ESTIMATING THE PROBABILITY OF INFORMED TRADING: FURTHER EVIDENCE FROM AN ORDER-DRIVEN MARKET*

David Abad and Antonio Rubia**

WP-AD 2004-38

Corresponding author: David Abad, Department of Financial Economics, University of Alicante, P.O. BOX 99, 03690 Alicante. E-mail: goliat@ua.es Tel: +34 965 90 36 21 Fax: +34 965 90 36 21.

Editor: Instituto Valenciano de Investigaciones Económicas, S.A.

Primera Edición Octubre 2004.

Depósito Legal: V-4394-2004

IVIE working papers offer in advance the results of economic research under way in order to encourage a discussion process before sending them to scientific journals for their final publication.

---

* This paper was completed while Antonio Rubia was a visiting scholar at the Anderson School at UCLA, whose hospitality is gratefully acknowledged. Financial support from the Spanish Department of Science and Technology through project BEC2002-03797 is acknowledged.

** David Abad, University of Alicante.

Antonio Rubia, University of Alicante. E-mail: antonio.rubia@ua.es

ESTIMATING THE PROBABILITY OF INFORMED TRADING: FURTHER EVIDENCE FROM AN ORDER-DRIVEN MARKET

DAVID ABAD AND ANTONIO RUBIA

ABSTRACT

Nyholm (2002, 2003) [*J. of Financial Research*, 25, pp. 485; *J. of Applied Econometrics*, 18, pp. 457] has proposed a new procedure to infer the probability of informed negotiation on a trade-to-trade basis through a regime-switching model. We provide further empirical evidence about the performance of this model by using trade-related information, such as the degree of aggressiveness and the trade size, on a pure order-driven market. It is evidenced that the switching scheme of the basic model is closely related to the arrival of different types of orders and not necessarily to information. This feature also applies when controlling for market variables other than order aggressiveness (e.g., trade size). The updating process in the non-linear setting proves so complex that it is necessary to account for a number of different microstructure effects to provide probabilities related to information arrivals. This evidence casts doubts about the general suitability of the procedure.

Key Words: Probability of Informed Trading, PIN, Order aggressiveness.

JEL Classification: C21, C52, D82

1 Introduction

One of the most important topics in modern microstructure literature is concerned with the analysis of market informational asymmetries and their implications for traded assets. Investors holding private information lead to permanent changes in prices as they negotiate optimally to profit from their advantage. On the contrary, uninformed traders mostly trade for consumption or liquidity reasons and do not affect equilibrium prices permanently. The advantageous behaviour of the informed investors is thus masked by the noisy activity of the uninformed traders, leading to the impossibility to identify their informational nature but in terms of probability. This uncertainty implies a risk for liquidity suppliers, who have to face an adverse selection problem.

Measuring the degree of asymmetry deserves attention for its economic implications on the transaction costs and on the price formation process. For instance, it is widely accepted that bid-ask spreads increase with the adverse selection risk, as liquidity suppliers try to compensate in this way the higher expected loss derived from dealing with informed agents (Bagehot, 1971). Consistent with this hypothesis, thinly-traded and lowest-priced stocks, which are less widely followed and hence subject to a greater degree of informational asymmetry, tend to carry larger adverse selection costs. The degree of informational asymmetry can thus be inferred indirectly by appraising the average size of this cost in the bid-ask spread. Alternatively, it is also possible to infer the probability of informed negotiation (PIN henceforth) as a telling measure of asymmetry in the price formation process, which rises, for instance, the average relative frequency in which new private information is incorporated into prices.

Recent microstructure literature has suggested some empirical procedures to address this issue. The seminar work is due to Easley, Kiefer, O'Hara and Paperman (1996), who developed a method based on a mixture of three Poisson processes to model arrivals of (relevant) information. More recently, Nyholm (2002, 2003) has proposed a method based on a regime switching model with the appealing of regarding the behaviour of the informed investors as a latent variable whose dynamics can be estimated endogenously from trade-to-trade data. The core of the model is a non-linear generalization of the well-known trade-indicator regression (Glosten and Harris, 1988), which states the predictability of the quote midpoint change on the basis of observable information such as the trade direction. The Nyholm's approach generalizes this idea upon the assumption of two states – informed and uninformed– that induce a different midquote revision. When the dynamics of the process is in 'excited' state, the changes in the midpoint prove more sensitive to the observed half-spread, and it is heuristically assumed that the trade at that time is initiated by an informed investor. Private information arrivals are to a great extent identified on the basis of large relative changes in midquotes. Nevertheless, it should be noted that this feature is potentially troublesome, since it is not clear that the immediate update process corresponds one-to-one with information arrivals and, moreover, what is regarded as a 'large' change potentially due to private information is clearly conditioned by the previous specification of the mean in the model. The estimates resultant from an oversimplified modelling could lead to misleading conclusions.

The main aim of this paper is to provide further insight on the empirical performance and the suitability of this new procedure. To extent the evidence presented by Nyholm on a quote-driven dealer market (the New York Stock Exchange, NYSE), we exploit the information conveyed by trade-related variables which are publicly observable in order-driven markets. This extension does not offer any theoretical inconvenience, because the aggregate behaviour of the limit-order book is similar to that of a market maker and, in fact, it has been evidenced the role of limit-order traders as liquidity suppliers.[1] The publicly-disclosed information from those markets is likely to be useful to address the occurrence of private trades, so the conclusions resultant from a deeper analysis are able to shed light on the empirical performance of the regime-switching model.

Specifically, we use the aggressiveness of the traded orders as a key notion to enhance estimates. This magnitude underlies the trader's decision and is related to impatience and willingness to trade facing less and less competitive prices. Despite its potential relevance, this variable has received little attention in the context of private information. Nevertheless, there are a number of reasons for which order aggressiveness seems worthwhile for this analysis. First, it is obviously tied to the informed process in which prices are updated. It is accepted in the wide literature related to the topic of order submission strategies that the more aggressive is the order, the more

[1] The possibility of applying this model on both specialist- and order-driven markets is remarked in Nyholm (2002). Furthermore, there exists a growing interest for order-driven market as new trading systems and recently restructured exchanges apply a limit-order book design.

information is conveyed. Second, it subsumes to some extent the information related to the market environment and the asset dynamics. Relevant trade-related variables such as the thickness of the book, the price volatility, the size of the book, among several others, are determinants of the degree of aggressiveness. Hence, the small number of variables needed to fit aggressiveness are linked to a large set of useful information, which allows for a fairly parsimonious analysis. This is specially important in the non-linear context of the regime switching model. Third, aggressiveness is a magnitude observable for market suppliers and, in the case of an order-driven market, it can easily be computed from the publicly visible information which is disseminated by the limit-order book. Investors can use this information at any time to improve their knowledge and define optimal strategies. Therefore, it seems legitimate and sensible to incorporate this information to refine and perhaps improve the results from the original model.[2]

We apply the standard model and several extensions on data from the Spanish Stock Exchange. This is a pure, electronic order-driven market without market makers that operates with a fully centralized computerized system similar to the Toronto's Computer Assisted Trading System (CATS) popular around the world. We form three groups of assets attending on the different levels of traded volume and proceed to infer the PIN dynamics on each asset. The importance of the degree of aggressiveness is supported by an overwhelming statistical evidence. It is seen that the switching scheme of the basic formulation, which does not acknowledge the order design explicitly, is in reality closely related to the arrivals of a particular type of aggressive order –those that impose, on average, the highest immediate revision in midquotes– and not necessarily to information disclosure. The effect due to order design in this context is far more important than that related to any other variable on the trader's decision, such as the size or the direction of the trade. Therefore, any attempt to enhance the specification by adding market variables still neglecting aggressiveness – as done in Nyholm (2003) – does not prevent the model to switch according to the above pattern. The revision process is sensitive to so many different microstructure factors that just controlling for order typology is not sufficient to isolate private information in the latent

[2]Note that under the philosophy of the model, the liquidity suppliers are particullarly interested in determining the nature of the trade-initiating investor as accurate as possible. These agents will be willing to observe and procese any signal that could improve their beliefs. Therefore, aggressiveness, among other potential variables, is likely to be useful in this context.

variable. The specification must be further generalized including additional microstructure effects (we identify the necessary, but probably not sufficient, role of trade size) in order to obtain probabilities measures coherent with information arrivals patterns. Extensions accounting for further effects are thus likely to yield better results, yet on the cost of putting considerable strain on an already heavily-parameterized model, which seems unfeasible in practice. Therefore, the evidence found on this paper casts doubts on the general suitability of the model.

This paper contributes to the previous microstructure literature in several ways. First, we provide an extensive discussion of the only procedure intended to approach the probability of informed trading on a trade-to-trade basis. We show that the information conveyed in some market-related variables observable by traders in real time must necessarily be exploited to get estimates which are related to some extent with the probability of private information arrivals. These measures might be suitable of being used as a proxies of information for further empirical applications on trade data, though more research at this point is deserved. Second, we show the fairly complexity of the immediate updating price process, which contrasts with the simplicity assumed in most empirical models. This evidence is similar to the results outlined in Kempf and Korn (1999), who based on a neural network model find that the assumption of a linear impact of orders on prices is highly questionable. We evidence that the price revision process displays a non-linear behaviour that is sensitive to a number of microstructure effects related to the market environment and the stock dynamics. This evidence is relevant to improve the knowledge about liquidity dynamics and traders' behavior. Finally, we underline the important role played by some variables in the price updating process, among which the degree of aggressiveness seems remarkably important. As remarked before, aggressiveness has received less comparative attention in the context of informational asymmetries than other trade-related variables such as the trade size, no doubt because of the limitations of the databases available in earlier works. The increasing availability of detailed information about trades and orders from exchanges over the world, as well as the overall evidence about the important role of aggressiveness in the price revision process, make this variable worthy of attention in further research.

The balance of the paper is organized as follows. Section two briefly describes the general theoretical background of the model. Section three states the basic setup of the Nyholm's model and discusses further generalizations.

Section four introduces the dataset and presents the usual descriptive analysis. Section five discusses the main findings and the implications from the estimation of several models. Finally, section six summarizes and concludes.

2 Theoretical background

Informational asymmetries were early studied by Bagehot (1971) and Jaffe and Winkler (1976), who suggested the distinction between informed and noise traders. Since then, a great deal of literature has focused on the consequences of market asymmetries. Of these issues, two are of special interest for market microstructure purposes.

One is to identify the different components (adverse selection, inventory-carrying and order-processing costs) that characterize the bid-ask spreads. The theoretical and empirical work in this area is extensive, so a concise revision of the literature is beyond the aim of the paper. Earlier statistical models used the simple time-series properties of the transaction prices (Roll, 1984; Choi, Salandro and Shastri, 1988; Stoll, 1989), whereas another category of models focus on structural models based on the trade-indicator regression model and its extensions: among others, Glosten and Harris (1988), George, Kaul and Nimalendran (1991), Lin, Sanger and Booth (1995), Madhavan and Cheng (1997) and Huang and Stoll (1997). Hasbrouck (1988, 1991a, 1991b) developed an alternative framework based on the vector autoregression analysis.

The second issue is related to determine the probability of informed transactions as a direct measure of asymmetric information. This is not an easy task, and the few empirical procedures put forward to address this topic is in contrast to the wide framework related to identify the adverse selection component. The original contribution is due to Easley *et al.* (1996) who developed a sequential model based on a mixture of three Poisson processes. The most attractive feature of this methodology is its simplicity: an averaged PIN measure over a given period can be inferred on the basis of the daily number of buyer- and seller-initiated trades. Applying this model to the U.S. market, the authors found a significant lower PIN measure for the group of actively-traded stocks. This procedure is widely accepted and has subsequently been used and extended to address a wide range of empirical issues in market microstructure (see Easley, Hvidkjaer, and O'Hara 2002 for a recent analysis and references therein).

Nyholm proposed a different approach to estimate the probability of informed trading. This model belongs to the framework of the trade-indicator model and hence inherits the advantages related to that formulation, but also most of its disadvantages. The most appealing feature is that a conditional PIN measure could be estimated for each trade, thus providing an interesting basis for further empirical applications. Nyholm applies the model on stocks from the NYSE and evidences as well a higher average PIN measure for the less liquid assets. We shall discuss in greater extent the main features of the trade-indicator model and its extension towards the regime switching model in the following section.

Finally, there exists an extensive body of literature related to aggressiveness, mainly focusing on the order submission strategies and their determinants. Cohen, Maier, Schwartz and Whitcomb (1981) early analyzed order submission. Biais, Hillion and Spatt (1995) outlined the relation between the order flow and the state of the limit order book in the Paris Bourse, finding evidence on strategies based on order placement: traders tend to place limit orders when the spread is large and the order book is thin, and the opposite is true for market orders. Harris and Hasbrouck (1996) provide evidence about the performance of order submission strategies on the NYSE as well. Recent papers have analyzed the determinants underlying the trader's decision to submit more aggressive orders. The execution probability, the transaction price and the risk of adverse selection plays a key role in this decision; as such, it is seen that trade- and market-related variables (such as spread, order size, thickness and transient volatility) condition the trader's decision as to which type of order to submit (Ranaldo, 2004). Griffiths, Smith, Alasdair, Turnbull and White (2000) evidenced that buy (sell) aggressive orders are more likely with small-firm stocks and when the limit-order book has a wide bid-ask spread and high (low) depth on the same (opposite) side as the order.

3 The basic Markov regime-switching model

Traders who hold private information can make a profit while fundamental prices do not reflect full information. They trade advantageously and force prices to change to correct unbalance. In addition, stock prices change as public information is released. The trade-indicator model combines both sources and assumes that quotes are adjusted to reflect the private information revealed by both the previous trade and the current random arrivals of public information,

$$\Delta M_t = \alpha \frac{S_{t-1}}{2} Q_{t-1} + \varepsilon_t; \quad \varepsilon_t \sim \left(0, \sigma^2\right); \ t = 1, ..., T \tag{1}$$

where ΔM_t denotes the variation of the midpoint of the bid-ask spread, S_t is the size of the quoted spread and Q_t is the trade-indicator variable signalling whether transactions are at the ask $(Q_{t-1} = 1)$ or at the bid side $(Q_{t-1} = -1)$. The arrival of public information is regarded as a white noise process, ε_t. The coefficient α measures the size of the adjustment on the half-spread and includes the effect attributable to both adverse selection and inventory holding costs. We shall denote $X_{t-1} = (S_{t-1}Q_{t-1})/2$ to avoid overloading unnecessarily the subsequent equations.

The Nyholm's approach extends the central idea of the trade-indicator model in (1) by allowing for non-linear dynamics. The midquote revision is conditioned to two latent states that are assumed to represent the particular nature of the trader. As liquidity suppliers face the risk of adverse selection, they are particularly interested on indentifying the hidden nature of the investor iniciating the trade. The model is thus defined on a stochastic process, say I_t, which takes a binary range of values depending on whether the trade at time t is initiated by an informed agent $(I_t = 1)$ or by a noise investor $(I_t = 0)$,

$$\begin{aligned} \Delta M_t &= \alpha_0 X_{t-1} + \varepsilon_t, & \text{if } I_{t-1} = 0; \quad \varepsilon_t \sim (0, \sigma^2) \\ \Delta M_t &= (\alpha_0 + \alpha_1) X_{t-1} + \varepsilon_t, & \text{if } I_{t-1} = 1; \quad \varepsilon_t \sim (0, \sigma^2) \end{aligned} \tag{2}$$

where α_0 gauges midquotes response to non-informational trades and α_1 reflects the additional effect when trades are initiated by informed investors. Private information arrivals lead to higher revisions on midquotes, so α_1 is expected to be significantly different from zero.

The latent variable I_t is assumed to follow a first-order Markov regime switching process. The probability of being in a particular state at t only depends on the state prevailing in period $t-1$, $t = 1, ..., T$. The dynamics of the process is characterized by a 2×2 transition matrix, here denoted as $\mathbb{P}$, which governs the whole process and determines the probability of being at each state at any time. The on-diagonal elements p_{ii} of this matrix are assumed to be unknown, constant parameters such that,

$$p_{ii} = \Pr\left(I_t = i | I_{t-1} = i\right); \ 0 < p_{ii} < 1, i = \{0, 1\} \tag{3}$$

and,

$$\mathbb{P} = \begin{pmatrix} p_{00} & 1 - p_{11} \\ 1 - p_{00} & p_{11} \end{pmatrix} \tag{4}$$

The specification is completed by assuming a particular distribution for the error term in (2). While the trade-indicator regression equation is usually estimated through GMM or least-squares without explicitly making this assumption, the regime-switching modelling requires of maximum likelihood or Bayesian methods. It is assumed that the disturbance term is driven by independent, identically distributed Gaussian innovations with variance σ^2 regardless of the particular value of the underlying latent variable.[3] The set of parameters $\theta = (\alpha_0, \alpha_1, \sigma, p_{00}, p_{11})'$ is then estimated by quasi-maximum likelihood (QML), as the assumption of normality cannot be regarded as realistic. In general, it provides consistent –though inefficient– estimates provided correct specification and moderate departures from the assumption of normality.

The QML estimates are then used to infer the dynamics of the hidden Markov chain that characterizes the arrival of private information. The conditional PIN estimate at any time t is inferred as the smoothed probability $\Pr\left(I_t = 1 | \Psi_{\mathrm{T}}; \hat{\theta}\right)$, where Ψ_{T} denotes the set of available information up to time T. Nyholm (2002) suggests a measure representative of the unconditional PIN over the entire period in the spirit of Easley *et al.* (1996), determined as the mean value of the smoothed probabilities. Note that under the implicit assumption of ergodicity in the Markov chain, this measure is representative of the unconditional probability of the process being at excited regime, $\Pr(I_t = 1)$, so it can readily be estimated as $\tilde{p}_{00} / (\tilde{p}_{00} + \tilde{p}_{11})$, with $\tilde{p}_{ii} = 1 - \hat{p}_{ii}$. Other alternatives might be possible as well, but we shall focus on this estimate because it is directly implied by the regime-switching dynamics and, therefore, seems the most natural measure.[4]

[3] The error term is aimed to capture public information shocks. Note that it is implicitly assumed independence between public and private information.

[4] Nyholm (2003) measures the inconditional probability of information arrivals after classificating each trade. This implies to make an exogenous decision about the threshold which determines whether a trade is informed-initiated or not.

3.1 Extending the model: a discussion

An implicit consequence of the model involved is that the midquote revision drives to *transitory* changes in prices when $I_t = 0$, and *permanent* changes when $I_t = 1$. Both components are consequently characterized through the estimates of α_0 and $(\alpha_0 + \alpha_1)$, so average transitory price updates are typically regarded as smaller than permanent ones. However, since the model does not involve an inter-temporal dimension beyond one trade, it is not easy to figure out how a change that lasts long in time can truly be identified.[5] It seems clear that these states are related to large and small *immediate* revisions, but not necessarily to permanent and transitory changes.

Actually, this argument constitutes the main criticize for structural models based on the trade-indicator setup and, in fact, there are some concerns about the ability of these models to identify precisely the adverse selection component (see the empirical analysis in Van Ness, Van Ness and Warr, 2001), and some authors advocate to use of more sophisticated, dynamic methodologies for appraising the long-run price impacts of trading (see for instance Hasbrouck, 1991). Nevertheless, the trade-indicator models have been applied intensively in the recent literature, finding a qualitative evidence largely consistent with the hypothesis conjectured a-priori.[6] Similarly, the evidenced reported by Nyholm on the empirical application of the regime-switching model on the NYSE is roughly consistent with the main patterns related to asymmetric information. Some caution should be exercised, as these estimates are based on estimates which are likely to be subject to measurement errors.

In estimating the Nyholm's model, the regime-switching scheme identifies each state through the prediction errors of the mean equation. If errors are relatively large, the underlying observations are much likely classified as excited so that a larger revision can reduce the error size. It is therefore remarkably important to define the mean equation as precise as possible in this framework. Yet the basic model assumes that a narrow set of information (the observed half-spread and the trade-indicator variable) is enough to characterize the hidden dynamics underlying the arrivals of private information.

[5] The model assumes that all price effects are incorporated in the first transaction price, as predicted by the semi-strong form market efficiency (Glosten and Milgrom,1985).

[6] Taking the measure of Easley *et al.* (1996), Chung and Lee (2003) verify that the estimates of the adverse-selection cost from several procedures is indeed related to the probability of information-based trading, providing empirical support for these procedures.

This makes the estimation of the model fairly feasible, but it could turn out to be too simplistic.

There exists a great deal of observable information relevant for the better comprehension of information arrivals in the quote-to-quote framework. Hasbrouck (1991), Huang and Stoll (1997), Dufuor and Engle (2000) and recently Pascual, Escribano and Tapia (2004), among others, have evidenced the empirical relevance of several variables related to market environment in the price revision process. Furthermore, in the conception of the model it is embedded the idea of a liquidity supplier trying to identify the nature of the investor by using the available information. A very appealing feature of the trade-indicator setup is that the basic formulation can readily be extended to incorporate any relevant microstructure effect by introducing indicator variables that are 1 under a specific condition and 0 otherwise. The effects related to trade conditions could thus be included in a simple way, so that the basic relation between ΔM_t and X_{t-1} stated in both (1) and (2) could be much enhanced by using the extra information related to market conditions.

From an econometric viewpoint, estimating the *linear* trade-indicator regression or an categories-extended model results only in different interpretations of the parameter estimates coupled with a potential gain in statistical efficiency. However, the picture can radically be different in the context of the regime-switching model, because the ultimate aim here is to characterize the dynamics of the hidden Markov chain. Any unexpected large shock in the midquote is likely regarded as an information arrival. Therefore, an oversimplified mean equation neglecting relevant microstructure effects could eventually lead to unreliable estimates of the regime switching dynamics, which is the whole purpose of the procedure.

Nyholm (2003) includes the effects of trade size in a further extension of the basic model, finding that volume affects quote reactions subsequent to normal-information, but no volume-effect seems to apply to private-information initiated trades. This point is, nevertheless, quite surprising and really unappealing, since a number of papers have reported the heterogenous response of the price updating process on trade size (among others, Huang and Stoll, 1997; and Ahn, Cai, Hamao and Ho, 2002). Furthermore, it is believed that large trades tend to convey more information because informed investors would be willing to trade large amounts at any given price (Easley and O'Hara, 1987) and because of the potential role of volume as a signal of the precision of beliefs (Blume, Easley and O'Hara, 1994). The trade size is an important variable because underlies the investor's decision, though it could be not as much important as other variables in the current context or should not be regarded solely. We propose to use another variable related to the investor's decision –degree of aggressiveness– as starting point to enhance the basic formulation, firstly, and assess the robustness of the results from the basic model when these effects are ignored.

3.1.1 Order aggressiveness

Investors trade through very different types of orders, thus generating a complex link between the dynamics of prices and the order submission process. As orders are placed following trading strategies, they convey information which could be processed to infer PIN measures more accurately. Traders can submit three basic types of standard orders in the Spanish Stock Exchange: *limit orders*, *market to limit orders* and *market orders*. This nomenclature is slightly different from that used in other exchanges. Limit orders specify conditions such as the quantity, the direction (purchase or sell), the price and the date when the order will be withdrawn. They typically provide liquidity to the system by either widening the depth of an existing quote or posting new quotes, but often can be traded immediately if there is a valid counterpart on the book (marketable limit orders). On the other hand, market or market to limit orders consume liquidity. Market to limit orders only specifies the quantity and direction of the trade, so they are executed immediately at the prevailing quote. Note that they are not allowed to walk up the book if the quantity available at the best price is not sufficient to fulfill the total volume required. Instead, the non-executed part keeps waiting for balancing entries in the other side of the book at the transaction price. Finally, market orders are intended to achieve a full execution by allowing the order to walk up the book till completion. They provide the fastest execution, but imply higher costs than more passive orders.

A suitable procedure would rank orders according to their degree of aggressiveness. Most aggressive orders trade at any given price, so they likely convey more information that less aggressive orders –in other words, they are more likely submitted by informed investors. We initially classify orders implying immediate execution in the same spirit than Bias *et al.* (1995). We consider three categories (we initially do not distinguish between purchase and sell orders) consistent with previous literature:

A1 The most aggressive orders, namely $A1$, demand a quantity larger than that available at the prevailing quote (either ask or bid). They are executed by walking up the book to complete the volume required or find a less competitive price. They imply a high degree of impatience because show the willingness to trade at any given price. The immediate effect is to widen the spread and change the midquote.

A2 The second level, namely $A2$, includes orders submitted by investors that are willing to trade at the current bid or ask quote a quantity larger than that available, so they do not allow to walk up the book. The depth at the current quote is then fully consumed to partially fill the order and the remaining part is transformed into a limit order at the transaction price. The immediate effect is to change the midpoint but, alike A1 orders, the effect on the spread depends on the size of the pre-trade spread size and the quotes.

A3 Finally, category $A3$ includes orders trading for a quantity lower or equal than that available at the current bid or ask quote. The immediate effect is a full execution and either a reduction or the full consumption of the available depth at the best quote.

All trades composing the database fall into one of these three categories.[7] So let $\mathbf{A} = \{A1, A2, A3\}$ be a set of subscripts and $D_{\mathbf{j},t},\ j \in \mathbf{A}$, an indicator taking value equal to one if the particular trade at time t is qualified as in its subscript and zero otherwise. We incorporate this potential valuable information into (2) to control for systematic changes in the midquote given the order typology in the two-stated latent variable framework,

$$\begin{aligned} \Delta M_t &= \sum_{j \in \mathbf{A}} \alpha_{0,\mathbf{j}} X_{t-1} D_{\mathbf{j},t-1} + \varepsilon_t, & \text{if } I_{t-1} = 0; \qquad (5) \\ \Delta M_t &= \sum_{j \in \mathbf{A}} \left(\alpha_{0,\mathbf{j}} + \alpha_{1,\mathbf{j}}\right) X_{t-1} D_{\mathbf{j},t-1} + \varepsilon_t, & \text{if } I_{t-1} = 1; \end{aligned}$$

where $\varepsilon_t \sim (0, \sigma^2)$. Hence, the latent variable is driven by the unexpected shocks given the particular type of order. Note that imposing the linear

[7] Trades are necesarilly initiated by some type of market order or marketable limit orders, so limit orders that not imply an inmediate execution are excluded. Keim and Madhavan (1995) show that liquidity traders are likely to use market orders, but that informed traders whose information value decays slowly tend to use limit orders. Nevertheless, Griffiths *et al.* (2000) evidence that small limit orders or orders that not generate an inmediate execution has in general a small or unsignificant price impact.

restrictions $\alpha_{0,j} = \alpha_0$ and $\alpha_{1,j} = \alpha_1$ leads to the basic formulation, so the suitability of this extension over the benchmark model can readily be tested with a standard likelihood ratio test.

This model allows for several degrees of aggressiveness regardless the state of nature. Informed agents could submit conservative market limit orders to avoid be early detected as a part of their strategies, and uninformed agents could submit aggressive orders due to imperative liquidity reasons.[8] Note that the factors underlying order placement depend on unobservable variables, such as the investor's information set and their personal preferences towards risk and portfolio allocation, but also on observable trade-related variables, as remarked previously. Despite the endogenous nature of the order placement, aggressiveness is regarded here as an exogenous variable. In the context of the trade-indicator model, in which the trade flow –and even the trade size– is also treated as exogenous, this assumption is not particularly strong.

4 Data description

The Spanish Stock Exchange (SSE hereafter) is a computerized limit order market. It uses a centralized electronic system known as SIBE (*Sistema de Interconexión Bursátil Español*) similar to that in Brussels (NTS), Paris (CAC), Stockholm (SAX) or Toronto (CATS). The exchange opens at 8.30 with a call auction after which stocks are traded on a continuous basis from 9.00 through 17.30. There are no market makers or floor traders, so liquidity is strictly supplied by a limit-order book which collects all the buy and sell proposals submitted by traders.

The dataset was facilitated by *Sociedad de Bolsas S.A* and contains information about every event that takes place in the first level of the limit order book. Both introductions and cancellations of limit orders as well as new transactions generate new records of trades and quotes. The sample period covers from September 1st to December 29th, 2000. For each stock, the quote-by-quote data set reports the transaction data (time stamp, price, volume in number of shares) and the order flow (time stamp, prevailing quotes, accumulated traded volume and depth in shares). Thus, the data set provides information on market orders and the best buy and sell prices (limit orders

[8] Griffiths *et al.* (2000) report that aggressive sells are morel likely motivated by liquidity than aggressive purchases.

at and within the previous quotes), but does not provide data outside the prevailing spread. The information of the state of the book is publicly visible and it is disseminated continuously during the trading session. It provides information in real time on the better quotes (pre-trade transparency) and transactions (post-trade transparency).

We processed the dataset to obtain the relevant variables for the analysis. We only use transactions, which are easily identified through the change in the accumulated volume at any time. Each transaction is exactly classified as either a buyer- or a seller-initiated trade by the location flags available from the dataset, without need of a classification algorithm. Also, the size of the spread immediate prior to the trade is readily collected. Finally, the order flow is processed by an algorithm that determines the degree of aggressiveness as a function of the traded size, the prevailing depth and the quotes.

We applied some filters to rule out potential anomalies. Stocks that not fulfill minimum daily activity requirements are excluded. We then apply a similar procedure than that in Easley *et al.* (1996) in order to make meaningful comparisons attending on the degree of trade activity. We rank stocks on traded volume in year 2000 and build seven groups. We then focus on stocks included in the second, fourth and sixth volume categories, namely Groups 1, 2 and 3, roughly representative of high, medium and low trading activity, respectively, and proceed to infer the PIN measure on each asset.

Table 1 presents some basic summary statistics over the sampled period for the 39 assets included in the sample. The variables refer to the liquidity levels, the traded volume, the price and the volatility of the assets involved. All these statistics underline the huge differences in terms of trading activity. The average traded volume in million of € ranges from 6.25 (Group 3, low liquidity) to 321.61 (Group 1, high liquidity). Overall, it can be seen than assets in Group 3 are on average more illiquid, more volatile and less priced than assets in Group 2 and 3. This feature also applies when comparing assets from Group 2 to those in Group 1. It should be expected, therefore, that the PIN measure decrease monotonically over the three groups, as there are clear, significant differences between the liquidity levels of the assets.

Table 1
Sample Descriptive Statistics

Ticker	Company Name	Sample Size	Relative Spread	Spread in ticks	Bid Depth €	Ask Depth €	Volume Mill €	% Price Change	Mid-point	Price Volat.
Group 1: High-volume assets										
ACE	Acesa S.A.	23091	0.0031	2.83	15745	16717	333.84	7.45	9.02	0.0014
ACR	Aceralia SA	27832	0.0034	3.13	14325	12092	189.13	-3.41	9.24	0.0017
ACS	Construcción y Servicios, S.A.	19137	0.0039	10.32	21657	18412	346.38	-9.16	26.59	0.0014
AGS	Aguas de Barcelona, S.A.	19714	0.0037	5.20	13752	11920	248.16	-2.62	14.24	0.0013
ALB	Alba, S.A.	14641	0.0058	16.00	14645	15860	383.98	-9.86	27.74	0.0017
ANA	Acciona, S.A.	22745	0.0033	12.61	18211	20670	466.08	4.34	37.99	0.0012
AUM	Aurea Conc. Infraestructuras, S.A.	6190	0.0049	8.37	20969	15162	168.06	7.43	17.21	0.0019
DRC	Dragados, S.A.	35401	0.0036	3.87	19463	19344	742.28	27.89	10.57	0.0015
MAP	Mapfre, S.A.	8234	0.0066	12.77	16776	15826	266.93	14.48	19.29	0.0022
NHH	NH Hoteles, S.A.	18319	0.0047	6.26	15010	18105	428.27	-5.39	13.29	00016
REE	Red Eléctrica de España, S.A.	21349	0.0043	4.52	12513	8525	133.17	-5.93	10.55	0.0017
SOL	Sol Melia, S.A.	21512	0.0045	4.68	12662	11823	261.62	-14.06	1062	0.0019
VAL	Vallehermoso, S.A.	19716	0.0044	2.98	12161	11876	213.05	-9.10	6.79	0.0018
	Cross-sectional averaged value	*19837*	*0.0043*	*7.20*	*15992*	*15103*	*321.61*	*0.16*	*16.40*	*0.0016*
Group 2: Medium-volume assets										
AEA	Azucarera Ebro Agrícola, S.A.	3243	0.0096	12.87	9295	8499	32.13	-8.03	13.51	0.0034
AZC	Asturiana del Zinc, S.A.	5115	0.0076	7.40	10503	8308	48.71	-11.71	9.90	0.0031
AZK	Azkoyen S.A.	4112	0.0093	6.35	64440	6400	22.71	-22.52	6.91	0.0039
CPF	Campofrío Alimentación, S.A.	3129	0.0080	9.84	11425	7475	26.09	6.24	12.25	0.0026
CRI	Cristaleria Española, S.A.	2662	0.0075	23.72	11159	11312	37.22	-17.91	31.97	0.0028
ENC	Grupo Empresarial Ence, S.A.	4606	0.0076	13.52	8538	9608	56.68	-15.73	17.84	0.0032
PAS	Banco Pastor, S.A.	1332	0.0048	22.04	23699	23268	16.41	0.11	45.97	0.0006
PQR	Parques Reunidos, S.A.	4368	0.0093	4.59	5059	6638	30.59	-12.04	4.93	0.0037
SOS	Sos Arana, S.A.	2591	0.0107	9.69	8023	8519	17.86	-16.18	9.06	0.0026
TAZ	Transportes Azkar, S.A.	3243	0.0121	8.28	4561	57811	38.02	-33.89	7.18	0.0047
VDR	Portland Valderrivas, S.A.	1264	0.0127	26.59	9601	8222	24.55	-7.04	21.00	0.0024
VIS	Viscofan, S.A.	9709	0.0058	3.56	7535	6254	58.65	-34.51	6.42	0.0028
ZOT	Zardoya Otis, S.A.	5200	0.0051	4.67	10840	13100	60.68	0.88	9.07	0.0021
	Cross-sectional averaged value	*3890*	*0.0085*	*11.78*	*9745*	*9491*	*36.18*	*-13.26*	*15.08*	*0.0029*
Group 3: Low-volume assets										
ASA	Tavex Algodonera, S.A.	1070	0.0161	3.79	4633	5017	4.34	-24.80	2.37	0.0035
BAM	Bami S.A.	1739	0.0107	3.20	9499	6958	8.61	0.99	3.02	0.0031
CAF	Cons.y Aux. de Ferrocarriles, S.A.	673	0.0173	37.81	8755	6339	4.80	-11.06	21.99	0.0024
DGI	Dogi International Fabrics, S.A.	2228	0.0127	8.18	5035	3827	7.65	-44.44	6.72	0.0044
ENA	Enaco, S.A.	2525	0.0216	10.92	3289	4724	13.60	27.98	5.11	0.0056
IBG	Iberpapel Gestión, S.A.	804	0.0195	20.09	4063	5455	2.48	-10.36	10.37	0.0034
MCM	Miquel y Costas & Miquel, S.A.	789	0.0228	54.29	12438	6440	7.52	11.36	23.62	0.0025
NEA	Nicolás Correa, S.A.	923	0.0172	5.18	8104	3469	2.09	-35.14	3.08	0.0035
PAC	Papeles y Cartones de Europa, S.A.	1122	0.0164	2.50	6184	5137	6.18	-36.71	1.60	0.0032
RIO	Bodegas Riojanas, S.A.	728	0.0196	17.35	6227	3652	6.85	-2.60	8.94	0.0025
UBS	Urbanizaciones y Transportes, S.A.	938	0.0219	1.63	8342	6229	2.44	-26.14	0.76	0.0033
VWG	Volkswagen Aktiengesellchft, S.A.	835	0.0244	46.55	8742	6869	10.63	11.58	54.14	0.0046
ZNC	Española del Zinc, S.A.	893	0.0211	5.55	5132	4579	4.06	-73.23	2.92	0.0040
	Cross-sectional averaged value	*1174*	*0.0186*	*16.70*	*6957*	*5284*	*6.25*	*-16.35*	*11.13*	*0.0035*

The table shows the name and ticker of all the companies included in the sample. The number of observations and the mean values related to the spread, depth, volume, price and volatility are also provided. *Relative Spread*, *Spread in ticks*, *Bid Depth €, Ask Depth €* and *Midqoute* are time-weighted means over the eighty day period. *Volume Mill €* is the total volume traded in € during this period of time. *% Price Change* shows the return calculated from the first transaction price in day 1 to the last transaction price in day 80. Finally, *Price Volatility* is measured as the standard deviation of transaction prices for the period. Means of these variables are also show for each of the three activity portfolio formed.

5 Empirical evidence

5.1 Basic model

We turn to analyzing the results from estimating (2). The mean, median, maximum and minimum of the parameter estimates as well as their mean asymptotic standard errors are reported in Table 2. The mean value of the estimates for α_0 ranks between 0.06 for the group of frequently-traded assets and 0.080 for the other two groups. Overall these estimates represent a half of those reported in Nyholm (2002). The estimates are much closer to those in Nyholm (2003), which correspond with twenty of the most actively traded assets from the NYSE. The estimates for α_1 in Table 2 are within 1.37 (Group 1, high volume) and 1.31 (Group 3, low volume). Though is tempting to make a direct comparison of these coefficients across volume categories, some care should apply in doing so. The higher estimate of α_1 seems to suggest that a greater adjustment is made for liquid assets, but these coefficient would apply on the average spread, which is smaller for liquid assets. A higher scale on a smaller margin could not lead necessarily to a higher price revision.

The estimations for the probabilities of informed trading over the analyzed period would be comprised between a 12.7% for the group of most liquid assets and a 16.7% for the group of thinly-traded stocks. These probabilities are higher than those reported in Nyholm –especially in the low volume category– essentially because the excited state is found here to be more persistent in all the volume groups. Setting apart the quantitative differences between the estimates from both different markets, the qualitative conclusions are similar, and a smaller probability of asymmetric information for frequently-traded assets would be evidenced.

5.2 Order aggressiveness

The arguments for the inclusion of order aggressiveness as a compelling variable in this framework find an overwhelming statistical support. The log-likelihood function value is largely increased for all stocks (see Appendix A), and the likelihood-ratio tests for equality of estimated coefficients is always strongly rejected. While an acceptation of the null of these tests would have

Table 2
Baseline Model

Parameter	Mean	Median	Minimum	Maximum	Mean St. Error.
		Group 1: High-volume assets			
α_0	0.066	0.066	0.046	0.077	0.005
α_1	1.372	1.384	1.278	1.464	0.032
σ	0.017	0.014	0.007	0.039	0.001
p_{00}	0.905	0.902	0.880	0.923	0.005
p_{11}	0.342	0.336	0.266	0.409	0.020
PIN	0.127	0.128	0.096	0.168	0.001
		Group 2: Medium-volume assets			
α_0	0.081	0.081	0.048	0.109	0.017
α_1	1.344	1.317	1.198	1.600	0.061
σ	0.031	0.021	0.010	0.070	0.002
p_{00}	0.878	0.890	0.834	0.916	0.012
p_{11}	0.380	0.387	0.291	0.470	0.040
PIN	0.165	0.153	0.122	0.221	0.002
		Group 3: Low-volume assets			
α_0	0.081	0.085	0.015	0.132	0.018
α_1	1.306	1.259	1.051	1.634	0.079
σ	0.059	0.026	0.004	0.368	0.004
p_{00}	0.887	0.879	0.861	0.932	0.019
p_{11}	0.427	0.428	0.323	0.499	0.068
PIN	0.166	0.173	0.096	0.215	0.005

The table summarizes the parameter estimates (average value, median, maximum, minimum and average asymptotic standard error) from estimating the baseline model (see equation (2) and description in the main text) and the inferred probabilities of informed negotiation (PIN). The subindexex 0 and 1 refers to the normal and excited state respectively.

implied that the price revision is independent of the architecture of the orders, it is evidenced that this issue is indeed relevant to set forth immediate changes in prices. There is therefore a great deal of valuable information conveyed in the order typology. For induction, variables other than aggressiveness could affect the immediate price revision as well. This evidence is in contrast to the basic model setup, which regards the price revision exclusively as a matter of information.

The results from the estimation of the extended model are presented in Table 3. The A2 orders, which do not allow to walk up the book, generate the largest revision in the midquote across the three volume portfolios. Most aggressive orders of the A1 type also induce important revisions, although not as large as A2. Finally A3 orders imply the lowest revision, much more moderate than the other categories. The most relevant feature is that the highest revision is not related to the most aggressive orders that, nevertheless, are more likely to convey information. This feature could be explained as that the immediate revision in prices is sensitive to both information and order design, and A2-type orders favor larger price revisions over A1 because of their design.[9] Immediate revisions include a possible effect due to the private information, but also a transitory effect due to microstructure effects, such as the order architecture. As this framework is based on immediate effects, it turns out that A2-type orders can on average exhibit higher revisions.[10]

[9] A similar feature is noted in Degryse, Jong, Van Ravenswaaij and Wuyts (2003) by using time windows around the aggressive order. To see this feature, consider this simple example. Let the pre-trade bid and ask quotes be b_0' and a_0' respectively. Assume for simplicity that the ask side includes only another quote, a_0'', and that a trader is willing to buy a quantity exceeding the depth available at a_0' and that would be partially filled with the depth at a_0''. This trader could submit either an A1 or an A2 order. The inmediate midquote change from the A1 order would be $\left(a_0'' - a_0'\right)/2$, while the change from the A2 order be $\left(a_0'' - b_0'\right)/2$. The latter excedes the former in $\left(a_0' - b_0'\right)/2$, the pre-trade half-spread. Therefore, A1 orders must walk up the book up to a far out quote in order to provoke a higher ceteris-paribus inmediate revision than A2 orders.

[10] Note that the inmediate relative price updating measured through the midquote change on the signed half-spread should not be misunderstood with the *price impact* as measured in Griffiths *et al.* (2000) The price impact is a more sophisticated measure given by the ratio of the realized price (the weighted average of the prices filling the order) to the pre-trade prevailing quote. Griffiths *et al.* (2000) evidence that the *price impact* increases monotically with order aggressiveness.

Table 3
Order I Model

Parameter	Mean	Median	Minimum	Maximum	Mean St. Eror.
		Group 1: High-volume assets			
$\alpha_{0,A1}$	0.328	0.312	0.238	0.414	0.025
$\alpha_{1,A1}$	2.180	1.995	1.475	4.058	0.211
$\alpha_{0,A2}$	1.223	1.240	1.062	1.327	0.029
$\alpha_{1,A2}$	4.127	4.294	2.369	5.702	0.511
$\alpha_{0,A3}$	0.020	0.019	0.003	0.043	0.005
$\alpha_{1,A3}$	1.096	0.906	0.562	2.366	0.188
σ	0.016	0.013	0.006	0.037	0.001
p_{00}	0.946	0.951	0.878	0.978	0.008
p_{11}	0.138	0.133	0.054	0.244	0.025
PIN	0.060	0.055	0.023	0.129	0.001
		Group 2: Medium-volume assets			
$\alpha_{0,A1}$	0.397	0.360	0.272	0.774	0.058
$\alpha_{1,A1}$	2.443	2.356	1.631	3.827	0.367
$\alpha_{0,A2}$	1.193	1.195	0.944	1.398	0.049
$\alpha_{1,A2}$	5.447	3.964	1.665	20.021	0.835
$\alpha_{0,A3}$	0.019	0.023	0.001	0.042	0.010
$\alpha_{1,A3}$	0.823	0.801	0.459	1.443	0.196
σ	0.029	0.019	0.009	0.062	0.001
p_{00}	0.937	0.933	0.918	0.959	0.016
p_{11}	0.159	0.143	0.082	0.240	0.058
PIN	0.070	0.071	0.049	0.095	0.003
		Group 3: Low-volume assets			
$\alpha_{0,A1}$	0.419	0.427	0.236	0.627	0.112
$\alpha_{1,A1}$	3.300	3.067	1.381	5.641	3.190
$\alpha_{0,A2}$	1.123	1.120	0.766	1.419	0.162
$\alpha_{1,A2}$	4.719	3.971	0.395	13.440	2.206
$\alpha_{0,A3}$	0.027	0.029	-0.010	0.058	0.022
$\alpha_{1,A3}$	0.644	0.828	0.007	1.570	0.270
σ	0.056	0.023	0.003	0.351	0.006
p_{00}	0.921	0.939	0.794	0.967	0.095
p_{11}	0.142	0.140	0.000	0.348	0.110
PIN	0.083	0.070	0.038	0.185	0.004

The table summarizes the parameter estimates (average value, median, maximum, minimum and average asymptotic standard error) from estimating the *Order I* model (see equation (5) and description in the main text) and the inferred probabilities of informed negotiation (PIN). The subindex 0 and 1 refers to the normal and excited state respectively, and the subindeces A1, A2 and A3 refer to the degree of aggressiveness.

Including the systematic effect due to the different types of orders has a severe impact on the inferred dynamics of the hidden Markov chain. The probabilities of transition show a very persistent non-excited state, specially in the group of active assets, and a fugacious excited state. The overall effect on the PIN estimates is a dramatic reduction across the three volume groups –the inferred PIN measures are on average halved and now range between 5.5% for frequently-traded assets and 7% for more illiquid assets. It is shown that the PIN dynamics are extremely sensitive to the specification of the mean equation.

There are several meaningful implications that arise from this analysis. The most important one is related to the consequences of the severe misspecification implied in the basic model. While it is not clear that the two states fully correspond to liquidity- and informed-initiated trades, it is evident on the other hand that order design is a key variable for the immediate price revision. The characteristic response proves quite heterogenous depending on the type of order, yet this feature is explicitly neglected in the basic specification. Because information is identified through large relative revisions, and the systematic chances due to the different degrees of aggressiveness are not acknowledged, what does it prevent the model to switch simply as a function of the arrival of different types of orders? Thus, the latent variable of the model would implicitly account for the heterogenous response in the updating process attributable to the degrees of aggressiveness.

The basis of the econometric formulation is so little restrictive that this hypothesis cannot be rejected at first sight –aggressiveness is able to generate differentiated enough responses in terms of 'small' and 'large' midquote changes. But, more important, the inferred probability measures would be related to some microstructure patterns commonly associated to information because so is aggressiveness. For instance, the basic model could find a smaller PIN measure for liquid assets because just because so is the frequency of aggressive orders arrivals, as observed in Griffiths *et al.* (2000).[11] Also, the inferred probability could exhibit meaningful intraday patterns, as reported in the studies of Nyholm, because the likelihood of observing different types of order also tends to exhibit intraday patterns (see, for instance, Chung, Van Ness, and Van Ness, 1999).

[11] Recall that average PIN for small, medium and high volume portfolios reported are 12.8%, 16.5% and 17.8%. The average probability of occurrence of A2 orders for those categories is 11.3%, 14.3% and 15.90%

If this conjecture is true, the unconditional PIN measure from the basic model should be closely related to the probability of observing aggressive orders that impose larger revisions, as the contrast between the two states is then more evident. It turns out illuminating to compare the cross-sectional PIN estimates from the basic model to the sample probability of A2 orders. The latter probability, say $\Pr(A2)$, is just computed as the average value of the indicator variable $D_{\mathbf{A2},t}$. Both measures are depicted in Figure 1. The linear correlation between both series is very high (over 83%) with the PIN estimates nearly matching the path followed by $\Pr(\mathrm{A2})$. Roughly speaking, the inferred measures seem to account for the probability of A2 occurrence plus a random term; this added component is no doubt related to the probability for which orders belonging to other groups, more likely A1, generate a similar response to that from A2 orders.

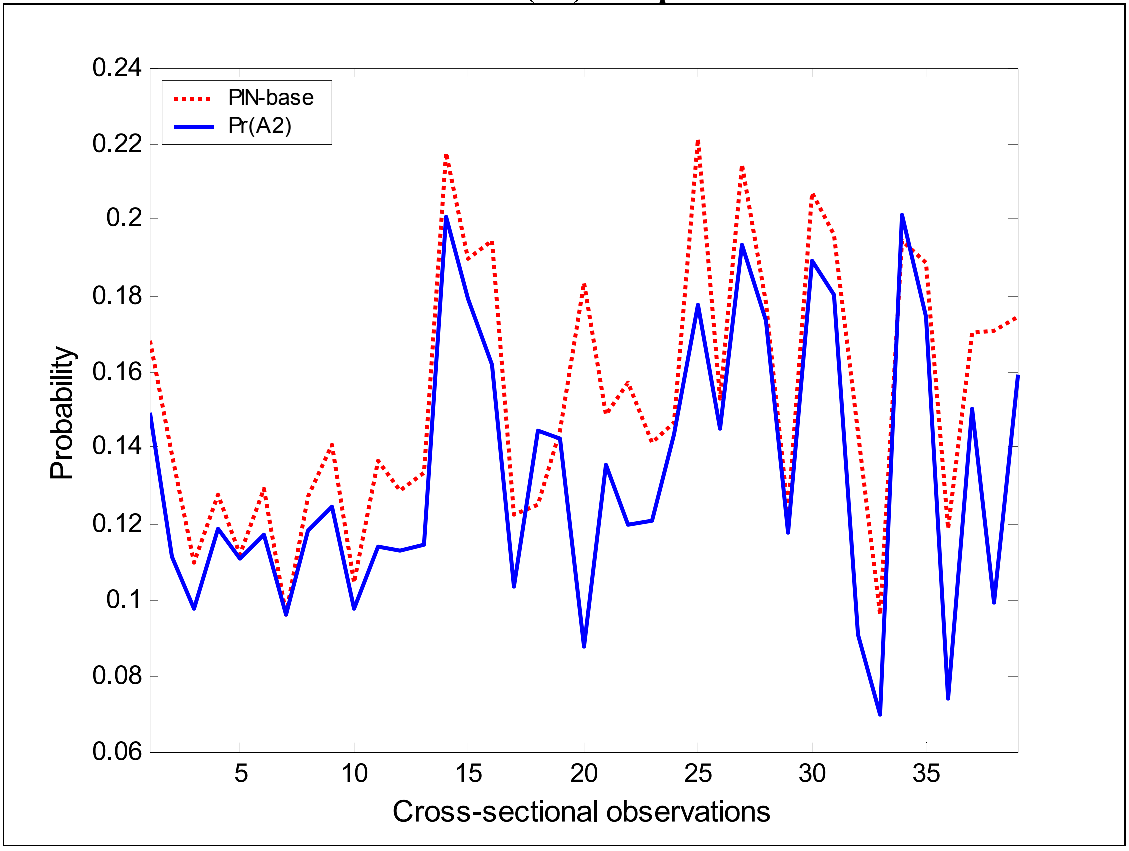
The conclusion is that the probability measure implied from the Nyholm's model basically tells how frequently A2 orders (and orders that generate similar effects) arrive at the market rather than how private information is disclosed. This measure can only be seen as a rough and unreliable estimation of the targeted probability. The relevant questions refer then to the consequences implied when including variables other than aggressiveness, and to what extent a model that (at least) includes aggressiveness would be able to measure properly information arrivals. We provide further results on both questions below.

5.3 Further results

A. Trade size

We analyze the effects of including variables other than aggressiveness in the regime-switching framework. Following Nyholm (2003), we consider three categories of trade size –small, medium and large– depending on a set of threshold points. As the sample includes stocks with quite different trading activities, rather than taking fixed points (e.g., 1000 shares) for all the assets we consider threshold values given by the empirical distribution of trade size for each asset. Thus, we define three indicator variables, $D_{\mathbf{j},t}$,

Figure 1
PIN and Pr(A2) Comparison



Pr(A2) denotes the probability of observing A2 orders for each asset in the sample, computed as the mean value of the indicator variable. PIN-base is the unconditional probability of being in excited state for any asset inferred from the baseline model.

with subscripts belonging to $\mathbf{S} = \{l, m, s\}$ and taking values 1/0 to indicate whether the order at time t traded a volume greater than the fourth quintile ($j = l$), within the first and fourth quintile ($j = m$), and less than the first quintile ($j = s$). The baseline model is then extended by incorporating these categories,

$$\begin{aligned}
\Delta M_t &= \sum_{j \in \mathbf{S}} \alpha_{0,\mathbf{j}} X_{t-1} D_{\mathbf{j},t-1} + \varepsilon_t, && \text{if } I_{t-1} = 0; \\
\Delta M_t &= \sum_{j \in \mathbf{S}} \left(\alpha_{0,\mathbf{j}} + \alpha_{1,\mathbf{j}}\right) X_{t-1} D_{\mathbf{j},t-1} + \varepsilon_t, && \text{if } I_{t-1} = 1;
\end{aligned} \tag{6}$$

As in the former case, imposing the set of linear restrictions $\alpha_{0,j} = \alpha_0$ and $\alpha_{1,j} = \alpha_1$ leads to the baseline formulation.

The outcomes from this model are reported in Table 4. The log-likelihood function increases (see Appendix A) yet not as much as when accounting for order aggressiveness.[12] The results are qualitatively similar to those reported in Nyholm (2003). First and most important, the regime-switching dynamics of the size-extended model do not differ from those of the baseline model, and therefore the resultant PIN estimates are basically the same.[13] A glance at Figure 2, where the cross-sectional PIN estimates from both the baseline and the volume-extended models are exhibited, confirms this feature. Clearly, the resultant probability estimates are still biased by the unconditional likelihood of A2 orders.

Second, the same unappealing pattern related to the coefficients of the size categories is evidenced: while the adjustment seems to be sensitive to size-effects in the normal state, the differences across coefficients are not so

[12] Furthermore, the joint hypothesis of overall equality of coefficients among the indicator categories cannot be rejected at the 1% confident in some assets belonging to the medium and low volume groups.

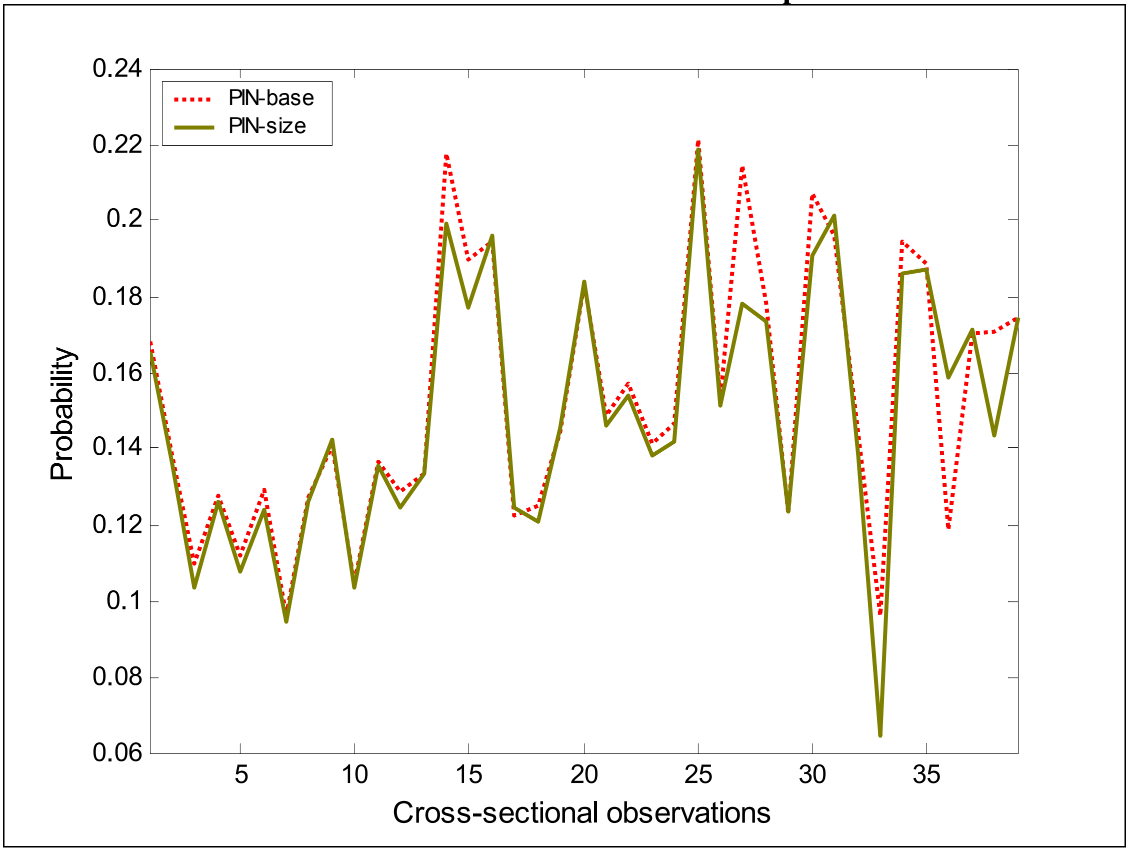
[13] Note that Nyholm (2003, tables 2, 3) reports averaged estimates for p_{00} and p_{11} of 0.907 and 0.096 in the basic model. The respective averaged estimates for the size-extended model are 0.908 and 0.093.

Table 4
Size-extended Model

Parameter	Mean	Median	Minimum	Maximum	Mean St. Error.
		Group 1: High-volume assets			
$\alpha_{0,L}$	0.174	0.174	0.146	0.205	0.017
$\alpha_{1,L}$	1.503	1.449	1.394	1.634	0.080
$\alpha_{0,M}$	0.033	0.035	0.007	0.053	0.008
$\alpha_{1,M}$	1.297	1.270	1.203	1.462	0.045
$\alpha_{0,S}$	0.057	0.057	0.039	0.077	0.007
$\alpha_{1,S}$	1.338	1.306	1.182	1.537	0.059
σ	0.017	0.014	0.001	0.007	0.039
p_{00}	0.906	0.904	0.882	0.924	0.005
p_{11}	0.344	0.343	0.272	0.410	0.020
PIN	0.125	0.126	0.095	0.167	0.001
		Group 2: Medium-volume assets			
$\alpha_{0,L}$	0.166	0.151	0.106	0.278	0.046
$\alpha_{1,L}$	1.593	1.571	1.252	2.346	0.265
$\alpha_{0,M}$	0.073	0.070	0.051	0.115	0.016
$\alpha_{1,M}$	1.263	1.244	1.027	1.499	0.069
$\alpha_{0,S}$	0.040	0.035	-0.008	0.091	0.020
$\alpha_{1,S}$	1.380	1.411	1.119	1.662	0.114
σ	0.031	0.021	0.010	0.068	0.002
p_{00}	0.881	0.890	0.839	0.915	0.015
p_{11}	0.381	0.393	0.282	0.463	0.042
PIN	0.161	0.152	0.121	0.219	0.004
		Group 3: Low-volume assets			
$\alpha_{0,L}$	0.205	0.153	0.095	0.362	0.049
$\alpha_{1,L}$	1.858	1.501	1.187	6.291	0.368
$\alpha_{0,M}$	0.076	0.080	0.010	0.126	0.022
$\alpha_{1,M}$	1.243	1.230	0.947	1.555	0.094
$\alpha_{0,S}$	0.035	0.026	-0.011	0.100	0.031
$\alpha_{1,S}$	1.406	1.347	0.134	2.624	0.177
σ	0.059	0.025	0.004	0.362	0.004
p_{00}	0.888	0.885	0.859	0.957	0.021
p_{11}	0.420	0.438	0.313	0.492	0.072
PIN	0.161	0.173	0.065	0.201	0.006

The table summarizes the parameter estimates (average value, median, maximum, minimum and average asymptotic standard error) from estimating the size-extended model (see equation (6) and description in the main text) and the inferred probabilities of informed negotiation (PIN). The subindex 0 and 1 refers to the normal and excited state respectively, and the subindeces L, M and S refer to the large, medium and small trade size categories.

Figure 2
Baseline and Size-extended Model Comparison



PIN-base and PIN-size denote, respectively, the cross-sectional unconditional probabilities of being in excited state inferred from the baseline and the size-extended models.

conclusive in the excited state. We applied t-tests as in Nyholm (2003) to analyze whether those differences are statistically significant. The results show that the difference between the coefficients related to small and large sizes in the normal state, $(\hat{\alpha}_{0,s} - \hat{\alpha}_{0,l})$, is not significant at the 5% confident in only 4 cases (medium, 3 and small, 1) but, on the contrary, the test on the difference $(\hat{\alpha}_{1,s} - \hat{\alpha}_{1,l})$ in the excited state cannot be rejected in 29 cases.[14]

The conclusions on the relation between private information and trade size based on the evidence from this model are necessarily misleading, because the latent variable is still related to the order typology, not to information arrivals. All the categories related to trade size include aggressive and non-aggressive orders, and consequently the pattern attributable to the different degree of aggressiveness is still present across those categories, thus driving the switching process. It seems clear that this effect will likely apply when accounting for other microstructural effects (e.g., hourly effects) through 1/0 categories still neglecting aggressiveness. The basic conclusion is therefore that any specification of the mean equation in this context should control, at least, for the systematic effect due to the order design, so that the regime-switching scheme can depend on unexpected shocks given this variable. Whether this is sufficient or not, is analyzed in the next section.

B. Extending the mean equation

The general conclusion from the previous analysis seems to be that it is crucial to filter the systematic effects in the mean equation as much as possible so that the latent variable can correspond to information arrivals. This aim suggests a potential conflict between tractability and the need of a precise specification, which could even be unattainable in practice –there are a large number of variables related to market and stock conditions that could be relevant in this analysis. Furthermore, since the framework of the model is intrinsically heuristic, there is no theoretical guidance that helps to identify which variables in particular should be added. The most obvious candidates are the other variables that rely on the investor's decision, namely direction (either buy or sell) and size. We conduct the further analysis on these variables, but the list of potential effects is not necessarily restricted to those variables.

[14]We avoid presenting these statistics for saving space, but they are available upon request.

We analyze several extensions in increasing complexity. To assess to what extent the inferred probabilities are related to information, we consider three basic properties in the spirit of Easley *et al.* (1996). First, the PIN measures should decrease with the trading activity. Our groups of assets show sizeable differences on the degree of liquidity, so the averaged estimated PIN should exhibit a similar pattern. Second, the cross-sectional estimates should be correlated positively with the spread. Spread is partially driven by informational asymmetries, and therefore the inferred PIN should have some predictive power on this variable. We apply a similar procedure than Easley *et al.* (1996) and compute the correlation between the relative average-spread of the assets and the cross-sectional unconditional PIN estimates implied by the different regime-switching models estimated.[15] Third, the PIN estimates should ideally be correlated with other measures approaching the probability of information. There are not alternative procedures to infer the PIN in the trade-to-trade frequency, but the method of Easley *et al.* (1996) provides unconditional estimates for each asset over the period analyzed (PIN_{EKOP} henceforth), so the cross-sectional unconditional probabilities between both procedures could be compared. These procedures are not expected to yield the same point estimates (they are based on very different methodologies and assumptions), but the PIN dynamics inferred from both methods should be correlated positively as they are supposed to approach the same phenomenon. Note that this strategy is the central idea in other empirical studies concerned with the suitability of bid-ask procedures, like those in Van Ness *et al.* (2003) and Chung and Lee (2003).

The simplest formulation corresponds to (5) , which is denoted as *Order I* to distinguish from further generalizations. We first consider a slightly more general classification that splits A3 orders into two subcategories: A3(D) is formed by orders that trade a quantity for the depth at the prevailing quote, and A3(L) include orders that trade a quantity lower than the depth. The reason is that it could be a different systematic effect for orders that consume the available depth over orders that just trade a lower volume, as it is later observed. The A3(L) orders are the least aggressive in the sample and typically provoke only changes of minor importance.[16] The set of categories

[15] The average spread is calculated for any asset as the mean value of the relative spread over the sample period (see Table 1 for descriptive statistics). Bootstrap tests easily reject the hypothesis that the different groups have the same mean average-spread.

[16] Market orders trading below the depth cannot generate an instantaneous change in the midquote. The change of midquotes is measured here between consecutive trades.

related to aggressiveness is now indexed as $\mathbf{A}^* = \{A1, A2, A3(D), A3(L)\}$. We denote the resultant model as Order *II*.

We further generalize this model and firstly account for order direction. Previous literature has evidenced the possibility of asymmetric responses in buy and sells. The resultant model, namely *Order III*, combines aggressiveness and direction (denote $\mathbf{D} = \{b, s\}$ the set of subindexes denoting whether the trade is a buy or a sell) and includes eight different cross-categories and 19 parameters to be estimated. The specification of this model is as follows,

$$\begin{aligned} \Delta M_t &= \sum_{i \in \mathbf{A}^*} \sum_{j \in \mathbf{D}} \alpha_{0,\mathbf{ij}} X_{t-1} D_{\mathbf{ij},t-1} + \varepsilon_t, && \text{if } I_{t-1} = 0; \\ \Delta M_t &= \sum_{i \in \mathbf{A}^*} \sum_{j \in \mathbf{D}} \left(\alpha_{0,\mathbf{ij}} + \alpha_{1,\mathbf{ij}}\right) X_{t-1} D_{\mathbf{ij},t-1} + \varepsilon_t, && \text{if } I_{t-1} = 1; \end{aligned} \tag{7}$$

where the double-indexed indicator variables $D_{\mathbf{ij},t}$ signals the degree of aggressiveness and the direction of the transaction.

On the other hand, we combine trade size and aggressiveness. The role of trade size has been discussed earlier. This model would include twelve cross-categories and require to estimate 27 parameters. Naturally, it would be of great interest the general specification accounting for all these effects simultaneously, but a total of 56 parameters to be estimated in a highly non-linear system simply makes it impracticable. In fact, the model extending for aggressiveness and order size is already troublesome, particularly for the group of less-frequently traded assets. We consider a slightly restricted model instead, namely *Order IV*, in which the size categories do not apply for A3(L) orders. Regardless the trade size, only minor effects are expected for these orders in any case.[17] This model requires of estimating 23 parameters. The specification of the model is as follows,

$$\begin{aligned} \Delta M_t &= \sum_{i \in \mathbf{A}^*} \sum_{j \in \mathbf{S}} \alpha_{0,\mathbf{ij}} X_{t-1} D_{\mathbf{ij},t-1} + \varepsilon_t, && \text{if } I_{t-1} = 0; \\ \Delta M_t &= \sum_{i \in \mathbf{A}^*} \sum_{j \in \mathbf{S}} \left(\alpha_{0,\mathbf{ij}} + \alpha_{1,\mathbf{ij}}\right) X_{t-1} D_{\mathbf{ij},t-1} + \varepsilon_t, && \text{if } I_{t-1} = 1; \end{aligned} \tag{8}$$

Thus, it is possible to observe a change related to a trade initiated by such an order if quotes were posted after the last trade and prior to the order.

[17] We considered a grid of 50 pre-estimated starting values to minimize the posibility of convergence to local extrema, but it seems virtually impossible to guarantee the convergence to the global optimum in such a heavily parameterized system.

where the restriction $\alpha_{0,\mathbf{A3(L)j}} = \alpha_{0,\mathbf{A3(L)}}$ and $\alpha_{1,\mathbf{A3(L)j}} = \alpha_{1,\mathbf{A3(L)}}$ is added in the estimation procedure.

Table 5 shows provides some descriptive statistics about the relative frequency in which aggressiveness, order size and order direction are jointly observed. It can be seen that there is a slight predominance of sell orders as the period was characterized by negative returns (see Table 1 above). The least aggressive orders predominate on the sample and it can be seen that aggressiveness increases as assets are less frequently traded.

The outcomes from estimating models *Order II*, *Order III* and *Order IV* are presented in Tables 6, 7 and 8. Table 9 summarizes the results from applying the three criteria commented above to compare the performance of the different models. It shows the average unconditional PIN estimates from each model –including the baseline and the size-extended model earlier estimated, the boostrap t-test and their p-values for the hypothesis of equality of means among the different groups, and the correlation between the different PIN estimates and both the relative spread and the PIN_{EKOP} measure across the three volume groups.

The estimation of different models yields very different outcomes, underlining once again the extreme sensitiveness of the regime-switching process to the specification of the mean equation. The huge increase in the log-likelihood function as the model tries to reflect new effects show the complexity of the revision process in the regime-switching setting, and the naiveness of the basic formulation. The largest relative increases are shown when splitting the A3 category into two sub-categories and when adding the effect of trade-size (see Appendix A). The effects of considering direction, though relevant, seems of second order.

We firstly comment the basic results evidenced for applying the procedure of Easley *et al.* (1996). The PIN_{EKOP} estimates show a decreasing pattern

Table 5
Sample Frequencies

	A1 Buy	A1 Sell	A2 Buy	A2 Sell	A3(D) Buy	A3(D) Sell	A3(L) Buy	A3(L) Sell
				Group 1: High-volume assets				
Small	1.48%	1.77%	1.12%	1.33%	0.71%	0.61%	5.15%	7.83%
Medium	2.13%	2.30%	3.32%	3.23%	2.78%	2.47%	18.45%	25.31%
Large	0.45%	0.43%	1.29%	1.10%	1.24%	1.14%	6.75%	7.60%
Total	***4.06%***	***4.50%***	***5.73%***	***5.66%***	***4.73%***	***4.22%***	***30.35%***	***40.75%***
				Group 2: Medium-volume assets				
Small	1.65%	1.95%	1.53%	1.94%	0.83%	0.77%	4.68%	6.66%
Medium	2.69%	3.02%	3.84%	4.47%	2.86%	2.43%	16.24%	24.42%
Large	0.57%	0.57%	1.26%	1.29%	1.23%	1.20%	5.48%	8.42%
Total	***4.91%***	***5.54%***	***6.63%***	***7.69%***	***4.92%***	***4.40%***	***26.40%***	***39.51%***
				Group 3: Low-volume assets				
Small	1.59%	2.12%	1.33%	2.61%	1.63%	0.69%	4.02%	6.05%
Medium	3.17%	3.52%	3.31%	4.72%	4.48%	2.25%	12.91%	25.58%
Large	0.64%	0.60%	1.22%	1.23%	2.40%	0.94%	4.26%	8.73%
Total	***5.40%***	***6.24%***	***5.86%***	***8.55%***	***8.51%***	***3.89%***	***21.19%***	***40.36%***

The table summarizes the relative frequencies of jointly observing the degree of aggressiveness (A1, A2, A3(D), A3(L)), the direction of the order (buy or sell) and the size of the trade (small, medium, large).

Table 6
Order II Model

Parameter	Mean	Median	Minimum	Maximum	Mean St. Error.
		Group 1: High-volume assets			
$\alpha_{0,A1}$	0.289	0.289	0.166	0.397	0.025
$\alpha_{1,A1}$	1.876	1.782	1.461	2.521	0.151
$\alpha_{0,A2}$	1.190	1.194	1.025	1.280	0.029
$\alpha_{1,A2}$	2.998	2.808	1.857	5.123	0.492
$\alpha_{0,A3(D)}$	0.351	0.357	0.136	0.539	0.033
$\alpha_{1,A3(D)}$	2.363	2.368	1.442	3.770	0.275
$\alpha_{0,A3(L)}$	0.041	0.039	0.028	0.061	0.006
$\alpha_{1,A3(L)}$	-0.111	-0.101	-0.187	-0.071	0.027
σ	0.015	0.012	0.006	0.036	0.000
p_{00}	0.805	0.812	0.732	0.859	0.022
p_{11}	0.157	0.140	0.097	0.288	0.027
PIN	0.188	0.182	0.149	0.254	0.001
		Group 2: Medium-volume assets			
$\alpha_{0,A1}$	0.370	0.326	0.210	0.770	0.051
$\alpha_{1,A1}$	2.318	2.103	1.583	3.748	0.283
$\alpha_{0,A2}$	1.158	1.137	0.932	1.364	0.053
$\alpha_{1,A2}$	2.938	3.134	1.055	4.886	0.681
$\alpha_{0,A3(D)}$	0.316	0.315	0.123	0.516	0.056
$\alpha_{1,A3(D)}$	3.102	2.794	1.153	8.807	0.495
$\alpha_{0,A3(L)}$	0.035	0.033	-0.015	0.067	0.013
$\alpha_{1,A3(L)}$	-0.082	-0.089	-0.503	0.456	0.091
σ	0.027	0.019	0.008	0.059	0.001
p_{00}	0.811	0.814	0.710	0.930	0.035
p_{11}	0.144	0.123	0.063	0.274	0.060
PIN	0.179	0.168	0.078	0.274	0.003
		Group 3: Low-volume assets			
$\alpha_{0,A1}$	0.370	0.389	0.204	0.519	0.069
$\alpha_{1,A1}$	2.743	2.611	1.390	5.478	0.461
$\alpha_{0,A2}$	1.099	1.097	0.738	1.327	0.084
$\alpha_{1,A2}$	4.332	2.401	0.930	13.280	0.642
$\alpha_{0,A3(D)}$	0.190	0.203	-0.066	0.392	0.067
$\alpha_{1,A3(D)}$	2.902	2.256	1.025	6.737	0.531
$\alpha_{0,A3(L)}$	0.033	0.016	0.000	0.141	0.025
$\alpha_{1,A3(L)}$	-0.081	-0.016	-0.461	0.011	0.074
σ	0.053	0.023	0.003	0.329	0.003
p_{00}	0.809	0.829	0.653	0.920	0.057
p_{11}	0.177	0.149	0.000	0.379	0.093
PIN	0.184	0.176	0.096	0.335	0.006

The table summarizes the parameter estimates (average value, median, maximum, minimum and average asymptotic standard error) from estimating the model with four categories of aggressiveness (see description in the main text) and the inferred probabilities of informed negotiation (PIN). The subindex 0 and 1 refers to the normal and excited state respectively, and the subindeces A1, A2 and A3(D) and A3(L) refer to the degree of aggressiveness.

Table 7
Order III Model

	Group 1					Group 2					Group 3				
	Mean	Med.	Min.	Max.	S.E.	Mean	Med.	Min.	Max.	S.E.	Mean	Med.	Min.	Max.	S.E.
$\alpha_{0,A1-B}$	0.298	0.311	0.166	0.405	0.037	0.381	0.347	0.240	0.699	0.071	0.406	0.361	0.215	0.785	0.108
$\alpha_{0,A1-S}$	0.270	0.287	0.196	0.377	0.034	0.370	0.354	0.184	0.806	0.064	0.341	0.320	0.146	0.625	0.091
$\alpha_{1,A1-B}$	1.827	1.800	1.412	2.530	0.190	2.577	2.392	1.377	4.133	0.390	3.409	2.887	1.127	10.039	0.501
$\alpha_{1,A1-S}$	2.077	1.966	1.398	3.665	0.278	2.452	2.198	1.323	3.809	0.352	2.357	2.210	1.171	4.466	0.513
$\alpha_{0,A2-B}$	1.176	1.209	0.925	1.271	0.043	1.130	1.123	0.891	1.355	0.083	1.069	1.020	0.485	1.457	0.141
$\alpha_{0,A2-S}$	1.204	1.213	0.985	1.341	0.033	1.155	1.180	0.951	1.372	0.077	1.094	1.101	0.801	1.270	0.129
$\alpha_{1,A2-B}$	3.513	2.954	1.276	6.966	0.860	2.998	2.533	0.893	6.923	0.759	3.655	2.891	0.415	15.328	0.902
$\alpha_{1,A2-S}$	2.835	2.633	0.745	4.429	0.401	3.883	2.637	0.503	20.796	0.818	5.019	3.480	0.547	18.179	1.209
$\alpha_{0,A3(D)-B}$	0.357	0.341	0.162	0.555	0.040	0.319	0.326	0.163	0.437	0.074	0.200	0.201	-0.004	0.503	0.076
$\alpha_{0,A3(D)-S}$	0.350	0.379	-0.005	0.527	0.072	0.324	0.319	0.047	0.502	0.082	0.241	0.276	-0.159	0.592	0.120
$\alpha_{1,A3(D)-B}$	2.227	2.318	1.462	3.773	0.266	2.770	2.386	1.418	5.055	0.532	3.364	2.557	0.991	8.366	0.744
$\alpha_{1,A3(D)-S}$	2.315	2.137	1.228	3.840	0.386	3.122	2.399	0.890	8.054	0.780	2.336	2.173	0.442	5.442	0.546
$\alpha_{0,A3(L)-B}$	-0.010	-0.007	-0.036	0.000	0.006	0.010	0.006	-0.066	0.145	0.024	0.004	0.003	-0.024	0.046	0.025
$\alpha_{0,A3(L)-S}$	0.040	0.035	0.025	0.067	0.008	0.035	0.042	-0.013	0.146	0.018	0.023	0.023	-0.007	0.075	0.023
$\alpha_{1,A3(L)-B}$	0.167	0.149	0.081	0.297	0.060	0.018	0.012	-0.204	0.283	0.072	0.066	0.000	-0.590	0.806	0.057
$\alpha_{1,A3(L)-S}$	-0.114	-0.114	-0.209	-0.034	0.035	-0.046	-0.041	-0.290	0.283	0.062	-0.015	-0.030	-0.296	0.409	0.115
σ	0.015	0.012	0.006	0.036	0.000	0.027	0.018	0.008	0.059	0.001	0.051	0.021	0.003	0.325	0.003
p_{00}	0.817	0.814	0.749	0.891	0.021	0.794	0.790	0.708	0.886	0.037	0.822	0.851	0.655	0.911	0.050
p_{11}	0.159	0.146	0.091	0.317	0.028	0.143	0.127	0.000	0.457	0.048	0.235	0.186	0.028	0.497	0.093
PIN	0.179	0.173	0.110	0.241	0.001	0.196	0.197	0.111	0.327	0.003	0.187	0.167	0.122	0.360	0.006

The table summarizes the parameter estimates (average value, median, maximum, minimum and asymptotic average standard error) from estimating the model with four categories of aggressiveness and the order direction (see equation (7) and the description in the main text) and the inferred probabilities of informed negotiation (PIN). The subindex 0 and 1 refers to the normal and excited state respectively. The subindeces refer to the degree of aggressiveness (A1,A2, A3(D) and A3(L)) and the direction (buy=B, sell=S) of the order.

Table 8
Order IV Model

	Group 1					Group 2					Group 3				
	MEAN	MED.	MIN.	MAX.	S.E.	MEAN	MED.	MIN.	MAX.	S.E.	MEAN	MED.	MIN.	MAX.	S.E.
$\alpha_{0,A1-S}$	0.213	0.221	0.129	0.269	0.024	0.241	0.236	0.122	0.426	0.048	0.233	0.212	0.100	0.462	0.052
$\alpha_{0,A1-M}$	0.520	0.514	0.351	0.641	0.047	0.488	0.469	0.300	0.734	0.074	0.439	0.446	0.224	0.579	0.088
$\alpha_{0,A1-L}$	1.383	1.238	0.966	2.143	0.185	1.406	1.168	0.685	4.528	0.313	1.307	1.000	0.087	5.530	0.205
$\alpha_{0,A2-S}$	1.059	1.105	0.828	1.156	0.041	0.940	0.959	0.133	1.523	0.059	0.940	1.103	0.070	1.227	0.093
$\alpha_{0,A2-M}$	1.349	1.343	1.252	1.431	0.033	1.265	1.293	1.078	1.497	0.068	1.256	1.243	1.045	1.582	0.092
$\alpha_{0,A2-L}$	1.964	1.884	1.574	2.317	0.126	1.960	1.889	1.216	2.971	0.265	2.106	1.969	1.148	3.175	0.309
$\alpha_{0,A3-S}$	0.209	0.227	-0.068	0.354	0.048	0.171	0.161	0.010	0.285	0.063	0.069	0.061	-0.181	0.400	0.079
$\alpha_{0,A3-M}$	0.428	0.429	0.237	0.644	0.032	0.379	0.406	0.140	0.563	0.079	0.277	0.284	0.054	0.520	0.065
$\alpha_{0,A3-L}$	0.715	0.686	0.466	1.035	0.079	0.641	0.630	0.181	1.321	0.185	0.565	0.663	-0.622	1.367	0.196
$\alpha_{1,A1-S}$	1.489	1.520	1.318	1.636	0.111	1.583	1.583	1.194	2.368	0.224	1.832	1.477	1.136	4.067	0.193
$\alpha_{1,A1-M}$	2.945	2.905	1.664	5.335	0.305	2.935	2.338	1.689	4.329	0.330	3.698	3.386	1.926	10.273	0.392
$\alpha_{1,A1-L}$	16.577	10.879	4.043	64.862	3.616	10.499	7.173	3.257	32.278	1.810	7.600	5.405	-5.468	31.363	0.819
$\alpha_{1,A2-S}$	1.603	1.454	0.988	3.904	0.307	0.965	0.887	0.559	1.573	0.191	0.987	1.126	-1.039	4.443	0.216
$\alpha_{1,A2-M}$	4.708	3.884	2.307	15.659	0.629	3.264	3.406	1.711	6.452	0.689	4.679	3.638	1.881	15.228	0.647
$\alpha_{1,A2-L}$	11.482	12.120	3.650	25.173	2.057	9.460	7.365	2.733	28.673	1.647	14.058	7.027	-0.010	48.481	1.960
$\alpha_{1,A3-S}$	1.504	1.431	0.754	2.352	0.172	1.414	1.268	0.822	2.289	0.324	1.078	1.038	-0.489	1.935	0.235
$\alpha_{1,A3-M}$	2.863	2.755	1.629	4.353	0.290	3.540	2.733	0.994	9.031	0.597	3.255	3.041	1.106	5.138	0.543
$\alpha_{1,A3-L}$	9.520	5.680	2.087	30.231	1.226	10.123	6.119	2.671	28.836	1.801	11.324	9.418	1.820	36.131	1.588
$\alpha_{0,A32}$	-0.011	-0.011	-0.026	-0.003	0.004	-0.017	-0.018	-0.061	0.006	0.011	-0.003	-0.006	-0.033	0.038	0.014
$\alpha_{1,A32}$	0.381	0.395	0.141	0.577	0.083	0.279	0.241	0.050	0.627	0.117	0.247	0.177	-0.190	0.800	0.159
σ	0.014	0.011	0.005	0.033	0.000	0.025	0.017	0.008	0.054	0.001	0.047	0.018	0.003	0.312	0.003
p_{00}	0.903	0.914	0.793	0.945	0.016	0.848	0.842	0.732	0.943	0.034	0.873	0.870	0.768	0.941	0.036
p_{11}	0.158	0.150	0.084	0.293	0.022	0.254	0.204	0.146	0.538	0.055	0.277	0.210	0.131	0.518	0.075
PIN	0.103	0.092	0.057	0.199	0.001	0.172	0.156	0.064	0.328	0.004	0.153	0.167	0.067	0.328	0.007

The table summarizes the parameter estimates (average value, median, maximum, minimum and asymptotic average standard error) from estimating the model with four categories of aggressiveness and the order size (see equation (8) and the description in the main text) and the inferred probabilities of informed negotiation (PIN). The subindex 0 and 1 refers to the normal and excited state respectively. The subindeces refer to the degree of aggressiveness (A1,A2, A3(D) and A3(L)) and the size (small, medium, large) of the order.

Table 9
Model Performance Analysis

PANEL A: CROSS-AVERAGE PIN

Trading Activity	Baseline	Size-ext.	Order1	Order2	Order3	Order4	EKOP
Group 1	12.7%	12.5%	6.0%	18.7%	17.9%	10.3%	17.9%
Group 2	16.5%	16.1%	7.0%	17.9%	19.6%	17.2%	21.8%
Group 3	16.7%	16.1%	7.0%	18.4%	18.7%	15.3%	26.8%

PANEL B: BOOTSTRAP TESTS AND SIGNIFICANCE

t-tests: Mean eq.	Baseline	Size-ext.	Order1	Order2	Order3	Order4	EKOP
Group1 vs Group2	-3.485	-3.451	-1.174	0.420	-0.982	-2.689	-1.560
	[0.00]	*[0.00]*	*[0.08]*	*[0.55]*	*[0.06*]	*[0.00]*	*[0.04]*
Group2 vs Group3	-0.196	0.036	-0.984	-0.198	0.343	0.655	-1.139
	[0.47]	*[0.54]*	*[0.18]*	*[0.50]*	*[0.77]*	*[0.82]*	*[0.09]*
Group1 vs Group3	-3.411	-2.973	-1.618	0.164	-0.39	-2.278	-2.278
	[0.00]	*[0.04]*	*[0.04]*	*[0.55]*	*[0.37]*	*[0.00]*	*[0.00]*

PANEL C: CORRELATIONS WITH SPREAD AND EKOP MEASURE

All groups	Baseline	Size-ext.	Order1	Order2	Order3	Order4	EKOP
SPREAD	25.8%	21.5%	-71.1%	7.2%	0.8%	26.7%	42.7%
EKOP	-20.5%	0.6%	-1.7%	2.6%	9.6%	15.4%	100%
Groups 1&2							
SPREAD	31.2%	29.9%	-71.1%	16.1%	1.4%	64.2%	54%
EKOP	-33.7%	-31.2%	-43.4%	-7.8%	7.7%	30.0%	100%

The table summarizes different results for all the estimated models. *Baseline* is the basic model in (2); *Size-ex* is the size-extended model in (6); *Order1 to Order4* are the models with i)three aggressiveness categories, ii) four aggressiveness categories, iii) four aggressiveness categories and order direction, iv) four aggressiveness categories and trade size. *EKOP* represents the estimates from the procedure of Easley et al. (1995). Panel A reports the cross-sectional mean values of the unconditional PIN estimates. Panel B tests presents the bootstrapped t-test (p-value between brackets) of the null hypothesis of equal mean values (i.e., the difference between the mean value of the i-th group and the j-th group is zero). Finally, Panel C presents the sample correlations between the PIN estimates of each model and both the relative spread (a time-weighted mean of the spread over the sample period) and the EKOP estimates. The correlation is computed by using the estimates and observations for each groups and excluding Group 3.

across the liquidity measures, and the estimates are correlated positively with the mean spread (42%) as expected.[18] These estimates are comparable in magnitude with those reported in Easley *et al.* (1996). The unsuitability of the baseline model is evidenced again through this analysis, as the inferred probabilities are negatively correlated with the $\mathrm{PIN}_{\mathrm{EKOP}}$ estimates. Intuitively, both measures cannot be approaching the same phenomenon. Models *Order I* and *Order II*, which only include degrees of aggressiveness, also exhibit inconsistent features with the hypothesis of a latent variable related to information arrivals. The PIN measures from Model *Order I* are strongly correlated negatively with the average spread, and those from model *Order II* do not differ significantly across the volume groups. It is remarkable the strong effect that splitting the A3 aggressiveness category into two subcategories has on the dynamics of the regime-switching process. The extension adding trade direction to order aggressiveness in model *Order III* provides gains in likelihood, and it is seen that there are indeed different revisions depending on the direction of the order, but again the PIN estimates do not differ significantly across volume categories. As conclusion, there is not sufficient evidence to consider that the latent variable is identified to information arrivals in any case.

The extension towards acknowledging effects related to aggressiveness and trade size yields some optimistic outcomes. We firstly focus on the results when considering only the high- and medium-volume groups. It is seen that the higher the trade size is, the higher the revision is for any of category of aggressiveness and for any value of the latent variable. The averaged values of the parameter estimates indicate that the major revisions are due to the A1 orders that trade for large volumes, but the differences seems of minor importance when comparing the median. The A2 orders still impose higher revision in the non-excited state and when the order trade small or medium sizes. The A3(L) orders show a negative, (significant in most cases) adjustment in the normal state, and a relatively small, positive revision in excited state.[19] The PIN estimates are on average much lower for assets

[18] We do not present the estimation of the relevant parameters to save space. These results are available upon request. The tests for the equality of mean values across volume groups are always rejected at the 90% confident level, but cannot be rejected in some case for higher levels. As the number of observations are small, the standard error of the test is still seizable. Demanding large confident levels in this context could imply large losses of power.

[19] The negative value mean that some bid (ask) quotes were posted beyond the midquote

in Group 1 and the bootstrap test rejects easily the hypothesis of an equal mean value. Furthermore, the inferred probabilities are highly correlated with the spread (64%), and show some significant degree of correlation with the PIN_{EKOP} estimates (30%). This suggests that including more structure in the mean equation begins to isolate successfully the information arrivals from the microstructure effects.

Now, observe the evidence for the assets in the less-frequently traded group. First, while the volume-effect still applies and the larger the size is the larger the revision, the more passive orders seem to gain power over the most aggressive ones when trading large volumes in excited state. The average values (both mean and median) in the revision for A3(D) orders are higher than those of A1 orders. This feature seems to make little sense. Also, the revision for A2 orders in excited state is (significantly) negative in two cases. Finally, the mean value of the PIN estimates is (not significantly) higher than that from the medium-volume group. The overall effect in relation to the spread and the PIN_{EKOP} consists of a large reduction which halves the correlations with these variables.

This feature shows a worse performance of the estimation procedure on the group of more illiquid assets. Note that there are far less observations available for those assets, and that their dynamics are more likely to exhibit irregular patterns and anomalities such outliers.[20] As the econometric specification involves a large number of parameters in a complex setting, the applicability of the model is not as straightforward as in the baseline case. Also, like in the case of just adding trade size on the basic model, the irregular results evidenced could indicate that some relevant microstructure effects are still neglected in the mean equation and the latent variable does not correspond completely with information. However, it seems unfeasible to overload more than that an already heavily-parameterized model.

value existing at the time of the last trade and prior to the buy (sell) order that later initiated the observed trade. Note that negative values can also be observed in model OrderII (Table 6), though corresponding to the revision in excited state. The latent variable, which determines which observations belong to each state, cannot measure the same thing over the two models.

[20]Note in Table 5 that the small proportion of A1 orders trading for a large volume over the total (less than 1%) implies a small number of observations to infer the parameters. For instance, the mean sample size for Group 3 is 1200 observations, meaning that only 12 observations are available on average for that category.

6 Concluding remarks

The main aim of this paper has been to provide insight on the performance of the procedure put forward by Nyholm (2002, 2003) for estimating the probability of informed trades on the basis of quote-to-quote information. The basic model assumes a very simple specification that allows for a feasible estimation, but which nevertheless proves to be an over-simplification. Conclusions based on that formulation are necessarily misleading.

The main problem is that the procedure infers information arrivals through large immediate revisions in the midquotes and, simultaneously, neglects the heterogenous response of the immediate midquote revision to microstructure variables. The immediate midquote updating is sensitive to information arrivals (permanent component), but also show transient effects, among which aggressiveness plays a major role in the two-staged conception of the model. The inferred measure from the misspecificated model are related to the probability of arrival of a type of orders, namely orders that trade volumes larger than the depth and which are not allowed to walk up the book. The reason is that these orders impose on average a larger immediate revision, and so the model implicitly regards this fact as a latent factor. It seems clear that this is not a valid measure of informed trading.

The relevant question is then whether a further extended model based on the regime-switching modelling could yield a reasonable approach of the PIN dynamics. The systematic responses due to microstructure effects must be filtered, which suggests add 1/0 categories in an extension of the basic model. However, the price updating process proves very complex in the regime-switching framework and is not sufficient with controlling solely for order aggressiveness to ensure estimates related to private information. The model should ideally account, among other potentially relevant effects, for important factors such as the endogenous variables in the trader's decision. However, the model becomes overparameterized very quickly and it seems impracticable to include further effects properly in addition to those related to aggressiveness and trade size, which prove be key variables in this context.

The method based on the regime-switching model imply a collection of theoretical drawbacks and technical limitations. First, the model inherits some weaknesses from the regression indicator-based modelling, including the exogenous conception of some variables. Second, the PIN estimates prove to be extremely sensitive to the model specification. Third, the model cannot be extended beyond some point to capture some potential relevant effects, as commented before. Furthermore, the extended model combining several categories (such as aggressiveness and trade size) already implies a large number of parameters to be estimated through maximum likelihood in the regime-switching framework. The procedure loses tractability and could turn out inapplicable when there are relatively few observations or if the data are not well-behaved, as it is often the case of the less-frequently traded assets. While it is clear what happens when neglecting the effects related to trade size and aggressiveness, the results evidenced for the extended regime switching model are not completely conclusive. The model successfully assigns less probability of information arrivals to the more liquid assets, and shows certain ability to yield estimates that are correlated with the spread and the PIN estimates from the model by Easley *et al.* (1996), regarded here as proxies of information. However, the PIN estimates do not differ across the medium- and low-volume categories, whereas the indicators of liquidity for those groups are clearly different. The poor results evidenced for the illiquid assets could be due to model misspecification. Though the estimates from an extended model might optimistically be used as a potential proxies of information in the continuous trading framework, some natural caution should be exercised. So far, the approach based on the regime-switching model is the only procedure intended to estimate private information arrivals on a trade-to-trade basis, a topic with relevant empirical applications and of interest for both traders and academics. Further research on this issue is deserved.

References

[1] Ahn, H.J., Cai, J., Hamao, Y. and R. Ho (2002), "The Components of the Bid-Ask Spread in a Limit-order Market: Evidence from the Tokio Stock Exchage", J*ournal of Empirical Finance*, 9, pp. 399-430.

[2] Bagehot, W. (1971), "The only Game in Town", *Financial Analysts Journal*, 27, pp. 1715-1729.

[3] Biais, B., Hillion, P. and C. Spatt, (1995), "An Empirical Analysis of the Limit Order Book and the Order Flow in the Paris Bourse", *Journal of Finance*, 50(5), pp. 1655-1689.

[4] Blume, L., Easley D. and M. O'Hara (1994), "Market Statistics and Technical Analysis: The Role of Volume", *Journal of Finance*, 49, pp. 153-181.

[5] Choi, J.Y., D. Salandro and K. Shastri (1988), "On the Estimation of Bid-Ask Spreads: Theory and Evidence", *Journal of Finance and Quantitative Analysis*, 23, pp. 219-230.

[6] Chung, K.H. and M. Lee (2003), "Adverse-Selection Costs and the Probability of Information-based Trading", *The Financial Reviews*, 38, pp. 257-272.

[7] Chung, K., B. Van Ness, and R. Van Ness, (1999), "Limit Orders and the Bid-Ask Spread," *Journal of Financial Economics,* 53, 255-287.

[8] Cohen, K., S. Maier, R. Schwartz, and D. Whitcomb (1981), "Transaction costs, order placement strategy, and existence of the bid-ask spread". *Journal of Political Economy* 89, pp. 287-305.

[9] Degryse, H.; de Jong, F.; Van Ravenswaaij, M. and G. Wuyts, (2003), "Aggressive Orders and the Resiliency of a Limit Order Market", mimeo.

[10] Dufuor, A. and R. F. Engle (2000), "Time and the Impact of a Trade", *Journal of Finance*, 55(6), pp. 2467-2479.

[11] Easley, D., S. Hvidkjaer, and M. O'Hara (2002), "Is Information Risk a Determinant of Asset Returns?", *Journal of Finance*, 57, pp. 2185-2221.

[12] Easley, D. and M. O'Hara, (1987), "Price, Trade Size and Information in Security Markets", *Journal of Financial Economics*, 19, pp. 69-90.

[13] Easley, D., N. Kiefer, M. O'Hara and J. Paperman (1996), "Liquidity, Information, and Infrequently Traded Stocks", *Journal of Finance*, 51, pp. 1405-1436.

[14] George, T., G. Kaul and M. Nimalendran (1991), "Estimation of the Bid-Ask Spread and its Components: A New Approach". *Review of Financial Studies*, 4(4), pp. 623-656.

[15] Glosten, L.R. and L.E. Harris (1988), "Estimating the Components of the Bid-Ask Spread", *Journal of Financial Economics*, 21, pp. 131-142.

[16] Griffiths, M.D.; Smith, B.F.; Alasdair, D.; Turnbull, S.; White, R.W. (2000), "The Cost and Determinants of Orders Aggressiveness", *Journal of Financial Economics*, 56, pp. 65-88.

[17] Harris, L. and J. Hasbrouck (1996), "Market versus Limit Orders: The SuperDOT Evidence on Order Submission Strategy", *Journal of Financial and Quantitative Analysis*, 31(2), pp. 213-231.

[18] Hasbrouck, J. (1988), "Trades, Quotes and Information", *Journal of Financial Econonomics*, 22, pp. 229.252.

[19] Hasbrouck, J. (1991a), "Measuring the Information Content of Stock Trades", *Journal of Finance*, 46, pp. 179-207.

[20] Hasbrouck, J. (1991b), "The Summary Informativeness of Stock Trades: An Econometric Analysis", *Review of Financial Studies*, 4, pp. 571-595.

[21] Huang, R.D. and H.R. Stoll (1997), "The Components of the Bid-Ask Spread: A General Approach", *Review of Financial Studies*, 10, pp. 995-1034.

[22] Jaffe, J.F. and R.L. Winkler (1976), "Optimal Speculation against an Efficient Market", *Journal of Finance*, 31, pp. 49-61.

[23] Keim, D.B. and A. Madhavan (1995). "Anatomy of the trading process: Empirical evidence on the behavior of institutional traders", *Journal of Financial Economics,* 37, pp. 371-398.

[24] Kempf, A. and O. Korn (1999), "Market Depth and Order Size", J*ournal of Financial Markets* 2, pp. 29–48

[25] Lin, J, G.C. Sanger and G.G. Booth (1995), "Trade Size and Components of the Bid-Ask Spread", *Review of Financial Studies*, 8, pp. 1153-1183.

[26] Madavhan, A.; Cheng, M.; (1997), "In Search for Liquidity: An Analysis of Upstairs and Downstairs Trades", *Review of Financial Studies*, 10, pp. 175-204.

[27] Nyholm, K. (2002), "Estimating the Probability of Informed Trading", *Journal of Financial Research*, 35(4), pp. 485-505.

[28] Nyholm, K. (2003), "Inferring the Private Information Content of Trades: a Regime-Switching Approach", *Journal of Applied Econometrics*, 18, pp. 457-470.

[29] Pascual, R., A. Escribano and M. Tapia (2004), "Adverse Selection Costs, Trading Activity and Liquidity in the NYSE: an Empirical Analysis", *Journal of Banking and Finance*, 28, pp. 107-128.

[30] Ranaldo, A.; (2004), "Order Aggressiveness in Limit Order Book Markets", *Journal of Financial Markets*, 7, pp. 53-74.

[31] Roll, R. (1984), "A Simply Implicit Measure of the Effective Bid-Ask Spread in a Efficient Market", *Journal of Finance*, 39, pp. 1127-1139.

[32] Stoll, H.R. (1989), "Inferring the Components of the Bid-Ask Spread: Theory and Empirical Tests", *Journal of Finance*, 19, pp. 115-134.

[33] Van Ness, B.F.; Van Ness, R.A.; Warr, R.S.; (2001), "How well Do Adverse Selection components Measure Adverse Selection?", *Financial Management*, 30, pp. 77-98.

Appendix: Mean Log-likelihood function

Group 1: High-volume assets													
Asset	***ACE***	***ACR***	***ACS***	***AGS***	***ALB***	***ANA***	***AUM***	***DRC***	***MAP***	***NHH***	***REE***	***SOL***	***VAL***
Sample Size	23091	27832	19137	19714	14641	22745	6190	35401	8234	18319	21349	21512	19716
Baseline	3.332	3.294	2.183	2.694	1.694	1.884	2.325	3.103	1.855	2.635	2.987	2.882	3.385
Size-ext.	3.336	3.301	2.198	2.698	1.702	1.890	2.336	3.108	1.862	2.646	2.993	2.893	3.393
Order I	3.591	3.514	2.312	2.863	1.819	2.051	2.454	3.300	1.983	2.805	3.162	3.060	3.593
Order II	3.648	3.593	2.351	2.902	1.849	2.085	2.485	3.351	2.023	2.858	3.232	3.128	3.660
Order III	3.651	3.595	2.354	2.911	1.851	2.089	2.502	3.356	2.028	2.868	3.233	3.131	3.662
Order IV	3.674	3.649	2.433	2.993	1.923	2.139	2.571	3.392	2.058	2.901	3.279	3.169	3.698
Group 2: Medium-volume assets													
Asset	***AEA***	***AZC***	***AZK***	***CPF***	***CRI***	***ENC***	***PAS***	***PQR***	***SOS***	***TAZ***	***VDR***	***VIS***	***ZOT***
Sample Size	3243	5115	4112	3129	2662	4606	1332	4368	2591	3243	1264	9709	5200
Baseline	1.639	2.245	2.431	2.095	1.095	1.649	1.256	2.806	2.301	2.264	1.176	2.936	2.822
Size-ext.	1.652	2.256	2.433	2.098	1.114	1.651	1.260	2.819	2.306	2.274	1.192	2.833	2.827
Order I	1.825	2.439	2.592	2.260	1.288	1.786	1.383	3.094	2.498	2.438	1.299	3.196	3.033
Order II	1.860	2.495	2.617	2.338	1.321	1.816	1.470	3.160	2.541	2.478	1.328	3.283	3.086
Order III	1.863	2.506	2.635	2.349	1.331	1.820	1.492	3.170	2.578	2.499	1.341	3.285	3.089
Order IV	1.946	2.544	2.674	2.383	1.403	1.902	1.552	3.187	2.571	2.517	1.385	3.336	3.149
Group 3: Low-volume assets													
Asset	***ASA***	***BAM***	***CAF***	***DGI***	***ENA***	***IBG***	***MCM***	***NEA***	***PAC***	***RIO***	***UBS***	***VWG***	***ZNC***
Sample Size	1070	1739	673	2228	2525	804	789	923	1122	728	938	835	893
Baseline	2.936	3.183	0.941	2.088	1.922	1.425	0.531	2.489	3.305	1.646	3.856	-0.570	2.733
Size-ext.	2.950	3.194	0.946	2.103	1.925	1.443	0.556	2.496	3.308	1.587	3.891	-0.552	2.735
Order I	3.113	3.459	1.091	2.280	2.123	1.589	0.612	2.740	3.581	1.715	4.203	-0.490	2.897
Order II	3.235	3.506	1.151	2.326	2.147	1.525	0.649	2.756	3.661	1.740	4.274	-0.430	2.945
Order III	3.248	3.521	1.195	2.334	2.165	1.694	0.678	2.847	3.680	1.826	4.276	-0.418	2.965
Order IV	3.304	3.592	1.204	2.416	2.250	1.713	0.815	2.947	3.732	1.988	4.353	-0.395	3.054

The table shows the mean value of the log-likelihood function of the estimated models. *Baseline* is the basic model in (2); *Size-ext* is the size-extended model in (6); *Order I to Order IV* are the models with i) three aggressiveness categories, ii) four aggressiveness categories, iii) four aggressiveness categories and order direction, iv) four aggressiveness categories and trade size.