaling coefficients for a lower level. At the end there are only detail coefficients and one scaling coefficient, which is the mean value of the signal (the scaling coefficients of each upper level can be reconstructed buy the detail coefficients of the lower level). Our methodology is based on the analysis of the distribution of the detail wavelet coefficients. The detail wavelet coefficients can be tiled or indexed as shown by Figure (17) (figure 3.8 in Addison 2002). In that same fashion they can be represented graphically - see Figure (18).

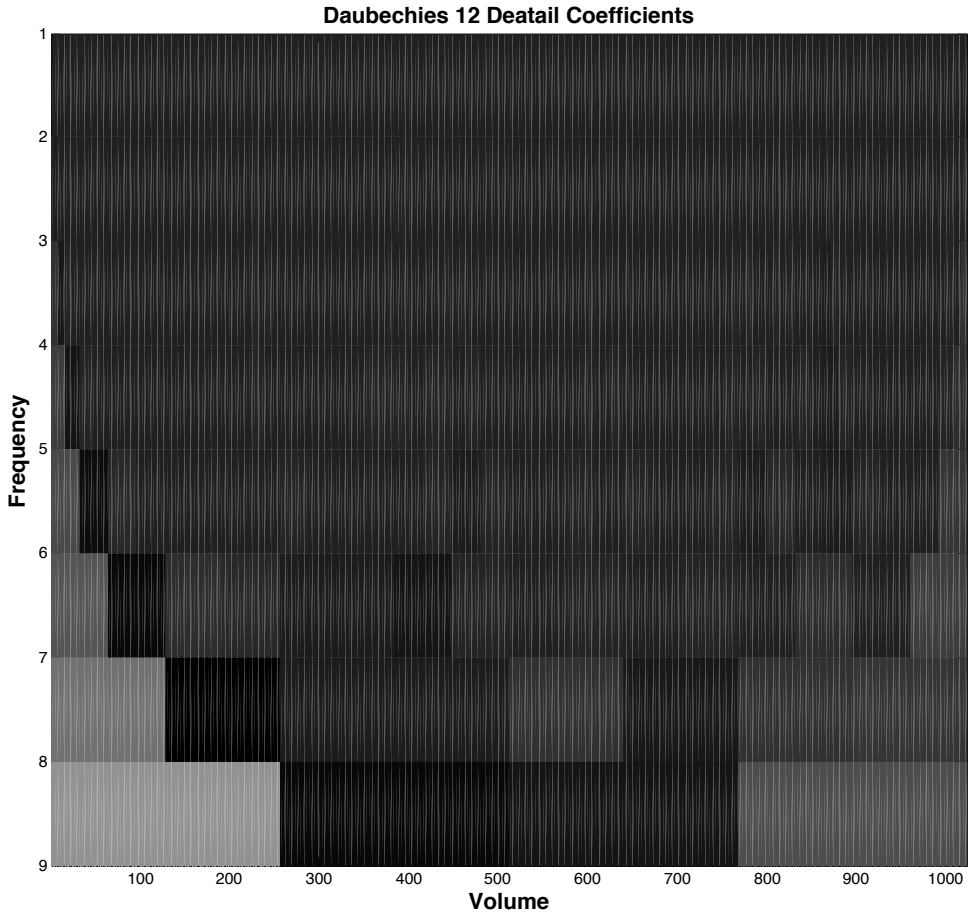

Figure 18: **Discrete Wavelet Detail Coefficients Tiling**

The figure shows Daubechies 12 wavelet decomposition of a limit order book. The coefficients are organized in volume (in times series it is time) frequency plane. Volume is on the horizontal axis and consists of 100-share units, meaning that the distance between each point on the axis is 100 shares. Frequency is on the vertical axis with highest frequencies on top lowest on bottom. See also Figure 17. The darker colors represent big negative coefficients and the lighter colors bigger positive coefficients.

9 Additional Graphs

Figure (19) presents the bid limit order books backward plotted for one day (7408 books). The books are plotted starting from the very back end of the book and proceeding towards the best quote. The new regime is plotted in green and he old regime is plotted in red.

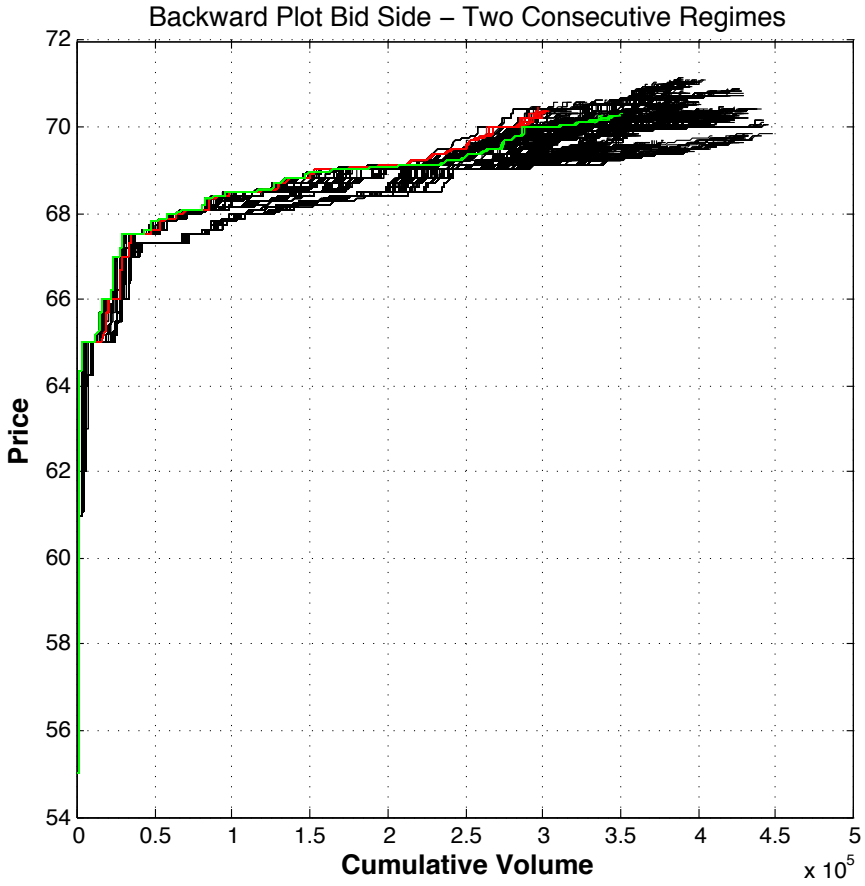

Figure 19: **Two Consecutive Regimes Vs All Books**

Figure shows bid limit order books backward plotted for one day - 7408 books. The books are plotted starting from the very back end of the book and proceeding towards the best quote. The new regime is plotted in green: books 755 - 788, regime size = 33, from sec 33354 to 33515 sec., duration: 161 sec. The old regime is plotted in red: books 604 - 647, regime size = 43, from sec 32916 to 33056 sec., duration: 140 sec. Transition size = 106 books, from sec 33060 to 33350 sec., duration: 290 sec.

References

Addison, P. (2002): *The Illustrated Wavelet Transform Handbook. Introductory Theory and Applications in Science, Engineering, Medicine and Finance.* lOP Publishing Ltd.

Fugal, D. L. (2009): *Conceptual Wavelets in Digital Signal Processing.* Space And Signals Technical Publishing.

Glosten, L. R. (1994): "Is the Electronic Open Limit Order Book Inevitable?," *The Journal of Finance*, 49(4), 1127–1161.

Grammig, J., A. Heinen, and E. Rengifo (2004): "Order submission on a pure limit order book: an analysis using a multivariate count data model," CORE Discussion

Papers No. 2004058.

Lehmann, B. (2008): "Arbitrage-free Limit Order Books and the Pricing of Order Flow Risk," NBER Working Paper No. 13848, March 2008.

Los, C. (2003): *Financial Market Risk Financial Market Risk: Measurement and analysis*. Routledge.