Information Asymmetries: Three Essays in Market Microstructure

Inaugural-Dissertation
zur Erlangung des Grades eines Doktors
der
Wirtschafts- und Gesellschaftswissenschaften
durch die
Rechts- und Staatswissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität
Bonn

vorgelegt von
Lars Simon Zehnder
aus Köln

Bonn 2015

DEKAN:	PROF. DR. RAINER HÜTTEMANN
ERSTREFERENT:	PROF. DR. HENDRIK HAKENES
ZWEITREFERENT:	PROF. DR. ERIK THEISSEN

TAG DER MÜNDLICHEN PRÜFUNG: 16TH NOVEMBER 2015

Contents

Acknowledgements

In preparing this thesis, I received support from many people to whom I am grateful to. First of all I wish to express my sincere gratitude to my supervisor Erik Theissen for his enduring and patient support. I learned so much from him in countless, interesting discussions about financial markets and financial modeling. Without his abiding inspiration and his willingness to challenge, this thesis would not have attained its current level. The year 2010 determined a watershed in my doctoral studies where I had to decide between giving up or starting all over. Erik Theissen supported me more than one can expect a doctoral supervisor to, and I could not have imagined a better supervisor for this dissertation.

I further want to thank Hendrik Hakenes who agreed on short notice to act as reviser and I appreciate his unselfish commitment, guidance, and valuable comments.

Furthermore, I want to thank Jörg Breitung for his support. In 2004, I started working in his institute and his dedication to research inspired me. I was able to learn a lot from him about econometric modeling and as a result, got my foundation for research in general. He advocated for my application to the doctoral studies program for economics at the University of Bonn and I am very thankful for his trust in my aptitude. Moreover, I also want to thank Joel Hasbrouck for being my sponsor at the Stern School of Business (NYU). He enabled me to make some of my most important personal and professional experiences which allowed for my extensive growth.

In addition, I owe many thanks to Christian Pigorsch for his countless comments and helpful dicussions. His advice leading to finite mixture distributions and their Bayesian estimation as the central modeling approach in my thesis, was especially crucial. Furthermore, this work profited from discussions with and comments from Joachim Grammig, Oliver Randall, Martin Schletter, and Larissa Bauer. I thank Elizabeth C. Romero for her unselfish support in editing this thesis. I want to express my special thanks to Christoph Wagner to whom I became close to during my studies in the doctoral program at the University of Bonn. Our discussions have been very worthful and motivating for the long way to go. I am very thankful for his honest opinions of my ideas and I was able to learn a lot from him about microeconomics. Finally, I am very indebted to my brother Jens Oliver Zehnder. With his continuous support in reading my thesis and giving his honest opinions and ideas on how to improve it, the quality of my thesis undoubtedly improved. He assisted me when he had more important things to do and I thank him very much for this.

Crucial to the process of developing my thesis was the financial support I received. The Friedrich-Naumann-Stiftung für die Freiheit (FNST) provided both, financial support and an excellent program for political-ethical education. Particularly, I would like to thank my fellow scholars and the administration team of the FNST. I also want to thank my godfather Karl Stamm. Without your constant financial and ideational support, my research would not have

become what it is today. Thank you for always being there for me.

As my research demanded much programming, I want to further thank Steffen Schuler for introducing me to object-oriented programming. He was my best teacher and gave me an unbudgeable understanding of programming. Furthermore, I would like to thank the R community and its core team and package developers' constant efforts, in providing one of the best statistical software to researchers for free. Special thanks goes out to the R-package developers Dirk Eddelbuettel, Romain Francois, Doug Bates, Matthew Dowle, and Arunkumar Srinivasan, for their endless input and patience. I thank the federal state North Rhine-Westphalia for providing me with access to the high-performance cluster at the RWTH Aachen. At the high-performance computing center of the RWTH Aachen, I would like to add my thanks to Christian Terboven, Tim Cramer, Hristo Iliev, Paul Kapinos, Frank Robel, Dirk Schmidl, and Sandra Wienke, for all the great workshops and their patience. I was able to learn so much from them about high-performance computing.

Most importantly, I would like to thank my family. Without you I would never have been able to be where I am right now. I am especially grateful to my parents for giving me so much love and freedom to pursue my goals. You taught me to value education by every act you did and the devotion to your children is endless. Next, I would also like to thank my grandmother, Annemarie Laqua, for her loveful support. Finally, I want to thank my siblings: Anne Nicola Zehnder for being there for me in my lonely moments and for pulling me right back into life when I needed it the most, Meike Cornelia Zehnder for her unconditional love and her trustiness and Ines Zehnder, as the peaceful and balanced woman she is, for showing me to value each moment without a haste.

Finally, I am greatly indebted to my friends. Especially to Lisa Ehlers for the great moments we had together and for her caring, emotional support. To Katrin Blüthner, for her endless fortitude, humor, patience and all she has shared with me. To Jenny Rubio-Braun, for her cordial and enduring friendship. To Michael Jin-Yob Kim who always believed in me and shattered all my doubts. I am very thankful to Tobias Kuhl, for our funny adventures together and for his immovable and unconditional succor in so many aspects of my life. Adriana Romero and Carlos Rubio helped me to settle down in New York City during my research at NYU and I thank them very much for their caring help and their loveful company. Furthermore, I would like to thank Elizabeth, Tania, and Gabriel Romero, and Amaru Landrón-Romero for their cultural-emotional support and for having been a second family to me during my stay in New York City. I am also thankful to Lea Maria Brandes, for our valuable dicussions concerning professional visions and her care. Finally, I would like to thank all of my friends who cared for and thought of me. I thank you for the special moments and the lots of laughter we had together. I am especially thankful for all the helpful and releasing conversations we had. Thank you for you accepting me as I am and always being there for me in your generous and caring ways, even when I might have been too much of a strain in these last years. I am eternally grateful for all of your nurturing and much needed support.

Introduction

Incipient models of economic interaction generally assumed perfect information among the interacting agents. Especially neoclassical economics contain perfect information as a key assumption in their approaches to describe the determination of prices, and income distributions in markets through supply and demand. Therein, market participants are considered to act independently on the basis of full and relevant information. As all market participants have identical information about the value of an asset and as their demand functions depend on the price of the asset, there exists a price at which the market clears and no participant regrets trading afterwards.

Since the seminal papers of Spence (1973), Akerlof (1970) and Rothschild and Stiglitz (1976) we know that information asymmetry can change the outcome of an economic game entirely. In particular, Akerlof (1970) demonstrates that a market might fail altogether from information asymmetry. For these reasons regulatory authorities have a special interest in monitoring markets for information asymmetries, like for example insider activities. In this thesis I reconsiderate several market microstructure models that measure information asymmetries on financial markets.

In the market for 'lemons' in Akerlof (1970) one party knows the true value of a good whereas the other party must act on an expectation of this value. The resulting situation can be determined by a total market breakdown through adverse selection: High-value assets have vanished from the market and low-value assets are offered but not traded, because uninformed participants would always regret trading since they would generally pay too much for a low-value asset.

Stiglitz (2002) characterizes in his nobel prize lecture the sources of information asymmetries. Some of the information asymmetries are inherent for as an insider naturally knows more about a financial asset than any other market participant. Further information evolves from the economic processes. For example: In trading a professional trader or a trading house might trade for years or even decades in certain securities and has acquired special knowledge about a company or is able to collect information from orderflow. Stiglitz (2002) points out that *"[w]hile such information asymmetries inevitably arise, the extent to which they do so and their consequences depend on how the market is structured [...]"* This hypothesis is actually considered as a chief argument in the debate about governmental intervention on security markets. Moreover, it defines the foundation for the field of market microstructure.

"[M]arket microstructure", as O'Hara (1995) states it, *"is the study of process and outcomes of exchanging assets under explicit trading rule[s]."* Whereas the rational expectations literature focuses exclusively on the analysis of equilibria by solving for an equilibrium price, market microstructure is interested in the processes in an economy that coordinate the desires of demanders and

suppliers, so that a price emerges and trade occurs. Furthermore, *"[u]nderlying much of the research in market microstructure is a shared focus on the information implicit in market data, and on the learning process that translates this information into price[s]." (O'Hara (1995))*

While former research in economics considered the price forming process most times as a Walrasian auction, Demsetz (1968) analyzed trading behavior and its consequences on prices in security markets and thereby set the stage for the formal study of market microstructure. The main contribution of Demsetz (1968) is the notice of time dimension in the trading process. Buyers and sellers rarely arrive at the same time at a market due to heterogeneity in their needs. This makes it impossible to find a market clearing price at a given point in time. Demsetz' idea is to introduce a cost for immediacy: Some traders wish to trade immediately, and as long as these traders cannot find a counterpart they must offer a better price. This results in two and not one price characterizing the equilibrium at each point in time: A price to sell and a price to buy. The difference between these two prices is called the 'bid-ask spread'. On most of today's security markets liquidity gets provided through traders and market makers who are willing to offer a free trading option in form of a so-called 'limit order', and it is considered that they provide this option only in exchange for an appropriate premia.[1] [2]

If some market participants are better informed, then trading with them results in a certain loss for an uninformed trader. Starting with an insightful paper by Bagehot (1971), a new theory emerged to model price processes that did not depend on transaction costs (e.g. cost of immediacy), but postulated an important role for information. As a result, this theory finds a spread to exist, even in the absence of transaction costs. Main contributions to this branch of market microstructure theory are Copeland and Galai (1983), Glosten and Milgrom (1985), and Easley and O'Hara (1987).

The one-period model of Copeland and Galai (1983) assumes a monopolistic market maker (a liquidity providing intermediary who is in the middle of all trades) and order arrival processes determined by an exogeneous probabilistic framework. There exist informed and uninformed traders in the market. Uninformed (or liquidity) market participants trade on both sides of the market for reasons left unspecified in the model and informed traders only trade, if they have information, and only on the market side specified by their private information. Since the market maker cannot know which type of trader she encounters, she weights her expected gains and losses by the probabilities of uninformed and informed trading. The bid and ask prices of the market maker then emerge from an optimization over these expected gains and losses.

The market maker's decision problem in Copeland and Galai (1983) is a static one: Trading happens only once, and hence, the market maker simply has to balance gains and losses. Introducing dynamics changes the nature of the order arrival process. This process cannot be exogeneous anymore, as now a single order itself may contain information which should be considered in the price-setting decision of the market maker. Furthermore, the information un-

[1] The free trading option was first noted by Copeland and Galai (1983).

[2] A limit buy order is an order to buy a certain amount of shares at a price not higher as indicated in the order's limit. A limit sell order is defined analogously.

derlying the orderflow can be extracted by uninformed traders during trading. This concept of inferring information from orderflow is developed in Glosten and Milgrom (1985).

To understand the contribution of the paper by Glosten and Milgrom (1985) it is important to see that an extension from the one-period to the multi-period case cannot be accomplished by merely accumulating gains and losses from the one-period solution in Copeland and Galai (1983). What is missing in this framework is that trades themselves could reveal the underlying information and thereby influence the price process.

Glosten and Milgrom (1985) consider the trade direction to be information revealing for the following reason: Informed traders buy the asset in the event of good news and sell it in case they have bad news about it. Therefore, selling to the market maker happens either, because a trader needs liquidity, or, because he received bad news about the asset. The market maker cannot tell which is the reason and minimizes her risk by adjusting her beliefs about the asset's value, conditional on the trade direction. During market hours the market maker receives trades and infers from their direction information, which causes her to permanently update her beliefs, and this causes prices to change. By time, the market maker eventually learns the information contained in trades, and her prices converge to the asset's true value. O'Hara (1995) writes: *"[T]his focus on the learning problem confronting the market maker was a new, and important, direction in market microstructure. [...] This linkage of price setting to underlying asset values meant that the process by which information was impounded into prices could be addressed. This issue, long the focus of both the efficient markets and the rational expectations literature, could now be addressed in the context of the actual mechanisms used to set prices in security market[s]."* In regard to the topic of this thesis, I like to emphasize that the model of Glosten and Milgrom (1985) also describes how information asymmetry dissolves through market participants learning the information contained in the trades of the informed.

Another important result from the Glosten and Milgrom model is that the adverse selection problem, induced by informed traders, can lead to a total market shutdown, similar to the breakdown of the market for 'lemons' in Akerlof (1970). As already mentioned above, the amount of informed trading is related to the size of the spread. The market maker sets a larger spread, if she beares a higher risk of trading with better-informed market participants. In case of a very high amount of informed traders, the spread is chosen so high that it precludes trading at all.

From the theoretical model of Glosten and Milgrom (1985) several approaches evolved to empirically measure information asymmetry and to estimate its influence on security prices. Such empirical testing can be found among others in Hasbrouck (1988), Glosten and Harris (1988), Madhavan et al. (1997), and Huang and Stoll (1997), and is summarized under the term 'trade indicator models'.

In the first part of this thesis I reconsiderate trade indicator models in a joint work with Erik Theissen. More precisely, we reconsiderate the models of Madhavan et al. (1997) and Huang and Stoll (1997). Trade indicator models divide the spread into an adverse selection component

and remaining components.[3] As a byproduct an estimate of the spread becomes available. It is a stylized fact that trade indicator models (e.g. Madhavan, Richardson, and Roomans (1997) and Huang and Stoll (1997)) underestimate the bid-ask spread. We argue that this negative bias is due to an endogeneity problem, which is caused by a negative correlation between the arrival of public information and trade direction. In our sample (the component stocks of the DAX 30 index) we find that the average correlation between these variables is -0.193. We develop modified estimators and show that they yield essentially unbiased spread estimates.

The second and third part of the thesis build an entity and consider another way to measure information asymmetries on financial markets. In a seminal paper Easley et al. (1996) proposed a sequential trade model to estimate the probability of informed trading (PIN). Their multi-period model assumes a market with risk-neutral and fully competitive market makers, informed traders and uninformed liquidity traders. The order arrival processes of informed traders and uninformed traders are described by Poisson processes. Nature chooses between news days and no-news days. In case of a news day, the news can be either bad or good. Liquidity traders trade for reasons exogeneous to the model and therefore on both sides of the market. Informed traders only trade on news days. If the news are bad, they sell the asset, and if the news are good, they buy the asset, i.e., they trade only on one side of the market. The empirical model of Easley, Kiefer, O'Hara and Paperman (from hereon EKOP model) relies on the theoretical ideas presented in Glosten and Milgrom (1985) in that the market maker is assumed to infer information from orderflow and to adjust his beliefs accordingly.

Information asymmetry in the EKOP model is measured differently from information asymmetry in the trade indicator models: The trade indicator models consider the adverse selection component in the spread and thereby follow the assumption that information asymmetry always gets priced. In contrast, the EKOP model estimates order arrival processes and computes the market makers beliefs (see Easley et al. (1996, p.1407)), i.e., the probability of informed trading as derived from the Poisson intensity rates.

In the second part, I present a collective paper with Joachim Grammig and Erik Theissen. We propose a methodology to estimate the probability of informed trading (PIN) that only requires data on the daily number of transactions (but not on the number of buyer-initiated and seller-initiated trades). Because maximum likelihood estimation of the model is problematic we propose a Bayesian estimation approach. We perform extensive simulations to evaluate the performance of our estimator. Our methodology increases the applicability of PIN estimation to situations, in which the data necessary for trade classification is unavailable, or in which trade classification is inaccurate.

In the third part of this thesis, I apply the Bayesian estimation framework from the second part to market data of the U.S. corporate bond market. The U.S. coporate bond market is a highly fragmented market while Easley et al. (1996) base their model on a centralized market structure. My results might resemble this discrepancy.

I use transaction data from the Trade Reporting and Compliance Engine (TRACE) for con-

[3] The remaining component in the Madhavan, Richardson and Roomans model is simply the transitory component. Huang and Stoll also add a component for inventory holding costs.

stituents of the S&P 500 in the first half-year of 2011. As a measurement of information asymmetry I employ the probability of informed trading (PIN) proposed by Easley, Kiefer, O'Hara, and Paperman (1996). In a cross-sectional regression of 4,155 fixed income securities on bond characteristics, market variables, and stock statistics I find that nearly 50% of the variation in PINs is explained. All estimated coefficients conform to expectations. While a comparison of PINs in bond and corresponding equity markets confirms prior findings of lower PINs on more active stock markets, it indicates the reverse for fixed-income securities: Less-frequently traded bonds exhibit lower PINs. These findings accord with there being lower transaction costs on less active bond markets as found by Goldstein, Hotchkiss, and Sirri (2007). However, as news probabilities for bonds from the same issuer and bonds and corresponding stocks differ significantly, I question the appropriateness of traditional models for measuring information asymmetries. The probability of informed trading might not be a suitable measure for highly fragmented markets such as the U.S. corporate bond market.

Understanding the sources of information asymmetries on markets is important for several reasons: First, a general understanding for the creation of information asymmetries has been already emphasized in Stiglitz (2002) and is one of the main goals of economic sciences. Second, any governmental intervention in the structure of financial markets should be based on an understanding of mechanisms and third, as has been demonstrated by Han and Zhou (2013), information asymmetries appear to be priced in interest rates by investors, and therefore an understanding of main influence factors in control of the issuer might help issuers to design a fixed security with bearable costs.

Bibliography

Akerlof, G. A., 1970. The market for 'lemons': quality uncertainty and the market mechanism. Aug 84 (3), 488–500.

Bagehot, W., 1971. The only game in town. Financial Analysts Journal 27 (2), 12–14.

Boehmer, E., Grammig, J., Theissen, E., 2007. Estimating the probability of informed trading—does trade misclassification matter? Journal of Financial Markets 10 (1), 26–47.

Copeland, T. E., Galai, D., 1983. Information effects on the bid-ask spread. the Journal of Finance 38 (5), 1457–1469.

Demsetz, H., 1968. The cost of transacting. The Quarterly Journal of Economics 82 (1), 33–53.

Dias, J. G., Wedel, M., 2004. An empirical comparison of em, sem and mcmc performance for problematic gaussian mixture likelihoods. Statistics and Computing 14 (4), 323–332.

Easley, D., Kiefer, N., O'Hara, M., Paperman, J., 1996. Liquidity, information, and infrequently traded stocks. Journal of Finance 51, 1405–1436.

Easley, D., O'Hara, M., 1987. Price, trade size, and information in securities markets. Journal of Financial economics 19 (1), 69–90.

Frühwirth-Schnatter, S., 2001. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association 96 (453), 194–209.

Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer Series in Statistics. Springer.

Glosten, L., Harris, L., 1988. Estimating the Components of the Bid-Ask Spread. Journal of Financial Economics 21 (1), 123–142.

Glosten, L. R., Milgrom, P. R., 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. Journal of financial economics 14 (1), 71–100.

Han, S., Zhou, X., 2013. Informed bond trading, corporate yield spreads, and corporate default prediction. Management Science 0 (0), 1–20.

Hartigan, J. A., Hartigan, P., 1985. The dip test of unimodality. The Annals of Statistics, 70–84.

Hasbrouck, J., 1988. Trades, quotes, inventories, and information. Journal of Financial Economics 22 (2), 229–252.

Huang, R., Stoll, H., 1997. The components of the bid-ask spread: A general approach. Review of Financial Studies 10 (4), 995–1034.

Kokot, S., 2004. The Econometrics of Sequential Trade Models: Theory and Applications Using High Frequency Data. Springer.

Lee, C., Ready, M. J., 1991. Inferring trade direction from intraday data. The Journal of Finance 46 (2), 733–746.

Madhavan, A., Richardson, M., Roomans, M., 1997. Why do security prices change? A transaction-level analysis of NYSE stocks. Review of Financial Studies 10 (4), 1035–1064.

McLachlan, G., Peel, D., 2000. Finite mixture models. Vol. 299. Wiley-Interscience.

O'Hara, M., 1995. Market microstructure theory. Vol. 108. Blackwell Cambridge.

Redner, R. A., Walker, H. F., 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM review 26 (2), 195–239.

Rothschild, M., Stiglitz, J., 1976. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. The Quarterly Journal of Economics 90 (4), 629–649.

Spence, M., 1973. Job market signaling. The quarterly journal of Economics 87 (3), 355–374.

Stiglitz, J. E., 2002. Information and the change in the paradigm in economics. American Economic Review, 460–501.

Part I.

Estimation of trading costs: Trade indicator models revisited

It is a stylized fact that trade indicator models (e.g. Madhavan, Richardson, and Roomans (1997) and Huang and Stoll (1997)) underestimate the bid-ask spread. We argue that this negative bias is due to an endogeneity problem which is caused by a negative correlation between the arrival of public information and trade direction. In our sample (the component stocks of the DAX 30 index) we find that the average correlation between these variables is -0.193. We develop modified estimators and show that they yield essentially unbiased spread estimates.

1. Introduction

For more than 30 years researchers have developed measures of the bid-ask spread and its components. Trade indicator models (as proposed by Glosten and Harris (1988), Huang and Stoll (1997) and Madhavan et al. (1997)) are a very important and popular class of models. The basic intuition of these models is that (1) because of bid-ask bounce the time series of transaction prices contains information on the size of the spread (as in Roll (1984)) and (2) suppliers of liquidity will adjust their bid and ask prices in response to the information content of trades they observe (as in Glosten and Milgrom (1985)). The data required to estimate a trade indicator model are a sequence of transaction prices and a trade indicator variable which indicates whether a trade was buyer-initiated or seller-initiated.[1]

It is well known that trade indicator models underestimate the actual bid-ask spread. Madhavan et al. (1997) report that their implied spread estimate obtained from their trade indicator model underestimates the spread by approximately one third. Grammig et al. (2006) estimate the Huang and Stoll (1997) trade indicator model for a sample of German firms and find that the implied spread is approximately 20% lower than the actual spread. We perform a similar analysis (using data from Germany) and confirm these results.

In this paper we analyze *why* trade indicator models underestimate the spread. We focus on the models proposed by Madhavan et al. (1997) and Huang and Stoll (1997). We hypothesize, and confirm empirically, that the trade indicator models suffer from an endogeneity problem. New public information is negatively correlated with the trade indicator variable. This negative correlation, in turn, results in a downward bias in the estimated spread and adverse selection component. We propose a modified estimator that accounts for this negative correlation. Application of this estimator largely eliminates the bias. The primary advantage of the modified estimator is that it allows us to identify the source of the bias in implied spreads estimated from trade indicator models. The modified estimator also lends itself to application in empirical research. However, it requires additional data, namely, a time series of quote midpoints.[2]

The bias we document suggests that the results of trade indicator models should be interpreted with care. The importance of our findings derives from the widespread use of trade indicator models in the empirical literature.[3] For example Weston (2000) uses the Huang and

[1] The model proposed by Glosten and Harris (1988) allows the transitory and permanent components of the spread to depend on transaction size. Consequently, data on transaction volume is required for estimation of the model.

[2] In many applications the trade indicator variable is constructed by applying the Lee and Ready (1991) algorithm to trade and quote data. In this case quote data is needed anyway, and the additional data requirement of our modified procedure is not a cause for concern.

[3] We do not intend to provide a complete list of papers that apply trade indicator models. When selecting the papers referred to in the text we confined ourselfes to research published in major journals.

Stoll (1997) model to analyze whether the (then) new Nasdaq order handling rules affected the components of the bid-ask spread. Green and Smart (1999) base their analysis of how an increase in noise trading affects the components of the spread on the Madhavan et al. (1997) model. Hatch and Johnson (2002) use the Madhavan et al. (1997) model to analyze whether mergers among NYSE specialist firms affect the spreads (and its components) of the stocks in which the specialist firms make a market. Chakravarty et al. (2012) compare the bid-ask spreads of trades triggered by intermarket sweep orders and trades triggered by other orders using the Madhavan et al. (1997) model. Brockman and Chung (2001) find that the adverse selection component of the spread, measured by the Huang and Stoll (1997) estimator, increases in periods during which firms repurchase shares.

The application of trade indicator models is not confined to the equity market. Green (2004) uses the Madhavan et al. (1997) model to analyze the impact of trading on bond prices after economic news announcements. Bessembinder et al. (2006) and Han and Zhou (2014) adapt the Madhavan et al. (1997) model to the corporate bond market while Bjønnes and Rime (2005) apply the Huang and Stoll (1997) model to the foreign exchange market.

Trade indicator models are not only used in market microstructure research but also in other areas of finance. Odders-White and Ready (2006) analyze the relation between credit ratings and stock liquidity. One of the liquidity measures they use is the Huang and Stoll (1997) estimator of the adverse selection component. Heflin and Shaw (2000) and Brockman et al. (44) use the Huang and Stoll (1997) model to analyze how block ownership of shares affects the spread and its components. Da et al. (2011) analyze the stock selection ability of fund managers. They find that fund managers are more likely to benefit from trading stocks that are more heavily affected by information risk. One of the measures of information risk that the authors use is the adverse selection component as estimated by the Madhavan et al. (1997) model. Cao et al. (2004) use the Madhavan et al. (1997) model to analyze how lockup expirations after IPOs affect liquidity.

The remainder of the paper is organized as follows. In section 2 we describe our data and present descriptive statistics. The effective spread and price impact we estimate directly from the data serve as benchmarks against which the implied spread and adverse selection component obtained from the trade indicator models are evaluated. In section 3 we present the trade indicator models (both the structural and the corresponding statistical models), explain the endogeneity bias and derive our modified estimator. In this section we also present empirical evidence which supports the presence of the endogeneity bias, and evidence which suggests that the modified estimator can largely alleviate this problem. Section 4 concludes.

2. Data

Our data set includes intraday data on the constituent stocks of the DAX-30 index.[4] These stocks are traded in Xetra, an electronic open limit order book. The sample period is the first quarter of 2004. The data set contains time-stamped transaction prices, the bid-ask spread in effect immediately prior to the transaction, and a trade indicator variable which indicates whether the trade was buyer-initiated (trade indicator = 1) or seller-initiated (-1). Note that the trade indicator variable was provided by the exchange. We thus did not need to apply the Lee and Ready (1991) algorithm to classify trades.

Table A.1 lists the 30 sample stocks. Even though they are the 30 most liquid German stocks, they cover a wide range in terms of market capitalization, trading volume, and transaction frequency. The largest stock (Deutsche Telekom) has more than 20 times the market capitalization of the smallest stock (TUI). Similarly, the most actively traded stock (Deutsche Telekom when trading activity is measured by volume and Allianz when it is measured by the transaction frequency) has more than 27 times the trading volume and more than 6 times the transaction frequency of the least active stock (Fresenius Medical Care). The last column of table A.1 shows the average effective spread measured in Euro-Cent. Spreads range from 1.12 Cents (Deutsche Telekom) to 6.51 Cents (Adidas-Salomon).

Table A.2 shows the average price impact for each stock. The price impact of a transaction t is measured by the product of the trade indicator and the change in the quote midpoint after the transaction. We estimate three alternative versions of the price impact measure. The first (which we denote the immediate price impact) is based on the change in the quote midpoint between transaction t and transaction $(t+1)$, the second (third) is based on the change in the quote midpoint between transaction t and the first transaction recorded at least one minute (five minutes) after transaction t. The immediate price impact is smaller than the 1- and 5-minute price impact in all 30 cases. The 5-minute price impact is larger than the 1-minute price impact for almost all stocks, but the increase is much smaller than the increase from the immediate to the 1-minute price impact. These results suggest that it takes at least one minute for the price impact of a trade to be fully reflected in the quote midpoint.

[4] Grammig et al. (2006) use the same data set.

3. Trade indicator models

As noted in the introduction we consider the trade indicator models proposed by Madhavan et al. (1997) (henceforth MRR) and Huang and Stoll (1997). The main difference between these models is that MRR assume that the trade indicator variable follows a Markov process. The resulting first-order serial correlation of the trade indicator variable is explicitly included in the model. Huang and Stoll (1997), in contrast, implicitly assume that the trade indicator variable is serially uncorrelated.[5] We start with the richer MRR model.

3.1. The model by Madhavan, Richardson and Roomans

3.1.1. Structural model

We consider a market in which suppliers of liquidity (who may be limit order traders or market makers) post bid and ask prices at which other traders can buy or sell assets. We develop the model with an electronic open limit order book (such as Xetra, the trading system from which we draw our sample) in mind. Therefore, we do not allow transactions at prices within the quoted spread. Put differently, we assume throughout the paper that the parameter λ is equal to zero.[6]

The model ticks in transaction time. Every time t a trade occurs, p_t denotes the transaction price and the trade indicator variable $q_t \in \{1, -1\}$ denotes trade direction. q_t is 1 for a buyer-initiated trade and -1 for a seller-initiated trade. Buys and sells are assumed to occur with the same (unconditional) frequency, i.e. the unconditional probability of the events $\{q_t = 1\}$ and $\{q_t = -1\}$ equals $\frac{1}{2}$. Consequently, the unconditional expected value of q_t is $\mathbb{E}[q_t] = 0$ and the variance is $\mathbb{V}[q_t] = 1$.

Denote by μ_t the expected value of the security conditional upon public information. Changes in μ_t depend on (i) new public information arrival modeled by the white noise process $\{u_t\}$ and (ii) (private) information contained in the order flow. In the spirit of Glosten and Milgrom (1985) the surprise in the order flow, $(q_t - \mathbb{E}[q_t \mid q_{t-1}])$, reveals some of the private information held by informed traders. The information content of the surprise component of the order flow is captured by the parameter $\theta_{\text{MRR}} \geq 0$. Thus, $\theta_{\text{MRR}}(q_t - \mathbb{E}[q_t \mid q_{t-1}])$ is the permanent impact of the order flow surprise on the expected value μ_t of the security. μ_t evolves according to

$$\mu_t = \mu_{t-1} + \theta_{\text{MRR}}(q_t - \mathbb{E}[q_t \mid q_{t-1}]) + u_t \,. \tag{3.1}$$

[5] This statement holds for their basic model (equation (5) in Huang and Stoll (1997)). In their three-way decomposition of the spread they allow for serial correlation in the trade indicator variable.

[6] MRR denote by λ the probability that a transaction occurs at a price inside the quoted spread.

Note that this process is a martingale and reduces to a simple random walk, if there were no informed traders (in which case $\theta_{\text{MRR}} = 0$). Suppliers of liquidity post quotes which are assumed to be ex-post rational (or regret-free). Consequently, the bid and the ask price are set conditional upon the next trade being seller-initiated ($q_t = -1$) and buyer-initiated ($q_t = 1$), respectively. Further, the liquidity providers require a compensation for their service. This compensation is assumed to be a constant amount $\phi_{\text{MRR}} \geq 0$ per share. The parameter ϕ_{MRR} models the transitory effect of order flow on prices.[7] The ask and bid prices evolve as

$$p_t^a = \mu_{t-1} + \theta_{\text{MRR}}(1 - \mathbb{E}[q_t \mid q_{t-1}]) + \phi_{\text{MRR}} + u_t \,, \tag{3.2}$$

$$p_t^b = \mu_{t-1} + \theta_{\text{MRR}}(-1 - \mathbb{E}[q_t \mid q_{t-1}]) - \phi_{\text{MRR}} + u_t \,. \tag{3.3}$$

The resulting transaction price is

$$p_t = \mu_t + \phi_{\text{MRR}} q_t + \eta_t \,, \tag{3.4}$$

where $\{\eta_t\}$ is a white-noise error process which, among other things, captures rounding errors caused by the existence of a discrete minimum tick size.[8] Combining equations (3.1) and (3.4) yields

$$p_t = \mu_{t-1} + \theta_{\text{MRR}}(q_t - \mathbb{E}[q_t \mid q_{t-1}]) + \phi_{\text{MRR}} q_t + u_t + \eta_t \,. \tag{3.5}$$

Next we specify the process for the trade indicator variable $\{q_t\}$. Like Madhavan et al. (1997) we assume a general Markov process for trade direction. However, as already noted above, we do not allow for transaction at prices within the quoted spread. The process is characterized by the fixed transition probability $\mathrm{P}[q_t = q_{t-1} \mid q_{t-1}] = \gamma$. If traders split up larger orders into several smaller trades we expect $\gamma > \frac{1}{2}$. Similarly, price continuity rules, trade reporting practices, and other institutional factors may cause γ to deviate from $\frac{1}{2}$.

With this specification the conditional expectation $\mathbb{E}[q_t \mid q_{t-1}]$ in eq. (3.5) is

$$\begin{aligned} \mathbb{E}[q_t \mid q_{t-1} = 1] &= \mathrm{P}[q_t = q_{t-1} \mid q_{t-1} = 1](1) + \mathrm{P}[q_t \neq q_{t-1} \mid q_{t-1} = 1](-1) \\ &= \gamma - (1 - \gamma) \,, \end{aligned} \tag{3.6}$$

$$\begin{aligned} \mathbb{E}[q_t \mid q_{t-1} = -1] &= \mathrm{P}[q_t \neq q_{t-1} \mid q_{t-1} = -1](1) + \mathrm{P}[q_t = q_{t-1} \mid q_{t-1} = -1](-1) \\ &= (1 - \gamma) - \gamma \,. \end{aligned} \tag{3.7}$$

The first-order autocorrelation of the trade indicator variable is $\rho_{\text{MRR}} = \frac{\mathbb{E}[q_t q_{t-1}]}{\mathbb{V}[q_t]} = \gamma - (1 - \gamma)$ and therefore eq. (3.6) and eq. (3.7) can be summarized by

$$\mathbb{E}[q_t \mid q_{t-1}] = \rho_{\text{MRR}} q_{t-1} \,. \tag{3.8}$$

[7] As also pointed out by Madhavan et al. (1997), ϕ_{MRR} may cover order processing costs, inventory holding costs, and possibly also rents earned by the suppliers of liquidity.

[8] Ball and Chordia (2001) proposes a model that takes the rounding of prices onto the tick grid explicitly into account.

Inserting eq. (3.8) into eq. (3.5) and recognizing that (from the first lag of eq. (3.4)) $\mu_{t-1} = p_{t-1} - \phi_{\text{MRR}} q_{t-1} - \eta_{t-1}$ yields the expression

$$\Delta p_t = (\phi_{\text{MRR}} + \theta_{\text{MRR}}) q_t - (\phi_{\text{MRR}} + \rho_{\text{MRR}} \theta_{\text{MRR}}) q_{t-1} + u_t + \Delta \eta_t \, . \tag{3.9}$$

This equation is identical to equation (4) in Madhavan et al. (1997) and is the basis for the empirical analysis.

Effective spread The expected effective spread implied by eq. (3.9) is $2(\phi_{\text{MRR}} + \theta_{\text{MRR}})$. To see this, start from the definition of the effective half-spread,

$$\frac{s_{\text{MRR}}}{2} = q_t (p_t - m_t) \, , \tag{3.10}$$

where m_t denotes the quote midpoint at time t,

$$m_t = \frac{p_t^a + p_t^b}{2} \, . \tag{3.11}$$

Inserting the ask and bid prices eq. (3.2) and eq. (3.3), respectively, and conditioning on q_{t-1} yields

$$\begin{aligned} m_t|_{\{q_{t-1}=1\}} &= \frac{p_t^a|_{\{q_{t-1}=1\}} + p_t^b|_{\{q_{t-1}=1\}}}{2} = \frac{2\mu_{t-1} - 2\rho_{\text{MRR}}\theta_{\text{MRR}} + 2u_t}{2} \\ &= \mu_{t-1} - \rho_{\text{MRR}}\theta_{\text{MRR}} + u_t \, , \end{aligned} \tag{3.12}$$

$$\begin{aligned} m_t|_{\{q_{t-1}=-1\}} &= \frac{p_t^a|_{\{q_{t-1}=-1\}} + p_t^b|_{\{q_{t-1}=-1\}}}{2} = \frac{2\mu_{t-1} + 2\rho_{\text{MRR}}\theta_{\text{MRR}} + 2u_t}{2} \\ &= \mu_{t-1} + \rho_{\text{MRR}}\theta_{\text{MRR}} + u_t \, . \end{aligned} \tag{3.13}$$

which implies

$$m_t = \mu_{t-1} - \rho_{\text{MRR}}\theta_{\text{MRR}} q_{t-1} + u_t \, . \tag{3.14}$$

This expression reflects the fact that the suppliers of liquidity take the serial correlation in the order flow into account when setting their quotes. The effective half-spread conditional on q_{t-1} is

$$\begin{aligned} \frac{s_{\text{MRR}}}{2}|_{\{q_{t-1}=1\}} &= q_t (p_t - m_t) \\ &= q_t (\mu_{t-1} + \theta_{\text{MRR}} q_t - \theta_{\text{MRR}}\rho_{\text{MRR}} + \phi_{\text{MRR}} q_t + u_t + \eta_t - \mu_{t-1} + \rho_{\text{MRR}}\theta_{\text{MRR}} - u_t) \\ &= (\phi_{\text{MRR}} + \theta_{\text{MRR}} + \eta_t) \, , \end{aligned} \tag{3.15}$$

$$\begin{aligned} \frac{s_{\text{MRR}}}{2}|_{\{q_{t-1}=-1\}} &= q_t (p_t - m_t) \\ &= q_t (\mu_{t-1} + \theta_{\text{MRR}} q_t + \theta_{\text{MRR}}\rho_{\text{MRR}} + \phi_{\text{MRR}} q_t + u_t + \eta_t - \mu_{t-1} - \rho_{\text{MRR}}\theta_{\text{MRR}} - u_t) \\ &= (\phi_{\text{MRR}} + \theta_{\text{MRR}} + \eta_t) \, , \end{aligned} \tag{3.16}$$

from which it immediately follows that the expected effective spread implied by the MRR model is

$$s_{\text{MRR}} = 2(\phi_{\text{MRR}} + \theta_{\text{MRR}}) \,. \tag{3.17}$$

3.1.2. Estimation of basic model

The basic model (3.9) is nonlinear in its parameters $(\phi_{\text{MRR}}, \theta_{\text{MRR}}, \rho_{\text{MRR}})'$. To estimate these parameters we follow Madhavan et al. (1997) and use the generalized method of moments (GMM) with moment conditions

$$\mathbb{E}\begin{bmatrix} q_t q_{t-1} - q_t^2 \rho_{\text{MRR}} \\ \xi_t q_t \\ \xi_t q_{t-1} \end{bmatrix} = \mathbf{0} \,, \tag{3.18}$$

where

$$\xi_t = \Delta p_t - (\phi_{\text{MRR}} + \theta_{\text{MRR}}) q_t + (\phi_{\text{MRR}} + \rho_{\text{MRR}} \theta_{\text{MRR}}) q_{t-1} \tag{3.19}$$

is the residual.

We estimate this model for the 30 stocks in our sample. Table A.3 reports, for each sample stock, the structural parameters $(\phi_{\text{MRR}}, \theta_{\text{MRR}}, \rho_{\text{MRR}})'$ and the implied effective spread, $s_{\text{MRR}} = 2(\phi_{\text{MRR}} + \theta_{\text{MRR}})$. The last three columns show the actual spread (taken from Table A.1), the difference between the implied spread and the actual spread in Cents, and the percentage difference between implied and actual spread. We report Newey-West standard errors to account for serial correlation and heteroskedasticity. The results are consistent with those reported by Madhavan et al. (1997) and Grammig et al. (2006). The spread implied by the trade indicator model is systematically lower than the actual spread measured directly from the data, $\bar{s} = \frac{1}{N} \sum_{t=1}^{N} 2 q_t (p_t - m_t)$ where N is the number of trades. The last column reveals that the bias amounts to approximately 20%. The objective of this paper is to analyze why such a bias occurs. We argue that there is an endogeneity problem in eq. (3.9), that is, one of the regressors (q_t and/or q_{t-1}) is correlated with the error term $(u_t + \Delta \eta_t)$. Our argument is best explained in the context of the statistical model corresponding to eq. (3.9) which we derive in the next section.

3.1.3. Statistical model

We search for a statistcal model that corresponds to our structural model (3.9). As pointed out in Hasbrouck (2007) the corresponding statistical model must be able to generate the full joint distribution of the variables under consideration (i.e. price changes and the trade indicator variable). The trade indicator model assumes a zero mean and stationary autocovariance for the stochastic processes $\{\Delta p_t\}$ and $\{q_t\}$. For models of this kind Wold's theorem (see e.g. Brockwell and Davis (2009) and the application to the Roll (1984) model in Hasbrouck (2007)) states that a corresponding moving average (MA) process exists and is of the form

$$x_t = \sum_{j=0}^{\infty} \delta_j \epsilon_{t-j} + \kappa_t \,, \tag{3.20}$$

where $\{\epsilon_t\}$ is a white-noise process, $\delta_0 = 1$ (a normalization), and $\sum_{j=0}^{\infty} \delta_j < \infty$. κ_t is a linearly deterministic process which, in this context, means that it can be predicted arbitrarily well by a linear projection on past observations of x_t. For a purely stochastic process $\kappa_t = 0$, and we are left with a moving average representation.

The structural model in (3.9) contains two stochastic processes, namely the price differences $\{\Delta p_t\}$ and the trade indicator $\{q_t\}$. The interaction of these variables is described by a multivariate linear model,

$$p_t = \mu_t + \phi^{mod}_{\text{MRR}} q_t \,, \qquad \phi^{mod}_{\text{MRR}} \geq 0 \,, \tag{3.21}$$

$$\mu_t = \mu_{t-1} + w_t \,, \tag{3.22}$$

$$w_t = u_t + \theta^{mod}_{\text{MRR}} v_t \,, \qquad \theta^{mod}_{\text{MRR}} \geq 0 \,, \tag{3.23}$$

$$q_t = \rho^{mod}_{\text{MRR}} q_{t-1} + v_t \,, \quad |\rho^{mod}_{\text{MRR}}| < 1 \,, \tag{3.24}$$

where $\{u_t\}$ and $\{v_t\}$ are white-noise processes, and μ_t denotes, as before, the expected value of the security conditional upon public information as of time t. To see that the system of linear equations in (3.21) - (3.24) indeed corresponds to the MRR trade indicator model we take first differences of the price eq. (3.21) and substitute $\Delta\mu_t = w_t$ from (3.22).

$$\begin{aligned}
\Delta p_t &= w_t + \phi^{mod}_{\text{MRR}}(q_t - q_{t-1}) \\
&= u_t + \theta^{mod}_{\text{MRR}} v_t + \phi^{mod}_{\text{MRR}}(q_t - q_{t-1}) \qquad (3.25) \\
&= \theta^{mod}_{\text{MRR}}(q_t - \rho^{mod}_{\text{MRR}} q_{t-1}) + \phi^{mod}_{\text{MRR}}(q_t - q_{t-1}) + u_t \\
&= (\phi^{mod}_{\text{MRR}} + \theta^{mod}_{\text{MRR}}) q_t - (\phi^{mod}_{\text{MRR}} + \rho^{mod}_{\text{MRR}} \theta^{mod}_{\text{MRR}}) q_{t-1} + u_t \,. \qquad (3.26)
\end{aligned}$$

[9]

The final expression is identical to eq. (3.9).

To arrive at an autoregressive moving average (ARMA) vector representation we transform (3.25) a little further:

$$\begin{aligned}
\Delta p_t &= u_t + \theta^{mod}_{\text{MRR}} v_t + \phi^{mod}_{\text{MRR}}(\rho^{mod}_{\text{MRR}} q_{t-1} - q_{t-1} + v_t) \\
&= u_t + (\phi^{mod}_{\text{MRR}} + \theta^{mod}_{\text{MRR}}) v_t + \phi^{mod}_{\text{MRR}}(\rho^{mod}_{\text{MRR}} - 1) q_{t-1} \,,
\end{aligned} \tag{3.27}$$

and use equation (3.27) together with (3.24) for the vector representation,

$$\begin{aligned}
y_t = \begin{bmatrix} \Delta p_t \\ q_t \end{bmatrix} &= \begin{bmatrix} 1 & \phi^{mod}_{\text{MRR}} + \theta^{mod}_{\text{MRR}} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_t \\ v_t \end{bmatrix} + \begin{bmatrix} 0 & \phi^{mod}_{\text{MRR}}(\rho^{mod}_{\text{MRR}} - 1) \\ 0 & \rho^{mod}_{\text{MRR}} \end{bmatrix} \begin{bmatrix} \Delta p_{t-1} \\ q_{t-1} \end{bmatrix} \\
&= \boldsymbol{\psi}_{\text{MRR}} \varepsilon_t + \boldsymbol{\varphi}_{\text{MRR}} y_{t-1} \\
&= \varepsilon^*_t + \boldsymbol{\varphi}_{\text{MRR}} y_{t-1} \,,
\end{aligned} \tag{3.28}$$

[9] The original model derived above also considers an error process $\{\eta_t\}$ accounting for rounding errors, etc. If we want this error term to be included in the statistical model we would add the error process $\{\eta_t\}$ to the price equation (3.21).

where $\boldsymbol{\psi}_{\text{MRR}}\varepsilon_t = \varepsilon_t^*$ is a standardization necessary for estimation of a VARMA model.[10] Define

$$\boldsymbol{\varphi}_{\text{MRR}} = \begin{bmatrix} 0 & \phi_{\text{MRR}}^{mod}(\rho_{\text{MRR}}^{mod} - 1) \\ 0 & \rho_{\text{MRR}}^{mod} \end{bmatrix} = \begin{bmatrix} \varphi_{11} & \varphi_{12} \\ \varphi_{21} & \varphi_{22} \end{bmatrix} \tag{3.29}$$

The vector moving average (VMA) representation of the model in equation (3.28) then is

$$y_t = (\mathbf{I} - \boldsymbol{\varphi}_{\text{MRR}} L)^{-1} \varepsilon_t^* \,, \tag{3.30}$$

where $\mathbf{I}$ is the identity matrix. For $\{y_t\}$ to be a stable process all the roots of the polynomial $(\mathbf{I} - \boldsymbol{\varphi}_{\text{MRR}} L)^{-1}$ must lie outside the unit circle.

Estimation of the VARMA model eq. (3.28) does not yield estimates of $(\phi_{\text{MRR}}^{mod} + \theta_{\text{MRR}}^{mod})$ and $\theta_{\text{MRR}}^{mod}$. However, inspection of the covariance matrix

$$\boldsymbol{\Omega}_{\varepsilon^*}^{\text{MRR}} = \mathbb{E}[\varepsilon^* \varepsilon^{*\prime}] = \begin{bmatrix} \mathbb{V}[u_t + (\phi_{\text{MRR}}^{mod} + \theta_{\text{MRR}}^{mod}) v_t] & \text{Cov}[u_t, v_t] + (\phi_{\text{MRR}}^{mod} + \theta_{\text{MRR}}^{mod})\sigma_v^2 \\ \text{Cov}[u_t, v_t] + (\phi_{\text{MRR}}^{mod} + \theta_{\text{MRR}}^{mod})\sigma_v^2 & \sigma_v^2 \end{bmatrix} \tag{3.31}$$

where σ_v^2 denotes the variance of the white-noise process $\{v_t\}$ yields the following insight. As long as the covariance between the trade innovation v_t and the public information arrival u_t is zero, the effective half-spread $(\phi_{\text{MRR}}^{mod} + \theta_{\text{MRR}}^{mod})$ can be estimated directly from the covariance matrix in eq. (3.31) by dividing the element (1, 2) by the element (2, 2).

However, if this covariance is non-zero, the MRR model suffers from an endogeneity problem. This can be seen from eq. (3.26) using the expression for the trade indicator process, $\{q_t\}$ in eq. (3.24). If $\{u_t\}$ and $\{v_t\}$ are correlated, so are $\{q_t\}$ and $\{u_t\}$. Whether $\{u_t\}$ and $\{v_t\}$ are correlated is an empirical question. Therefore, we now go back to the data in order to obtain an estimator of the covariance between $\{u_t\}$ and $\{v_t\}$.

3.1.4. Estimation of covariances

To obtain an estimator for $(\phi_{\text{MRR}}^{mod} + \theta_{\text{MRR}}^{mod})$ we need an estimate of the variance of the white-noise process $\{v_t\}$ and an estimate of the covariance between $\{u_t\}$ and $\{v_t\}$. Remember that $\{u_t\}$ is public information arrival and $\{v_t\}$ the order flow surprise. To obtain an estimate of the variance of $\{v_t\}$ we can estimate eq. (3.24) by OLS. This regression also provides us with the time series of $\{v_t\}$. In order to estimate the covariance between $\{u_t\}$ and $\{v_t\}$ we also need the time series of $\{u_t\}$. To obtain this time series we first-difference the quote midpoint equation (3.14) and substitute $\Delta\mu_{t-1} = u_{t-1} + \theta_{\text{MRR}}^{mod} v_{t-1}$ from eq. (3.22) and eq. (3.23).

[10] See for instance Lütkepohl (2005, p.448).

$$\begin{aligned}\Delta m_t &= \Delta\mu_{t-1} - \rho^{mod}_{\text{MRR}}\theta^{mod}_{\text{MRR}}\Delta q_{t-1} + \Delta u_t \\ &= -\rho^{mod}_{\text{MRR}}\theta^{mod}_{\text{MRR}}\Delta q_{t-1} + \theta^{mod}_{\text{MRR}}v_{t-1} + u_{t-1} + \Delta u_t \qquad (3.32)\\ &= \alpha_1\Delta q_{t-1} + \alpha_2 v_{t-1} + u_t\ . \qquad (3.33)\end{aligned}$$

We estimate eq. (3.33) by OLS and retain the residuals. Now, equipped with the time series of $\{u_t\}$ and $\{v_t\}$, we can estimate the covariance [11]

$$\widehat{\text{Cov}}[u_t, v_t] = \text{Cov}[\hat{u}_t, \hat{v}_t]\ . \qquad (3.34)$$

We apply this procedure to our data. The results are shown in Table A.4. The covariance between $\{u_t\}$ and $\{v_t\}$ is negative for all 30 sample stocks. Estimates of the correlation range from -0.041 to -0.257 with a mean value of -0.193. These results provide strong evidence that $\{u_t\}$ and $\{v_t\}$ are correlated and that, therefore, the trade indicator model suffers from an endogeneity problem. Economically, the non-zero correlation between $\{u_t\}$ and $\{v_t\}$ implies that the arrival of public information is correlated with the surprise component of the order flow. A possible interpretation of a non-zero correlation is that the suppliers of liquidity make errors when adjusting their quotes to new public information (or are simply adjusting their quotes too slowly). Other traders who observe these errors (or who are faster than the suppliers of liquidity, e.g. high frequency traders) react by submitting orders. These orders will be buy orders ($q_t = 1$) when the actual quote change is too small (u_t smaller than the true price impact of new public information) and will be sell orders ($q_t = -1$) when the actual quote change is too large (u_t larger than the true price impact of new public information). This will then result in a negative correlation between public information arrival and the order flow surprise.

3.1.5. Estimation of VAR model

So far we have shown that the MRR model suffers from an endogeneity problem because $\{u_t\}$ and $\{v_t\}$ and, in consequence, q_t and u_t are correlated. We now want to analyze whether this endogeneity problem is indeed the reason why the implied spread from the trade indicator model underestimates the actual spread. To this end we propose the following three-step estimation procedure to obtain (i) an estimate of the adverse-selection component $\theta^{mod}_{\text{MRR}}$ and (ii) an unbiased estimate of the effective spread.

1. Estimate via least-squares an AR(1) model of the trade indicator variable $\{q_t\}$ (eq. (3.24))

$$q_t = \rho^{mod}_{\text{MRR}} q_{t-1} + v_t\ .$$

[11] Note that $\text{Cov}[u_t, v_t] \neq 0$ does not imply $\text{Cov}[u_t, v_{t-1}] \neq 0$ because the error processes $\{u_t\}$ and $\{v_t\}$ are assumed to be serially uncorrelated. Consequently, estimating eq. (3.33) by OLS yields unbiased estimates.

2. Estimate the following model by OLS (eq. (3.33))

$$\Delta m_t = \alpha_1 \Delta q_{t-1} + \alpha_2 \hat{v}_{t-1} + u_t \ .$$

Use the residuals from this regression together with those from step 1, $\hat{v}_t$, to compute the covariance $\text{Cov}[\hat{u}_t, \hat{v}_t]$.

3. Next, use a maximum likelihood approach (see for example Lütkepohl (2005)) to estimate the VAR model (eq. (3.28))

$$y_t = \varepsilon_t^* + \boldsymbol{\varphi}_{\text{MRR}} y_{t-1} \ .$$

Obtain the variance-covariance matrix of the residuals (eq. (3.31)), $\hat{\boldsymbol{\Omega}}_{\varepsilon^*}^{\text{MRR}}$, and compute the effective spread by inserting the covariance, $\text{Cov}[\hat{u}_t, \hat{v}_t]$, from step 2 into the expression

$$\hat{s}_{\text{MRR}}^{mod} = 2\frac{(\hat{\boldsymbol{\Omega}}_{\varepsilon^*}^{\text{MRR}}(1,2) - \text{Cov}[\hat{u}_t, \hat{v}_t])}{\hat{\boldsymbol{\Omega}}_{\varepsilon^*}^{\text{MRR}}(2,2)} \ .$$

The parameter $\hat{\alpha}_2$ from step 2 provides an estimate of the adverse-selection component $\theta_{\text{MRR}}^{mod}$ and $\frac{\hat{s}_{\text{MRR}}^{mod} - 2\hat{\alpha}_2}{2} = \hat{\phi}_{\text{MRR}}^{mod}$ is an estimate of the transitory component of the spread.

We apply this three-step procedure to our 30 sample stocks.[12] The results are shown in table A.5. All parameters exhibit the expected signs: $\hat{\rho}_{\text{MRR}}^{mod} = \hat{\varphi}_{22} > 0$, $\hat{\alpha}_1 = -\widehat{\rho_{\text{MRR}}^{mod}\theta_{\text{MRR}}^{mod}} < 0$, $\hat{\alpha}_2 = \hat{\theta}_{\text{MRR}}^{mod} > 0$, and $\hat{\varphi}_{12} = \phi_{\text{MRR}}^{mod}\widehat{(\rho_{\text{MRR}}^{mod} - 1)} < 0$. Furthermore, estimates from all three steps are significant at the 1% level. We also tested each polynomial for roots outside the unit circle and found all estimated polynomials to be stable.

The last three columns of table A.5 show three estimates of the effective bid-ask spread. $\bar{s}$, taken from Table A.1 is the spread estimated directly from the data and serves as benchmark, as before. $\hat{s}_{\text{MRR}}^0$ is the implied spread obtained under the assumption that $\text{Cov}[u_t, v_t] = 0$. It is almost identical to the implied spread obtained from the structural MRR model (eq. (3.9)) shown in Table A.3. Most importantly, it exhibits the same 20% downward bias documented earlier. In contrast, the estimate of the effective spread obtained under the assumption that $\text{Cov}[u_t, v_t] = \text{Cov}[\hat{u}_t, \hat{v}_t]$, denoted $\hat{s}_{\text{MRR}}^{mod}$, approximates the actual spread $\bar{s}$ very well. It does not show the downward bias that plagues the MRR implied spread. In fact, $\hat{s}_{\text{MRR}}^{mod}$ is smaller than $\bar{s}$ in 14 cases, larger in 15 cases, and in one case the values (rounded to the third digit) are identical. The mean implied spread is 2.933 which is indeed very close to the average actual spread of 2.955. The largest relative deviation between the implied spread and the actual spread for any individual stock is 3.57% (as compared to an *average* relative deviation of 19.9% for the biased estimator $\hat{s}_{\text{MRR}}^0$). From these results we conclude that our modified estimator yields an unbiased estimate of the effective bid-ask spread.

As noted above, $\hat{\alpha}_2$ shown in table A.5 is an estimate of the adverse-selection component and can be compared to the estimate of θ_{MRR} shown in Table A.3. This comparison reveals that our

[12] All estimations were conducted in `R-3.0.1` using the package `dse` (version `2013.3.2`). See Petris (2010) for further information about the `dse`-package.

modified estimator yields significantly larger estimates of the adverse selection component. In fact, while the θ_{MRR} estimates are similar in magnitude to the immediate price impacts shown in table A.2, the estimates we obtain when using our modified estimator are much closer to the 1-minute and 5-minute price impacts. In contrast, the transitory component obtained when using our estimator (not shown in table A.5 but obtainable using the expression $\frac{\hat{s}^{mod}_{\text{MRR}} - 2\hat{\alpha}_2}{2} = \hat{\phi}^{mod}_{\text{MRR}}$) is similar in magnitude to the MRR estimate of ϕ_{MRR} shown in Table A.3.

3.2. The model by Huang and Stoll

In this section we repeat our analysis for the Huang and Stoll (1997) trade indicator model. As noted earlier, Huang and Stoll (1997) assume that the trade indicator variable is serially uncorrelated. Their model can be derived from the MRR model by setting the autocorrelation of the trade indicator variable, ρ_{MRR} in eq. (3.9), to zero.

$$\begin{aligned}
\Delta p_t &= (\phi_{\text{HS}} + \theta_{\text{HS}})q_t - (\phi_{\text{HS}} + 0 \cdot \theta_{\text{HS}})q_{t-1} + u_t + \Delta\eta_t \\
&= \phi_{\text{HS}}\Delta q_t + \theta_{\text{HS}}q_t + \theta_{\text{HS}}q_{t-1} - \theta_{\text{HS}}q_{t-1} + u_t + \Delta\eta_t \\
&= (\phi_{\text{HS}} + \theta_{\text{HS}})\Delta q_t + \theta_{\text{HS}}q_{t-1} + u_t + \Delta\eta_t \,. \qquad (3.35)
\end{aligned}$$

3.2.1. Estimating the basic model

We stimate the basic Huang and Stoll (1997) model for our 30 sample stocks. The results are shown in Table A.6. All parameter estimates are significant at the 1% level. The effective spread estimates (and, by implication, the bias relative to the effective spread estimated directly from the data) implied by the model are virtually identical to those obtained from the MRR model (table A.3). However, the components of the spread estimated by the Huang and Stoll (1997) model differ from those obtained from the MRR model. The transitory component is smaller and the adverse selection component larger than the corresponding MRR estimates.

3.2.2. Statistical model

We now derive the statistical model corresponding to the Huang and Stoll (1997) structural model. We start from eq. (3.28) and set $\rho^{mod}_{\text{MRR}} = 0$. This results in $q_t = v_t$ and we get the following

VAR representation for the Huang and Stoll model[13]

$$y_t = \begin{bmatrix} \Delta p_t \\ q_t \end{bmatrix} = \begin{bmatrix} 1 & \phi_{\text{HS}}^{mod} + \theta_{\text{HS}}^{mod} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_t \\ v_t \end{bmatrix} + \begin{bmatrix} 0 & -\phi_{\text{HS}}^{mod} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta p_{t-1} \\ q_{t-1} \end{bmatrix} \tag{3.39}$$

$$= \boldsymbol{\psi}_{\text{HS}} \varepsilon_t + \boldsymbol{\varphi}_{\text{HS}} y_{t-1} \, . \tag{3.40}$$

A normalization of the error term $\varepsilon_t^* = \boldsymbol{\psi}_{\text{HS}} \varepsilon_t$ results in the equation system

$$y_t = \varepsilon_t^* + \boldsymbol{\varphi}_{\text{HS}} y_{t-1} \tag{3.41}$$

which can be estimated by maximum likelihood.

As in the VAR model in eq. (3.28) the variance-covariance matrix contains in its element $(1,2)$ the term $\text{Cov}[u_t, v_t] + \phi_{\text{HS}}^{mod} \sigma_v^2$ and in its element $(2,2)$ the variance of the trade indicator variable, σ_v^2 (remember that in the Huang and Stoll (1997) model $v_t = q_t$).

3.2.3. Estimation of covariances

As in the MRR model we wish to estimate the covariance between public information arrival and the order flow surprise (which here is equal to the trade indicator variable because the expected value of q_t is zero). We obtain an estimate of $\{u_t\}$ from the equation for the first difference of the quote midpoint (see equation (3) in Huang and Stoll (1997)),

$$\Delta m_t = \theta_{\text{HS}}^{mod} q_{t-1} + u_t \, . \tag{3.42}$$

Least-squares estimation gives us residuals $\hat{u}_t$ which can be used to calculate an estimate of the covariance between the trade indicator process $\{q_t\}$ and the process of new public information arrival $\{u_t\}$,

$$\widehat{\text{Cov}}[u_t, q_t] = \text{Cov}[\hat{u}_t, q_t] \, . \tag{3.43}$$

Table A.7 shows the covariance estimates that we obtain when we apply this procedure to our data. All covariances are negative and similar in magnitude to the estimates obtained from the MRR model. This confirms our previous evidence that new public information arrival and the trade indicator are negatively correlated.

[13] It can be shown that this VAR model derives directly from the statistical model,

$$p_t = \mu_t + \phi_{\text{HS}}^{mod} q_t \, , \tag{3.36}$$

$$\mu_t = \mu_{t-1} + w_t \, , \tag{3.37}$$

$$w_t = u_t + \theta_{\text{HS}}^{mod} q_t \, , \tag{3.38}$$

and further that this statistical model is identical to the structural model in eq. (3.35).

3.2.4. Estimation of VAR model

Building on the statistical model derived above we now propose a two-step procedure for a modified estimate of the effective spread from the HS model.

1. Estimate via least squares the model

$$\Delta m_t = \theta_{\mathrm{HS}}^{mod} q_{t-1} + u_t \ ,$$

   and use the residuals to compute the covariance $\mathrm{Cov}[\hat{u}_t, q_t]$.

2. Use maximum likelihood to estimate the VAR model (eq. (3.41)),

$$y_t = \varepsilon_t^* + \boldsymbol{\varphi}_{\mathrm{HS}} y_{t-1} \ .$$

   Then take the variance-covariance matrix of the residuals, $\hat{\boldsymbol{\Omega}}_{\varepsilon^*}^{\mathrm{HS}}$, and, by using the covariance $\mathrm{Cov}[\hat{u}_t, q_t]$ from step 1, compute the effective spread as

$$\hat{s}_{\mathrm{HS}}^{mod} = 2\frac{\hat{\boldsymbol{\Omega}}_{\varepsilon^*}^{\mathrm{HS}}(1,2) - \mathrm{Cov}[\hat{u}_t, q_t]}{\hat{\boldsymbol{\Omega}}_{\varepsilon^*}^{\mathrm{HS}}(2,2)} \ .$$

We apply this two-step procedure to our data. The results are summarized in Table A.8. All parameters, without exception, possess the expected sign ($\hat{\theta}_{\mathrm{HS}}^{mod} > 0$ and $\hat{\varphi}_{12} = -\hat{\phi}_{\mathrm{HS}}^{mod} < 0$). The effective spread $\hat{s}_{\mathrm{HS}}^{0}$ estimated under the assumption that $\mathrm{Cov}[u_t, q_t] = 0$ is similar to the implied effective spread $\hat{s}_{\mathrm{HS}}$ from the structural model (for these results see table A.6). When comparing the implied spreads from the two-step procedure for the HS model $\hat{s}_{\mathrm{HS}}^{mod}$ and the estimates from the three-step procedure for the MRR model ($\hat{s}_{\mathrm{MRR}}^{mod}$ in table A.5) we find that the former exhibit a small but systematic negative bias of about 1%. This bias is due to the fact that the Huang and Stoll (1997) model does not take the serial correlation of the trade indicator variable into account.

Note that the adverse selection component estimated by our two-step procedure is simply the slope of a regression of changes in the quote midpoint on the lagged trade indicator variable and is thus identical by definition to the immediate price impact shown in table A.2.

To conclude, the modified Huang and Stoll (1997) model also corrects the bias of the structural model to a large extent, but does not perform as well as the modified Madhavan et al. (1997) model.

4. Conclusion

This paper is motivated by the stylized fact that trade indicator models, such as the popular models by Madhavan et al. (1997) and Huang and Stoll (1997), underestimate the bid-ask

spread. We argue that this negative bias is due to an endogeneity problem. In order to substantiate our claim we develop the statistical models that correspond to the structural Madhavan et al. (1997) and Huang and Stoll (1997) models. The VARMA representation of these models reveals that, in both cases, the spread implied by the model depends on the covariance between public information arrival and the surprise in the trade indicator variable. If this covariance is different from zero the structural models suffer from an endogeneity problem which results in biased spread estimates. We use data for the component stocks of the DAX30 index and the first quarter of 2004 and find that the covariance is negative and substantial (the average correlation is -0.193).

We then develop modified estimators which take the covariance between public information arrival and the surprise component in the order flow explicitly into account. The modified Huang and Stoll (1997) model has a bias of only about 1% (as compared to almost 19% for the original model). The modified Madhavan et al. (1997) is essentially unbiased. A potential drawback of the modified estimator is that it requires additional data, namely, a time series of quote midpoints. In many applications this will not be a cause for concern, though. Estimation of any trade indicator model requires a trade indicator variable. This variable, in turn, is usually obtained by applying the Lee and Ready (1991) algorithm to trade and quote data (as is done in Huang and Stoll (1997) and Madhavan et al. (1997)). Thus, quote data is required anyway.

A. Tables

Table A.1.: **Descriptive statistics.** The table shows the stocks contained in the DAX-30 index together with their ticker symbols, market capitalization (31st December, 2003), trading volume (Q1, 2004), number of transactions (Q1, 2004), and effective spread (Q1, 2004). The latter two columns contain solely trading on the electronic limit order market Xetra.

Stock	Ticker	Market Cap [bio. Euro]	Trading Volume [bio. Euro]	Transactions	Eff. Spread [Euro-Cent]
Adidas-Salomon	adsft	4.12	2.04	62,394	6.51
Altana	altft	6.72	1.98	69,721	3.87
Allianz	alvft	38.51	18.54	288,276	4.86
BASF	basft	25.52	7.96	164,692	2.18
Bayer	bayft	17.09	5.67	153,092	1.71
BMW	bmwft	22.99	5.62	134,603	2.07
Commerzbank	cbkft	9.25	3.40	92,285	1.52
Continental	contft	4.07	1.64	63,703	2.89
Deutsche Boerse	db1ft	4.86	2.28	62,518	3.51
Deutsche Bank	dbkft	38.34	19.78	252,666	2.96
DaimlerChrysler	dcxft	37.80	12.00	211,053	1.98
Deutsche Post	dpwft	18.16	2.80	83,772	1.76
Deutsche Telekom	dteft	61.04	22.42	283,502	1.12
E.ON	eoaft	35.94	10.27	183,284	2.54
Fresenius MC	fmeft	3.96	0.82	39,538	5.29
Henkel	hen3ft	3.68	1.16	44,704	5.05
Hypovereinsbank	hvmft	9.65	6.29	123,296	1.82
Infineon	ifxft	7.99	9.37	178,506	1.20
Lufthansa	lhaft	5.06	2.81	85,964	1.55
Linde	linft	5.09	1.43	57,135	3.50
MAN	manft	3.39	1.77	67,326	2.67
Metro	meoft	11.34	2.48	78,735	3.10
Muenchener Rueck	muv2ft	22.23	13.26	218,564	4.62
RWE	rweft	16.54	6.24	147,573	2.11
SAP	sapft	42.14	11.80	178,952	6.48
Schering	schft	7.79	3.28	97,004	2.89
Siemens	sieft	56.84	20.57	281,927	2.63
Thyssen Krupp	tkaft	8.08	2.42	80,258	1.76
TUI	tuift	2.96	1.68	67,646	2.32
VW	vowft	14.20	6.67	162,360	2.17
Mean		**18.18**	**6.95**	**133,835**	**2.95**

Table A.2.: **Price impacts.** The table shows the price impacts for all stocks in the DAX-30 index where N denotes the number of transactions. Impacts are calculated in three different ways: (1) with the next sequential midquote, (2) the next midquote after 1 minute and (3) the next midquote after 5 minutes. Results were obtained by using `R-3.0.2` and the `xts`-package (`version 0.9.7`).

Stock	Ticker	N	Next Transaction [Euro Cent]	1 Minute [Euro Cent]	5 Minutes [Euro Cent]
Adidas-Salomon	adsft	62,394	1.7841	2.4910	2.9337
Altana	altft	69,721	1.0114	1.4035	1.5530
Allianz	alvft	288,276	1.1330	1.6811	2.1625
BASF	basft	164,692	0.5774	0.9584	1.0396
Bayer	bayft	153,092	0.4265	0.5950	0.5706
BMW	bmwft	134,603	0.5378	0.7609	0.8049
Commerzbank	cbkft	92,285	0.3684	0.4591	0.5281
Continental	contft	63,703	0.7793	1.1076	1.6022
Deutsche Boerse	db1ft	62,518	0.8626	1.2821	1.3380
Deutsche Bank	dbkft	252,666	0.7329	1.1480	1.3767
DaimlerChrysler	dcxft	211,053	0.4688	0.7273	0.8260
Deutsche Post	dpwft	83,772	0.3756	0.4721	0.4951
Deutsche Telekom	dteft	283,502	0.1518	0.2306	0.2811
E.ON	eoaft	183,284	0.6386	1.0643	1.0768
Fresenius MC	fmeft	39,538	1.4475	1.8328	2.1103
Henkel	hen3ft	44,704	1.3155	1.8675	2.3577
Hypovereinsbank	hvmft	123,296	0.4433	0.5670	0.7075
Infineon	ifxft	178,506	0.2160	0.2682	0.2769
Lufthansa	lhaft	85,964	0.3620	0.4901	0.6102
Linde	linft	57,135	1.0203	1.5109	1.9072
MAN	manft	67,326	0.7039	0.9773	1.1019
Metro	meoft	78,735	0.9246	1.3071	1.6146
Muenchener Rueck	muv2ft	218,564	1.1587	1.9262	2.2624
RWE	rweft	147,573	0.5479	0.8294	0.9195
SAP	sapft	178,952	1.6708	2.6361	2.5552
Schering	schft	97,004	0.7274	0.9894	1.2963
Siemens	sieft	281,927	0.6306	1.0125	1.0951
Thyssen Krupp	tkaft	80,258	0.4030	0.4963	0.5726
TUI	tuift	67,646	0.5696	0.7699	0.9512
VW	vowft	162,360	0.5664	0.8709	0.8856
Mean		**133,835**	**0.7519**	**1.0911**	**1.2604**

Table A.3.: **MRR basic model.** The table shows the results from a GMM estimation of the model by Madhavan et al. (1997). Newey-West standard errors are shown below each estimate and the standard error for the arithmetic average is used for the empirical spread. All estimations were performed in the statistical programming language `R-3.0.2` using the package `gmm (version 1.4.5)`.

		Parameters			**Spread**			
Ticker	N	ϕ_{MRR}	θ_{MRR}	ρ_{MRR}	$\hat{s}_{MRR}$	$\bar{s}$	Bias	Rel. Bias
		Std. err.			[Euro-Cent]	[Euro-Cent]	[Euro-Cent]	[%]
adsft	62,394	0.008822	0.016643	0.2079	5.093	6.51	-1.416	-21.8
		2.956e-04	2.887e-04	4.567e-03	6.142e-04	2.619e-04		
altft	69,721	0.006137	0.009613	0.2142	3.150	3.87	-0.722	-18.7
		1.763e-04	1.691e-04	4.671e-03	3.978e-04	1.313e-04		
alvft	288,276	0.009059	0.010127	0.1977	3.837	4.86	-1.027	-21.1
		1.082e-04	9.642e-05	2.675e-03	2.694e-04	1.336e-04		
basft	164,692	0.003308	0.005569	0.2403	1.775	2.18	-0.405	-18.6
		5.978e-05	6.262e-05	3.269e-03	1.388e-04	4.102e-05		
bayft	153,092	0.003203	0.003620	0.1857	1.364	1.71	-0.349	-20.4
		3.997e-05	4.248e-05	3.195e-03	9.296e-05	3.792e-05		
bmwft	134,603	0.003532	0.004885	0.2025	1.683	2.07	-0.386	-18.7
		5.900e-05	6.388e-05	3.496e-03	1.357e-04	4.292e-05		
cbkft	92,285	0.002941	0.002992	0.2058	1.187	1.52	-0.334	-22.0
		4.236e-05	4.499e-05	4.265e-03	9.293e-05	3.864e-05		
contft	63,703	0.003454	0.007834	0.2414	2.258	2.89	-0.636	-22.0
		1.345e-04	1.437e-04	5.027e-03	3.252e-04	1.226e-04		
db1ft	62,518	0.005320	0.009083	0.2679	2.881	3.51	-0.625	-17.8
		1.563e-04	1.698e-04	4.984e-03	3.467e-04	1.174e-04		
dbkft	252,666	0.005337	0.006804	0.2165	2.428	2.96	-0.534	-18.0
		6.752e-05	6.532e-05	2.654e-03	1.623e-04	4.923e-05		
dcxft	211,053	0.003796	0.004317	0.2281	1.623	1.98	-0.354	-17.9
		4.316e-05	4.221e-05	2.893e-03	9.755e-05	3.222e-05		
dpwft	83,772	0.003908	0.003203	0.1983	1.422	1.76	-0.333	-19.0
		5.420e-05	5.359e-05	4.316e-03	1.279e-04	5.583e-05		
dteft	283,502	0.003647	0.001089	0.2242	0.947	1.12	-0.175	-15.6
		1.418e-05	1.269e-05	2.703e-03	2.484e-05	8.167e-06		
eoaft	183,284	0.004137	0.006073	0.2440	2.042	2.54	-0.498	-19.6
		6.414e-05	6.730e-05	3.036e-03	1.461e-04	4.615e-05		
fmeft	39,538	0.006392	0.013644	0.2305	4.007	5.29	-1.278	-24.2
		2.836e-04	3.011e-04	5.760e-03	5.922e-04	2.769e-04		
hen3ft	44,704	0.006125	0.014285	0.2666	4.082	5.05	-0.970	-19.2
		2.633e-04	2.885e-04	5.616e-03	5.649e-04	2.041e-04		
hvmft	123,296	0.003713	0.003486	0.1862	1.440	1.82	-0.378	-20.8
		4.790e-05	4.928e-05	3.508e-03	1.213e-04	4.089e-05		
ifxft	178,506	0.003306	0.001457	0.1992	0.953	1.20	-0.243	-20.3
		2.007e-05	1.854e-05	3.307e-03	4.300e-05	1.373e-05		
lhaft	85,964	0.002983	0.003219	0.2259	1.240	1.55	-0.311	-20.0
		4.588e-05	4.640e-05	4.299e-03	9.561e-05	4.346e-05		
linft	57,135	0.002961	0.010837	0.2594	2.760	3.50	-0.736	-21.1
		1.689e-04	1.883e-04	5.104e-03	3.580e-04	1.316e-04		
manft	67,326	0.003968	0.006794	0.2477	2.152	2.67	-0.515	-19.3
		1.159e-04	1.218e-04	4.831e-03	2.641e-04	8.581e-05		
meoft	78,735	0.003181	0.008790	0.2311	2.394	3.10	-0.710	-22.9
		1.242e-04	1.331e-04	4.447e-03	2.591e-04	9.356e-05		

Table A.3.: *(continued)*

		Parameters			Spread			
Ticker	N	ϕ_{MRR}	θ_{MRR}	ρ_{MRR}	$\hat{s}_{MRR}$	$\bar{s}$	Bias	Rel. Bias
		Std. err.			[Euro-Cent]	[Euro-Cent]	[Euro-Cent]	[%]
muv2ft	218,564	0.008387	0.010666	0.2104	3.810	4.62	-0.814	-17.6
		1.096e-04	1.158e-04	2.775e-03	2.767e-04	8.312e-05		
rweft	147,573	0.003510	0.004820	0.2163	1.666	2.11	-0.439	-20.9
		5.822e-05	6.143e-05	3.359e-03	1.399e-04	4.164e-05		
sapft	178,952	0.010699	0.014314	0.1954	5.003	6.48	-1.474	-22.8
		1.720e-04	1.604e-04	3.015e-03	4.225e-04	1.359e-04		
schft	97,004	0.005012	0.006597	0.2052	2.322	2.89	-0.568	-19.7
		9.979e-05	1.059e-04	3.985e-03	2.412e-04	9.056e-05		
sieft	281,927	0.004987	0.005709	0.2141	2.139	2.63	-0.495	-18.8
		5.224e-05	5.106e-05	2.638e-03	1.211e-04	4.184e-05		
tkaft	80,258	0.003653	0.003511	0.1943	1.433	1.76	-0.324	-18.4
		5.269e-05	5.344e-05	4.141e-03	1.169e-04	4.461e-05		
tuift	67,646	0.003946	0.005369	0.2142	1.863	2.32	-0.457	-19.7
		8.749e-05	8.801e-05	4.716e-03	1.883e-04	8.246e-05		
vowft	162,360	0.003686	0.005070	0.2274	1.751	2.17	-0.424	-19.5
		5.691e-05	5.973e-05	3.333e-03	1.265e-04	4.119e-05		
Mean	**133,835**	**0.004770**	**0.007014**	**0.2199**	**2.357**	**2.95**	**-0.598**	**-19.9**

Table A.4.: **MRR covariances.** The table shows for all stocks in the DAX-30 index the variances of the new public information announcement, u_t, and the trade innovation, v_t, as well as their common covariance and the correlation coefficient. All estimations were performed in the statistical progrmming language `R-3.0.2`.

Ticker	N	σ_u^2	σ_v^2	Cov$[u_t, v_t]$	Corr$[u_t, v_t]$
adsft	62,394	0.0012492	0.957	-0.006303	-0.1823
altft	69,721	0.0003948	0.954	-0.003429	-0.1767
alvft	288,276	0.0102886	0.960	-0.004097	-0.0412
basft	164,692	0.0001225	0.942	-0.002085	-0.1941
bayft	153,092	0.0000633	0.965	-0.001690	-0.2163
bmwft	134,603	0.0001139	0.959	-0.001996	-0.1910
cbkft	92,285	0.0000484	0.958	-0.001687	-0.2478
contft	63,703	0.0002438	0.941	-0.002624	-0.1732
db1ft	62,518	0.0003256	0.928	-0.002735	-0.1573
dbkft	252,666	0.0002077	0.953	-0.002616	-0.1859
dcxft	211,053	0.0000854	0.947	-0.001782	-0.1982
dpwft	83,772	0.0000591	0.961	-0.001508	-0.2002
dteft	283,502	0.0000142	0.950	-0.000838	-0.2282
eoaft	183,284	0.0001564	0.940	-0.002374	-0.1957
fmeft	39,538	0.0008230	0.947	-0.005190	-0.1859
hen3ft	44,704	0.0006795	0.929	-0.003982	-0.1585
hvmft	123,296	0.0000696	0.965	-0.001923	-0.2347
ifxft	178,506	0.0000220	0.959	-0.001180	-0.2569
lhaft	85,964	0.0000497	0.949	-0.001495	-0.2177
linft	57,135	0.0003416	0.932	-0.003078	-0.1725
manft	67,326	0.0001933	0.938	-0.002501	-0.1857
meoft	78,735	0.0002839	0.947	-0.003297	-0.2011
muv2ft	218,564	0.0005249	0.955	-0.004163	-0.1859
rweft	147,573	0.0001126	0.952	-0.002220	-0.2144
sapft	178,952	0.0010630	0.962	-0.006635	-0.2075
schft	97,004	0.0002112	0.958	-0.002633	-0.1851
sieft	281,927	0.0001445	0.954	-0.002388	-0.2034
tkaft	80,258	0.0000585	0.962	-0.001545	-0.2058
tuift	67,646	0.0001189	0.953	-0.001959	-0.1840
vowft	162,360	0.0001254	0.948	-0.002260	-0.2073
Mean	**133,835**	**0.0006065**	**0.951**	**-0.002740**	**-0.1932**

Table A.5.: **MRR estimation results VAR model.** The table shows the results from the three step estimation for each stock of the DAX-30 index. $\hat{s}^0_{\text{MRR}}$ is the spread estimated under the assumption that $\text{Cov}[u_t, v_t] = 0$ and $\hat{s}^{mod}_{\text{MRR}}$ is the spread estimated with an estimate of the empirical $\text{Cov}[u_t, v_t]$. All estimations have been performed in the statistical programming language `R-3.0.2` and for the VAR model the `R`-package `dse (version 2013.3.2)` has been used.

		Step 1	**Step 2**		**Step 3**		**Spread**		
Ticker	N	ρ^{mod}_{MRR}	α_1 $(-\rho^{mod}_{\text{MRR}}\theta^{mod}_{\text{MRR}})$	α_2 $(\theta^{mod}_{\text{MRR}})$	φ_{12}	φ_{22}	$\hat{s}^0_{\text{MRR}}$ [Euro-Cent]	$\hat{s}^{mod}_{\text{MRR}}$ [Euro-Cent]	$\bar{s}$ [Euro-Cent]
adsft	62,394	0.2079	-0.0056873	0.023242	-0.006969	0.2077	5.091	6.408	6.509
		4.570e-03	2.043e-04	3.961e-04	2.058e-04	3.916e-03			
altft	69,721	0.2142	-0.0028208	0.012892	-0.004807	0.2138	3.149	3.867	3.873
		4.674e-03	1.055e-04	2.165e-04	1.155e-04	3.700e-03			
alvft	288,276	0.1977	-0.0018046	0.012811	-0.007251	0.1977	3.837	4.690	4.864
		2.683e-03	8.026e-04	9.563e-04	6.712e-05	1.826e-03			
basft	164,692	0.2403	-0.0017277	0.007509	-0.002512	0.2401	1.775	2.218	2.180
		3.275e-03	4.375e-05	8.166e-05	3.996e-05	2.392e-03			
bayft	153,092	0.1857	-0.0009004	0.005172	-0.002603	0.1857	1.364	1.714	1.714
		3.199e-03	2.840e-05	5.459e-05	3.075e-05	2.511e-03			
bmwft	134,603	0.2025	-0.0013274	0.006701	-0.002811	0.2025	1.683	2.099	2.069
		3.499e-03	4.288e-05	8.669e-05	4.187e-05	2.669e-03			
cbkft	92,285	0.2058	-0.0007368	0.004450	-0.002328	0.2058	1.187	1.539	1.521
		4.274e-03	3.161e-05	5.792e-05	3.364e-05	3.221e-03			
contft	63,703	0.2414	-0.0027634	0.010472	-0.002609	0.2413	2.257	2.814	2.894
		5.034e-03	8.927e-05	1.704e-04	8.901e-05	3.845e-03			
db1ft	62,518	0.2679	-0.0029509	0.011577	-0.003875	0.2678	2.879	3.468	3.506
		4.989e-03	1.164e-04	2.141e-04	1.061e-04	3.853e-03			
dbkft	252,666	0.2165	-0.0017910	0.009157	-0.004176	0.2165	2.428	2.977	2.963
		2.661e-03	4.114e-05	9.033e-05	4.367e-05	1.942e-03			
dcxft	211,053	0.2281	-0.0012290	0.005940	-0.002930	0.2280	1.622	1.998	1.977
		2.905e-03	3.147e-05	5.858e-05	3.021e-05	2.119e-03			
dpwft	83,772	0.1983	-0.0007756	0.004544	-0.003129	0.1982	1.423	1.736	1.755
		4.323e-03	3.427e-05	6.712e-05	4.175e-05	3.386e-03			
dteft	283,502	0.2242	-0.0004575	0.001970	-0.002828	0.2241	0.947	1.123	1.122
		2.707e-03	9.483e-06	1.775e-05	1.229e-05	1.823e-03			
eoaft	183,284	0.2440	-0.0019933	0.008365	-0.003125	0.2439	2.042	2.547	2.540
		3.040e-03	5.111e-05	9.402e-05	4.306e-05	2.265e-03			

Table A.5.: *(continued)*

		Step 1	**Step 2**		**Step 3**		**Spread**		
Ticker	N	ρ_{MRR}^{mod}	α_1 $(-\rho_{\text{MRR}}^{mod}\theta_{\text{MRR}}^{mod})$	α_2 $(\theta_{\text{MRR}}^{mod})$	φ_{12}	φ_{22}	$\hat{s}_{\text{MRR}}^{0}$ [Euro-Cent]	$\hat{s}_{\text{MRR}}^{mod}$ [Euro-Cent]	$\bar{s}$ [Euro-Cent]
fmeft	39,538	0.2305	-0.0041443	0.018533	-0.004886	0.2304	4.006	5.102	5.285
		5.766e-03	1.980e-04	3.492e-04	2.059e-04	4.894e-03			
hen3ft	44,704	0.2666	-0.0047732	0.017840	-0.004489	0.2661	4.080	4.937	5.052
		5.631e-03	1.872e-04	3.370e-04	1.781e-04	4.559e-03			
hvmft	123,296	0.1862	-0.0006756	0.005155	-0.003018	0.1862	1.439	1.838	1.818
		3.524e-03	3.273e-05	7.266e-05	3.631e-05	2.798e-03			
ifxft	178,506	0.1992	-0.0003704	0.002556	-0.002644	0.1996	0.952	1.198	1.196
		3.314e-03	1.494e-05	2.577e-05	1.718e-05	2.317e-03			
lhaft	85,964	0.2259	-0.0007677	0.004425	-0.002304	0.2258	1.240	1.555	1.551
		4.313e-03	3.522e-05	5.756e-05	3.585e-05	3.322e-03			
linft	57,135	0.2594	-0.0037667	0.013863	-0.002163	0.2591	2.757	3.417	3.496
		5.115e-03	1.227e-04	2.164e-04	1.120e-04	4.041e-03			
manft	67,326	0.2477	-0.0022059	0.009229	-0.002971	0.2476	2.152	2.685	2.667
		4.837e-03	8.110e-05	1.562e-04	7.810e-05	3.734e-03			
meoft	78,735	0.2311	-0.0028111	0.011995	-0.002452	0.2308	2.394	3.091	3.104
		4.452e-03	9.332e-05	1.687e-04	8.414e-05	3.468e-03			
muv2ft	218,564	0.2104	-0.0030833	0.014645	-0.006623	0.2104	3.810	4.681	4.625
		2.781e-03	7.524e-05	1.648e-04	7.447e-05	2.091e-03			
rweft	147,573	0.2163	-0.0014486	0.006929	-0.002750	0.2162	1.666	2.132	2.105
		3.364e-03	4.198e-05	8.435e-05	3.991e-05	2.542e-03			
sapft	178,952	0.1954	-0.0040698	0.020747	-0.008577	0.1953	5.001	6.381	6.476
		3.023e-03	1.129e-04	2.249e-04	1.146e-04	2.318e-03			
schft	97,004	0.2052	-0.0017597	0.009021	-0.003979	0.2050	2.321	2.871	2.890
		3.989e-03	6.773e-05	1.354e-04	7.010e-05	3.143e-03			
sieft	281,927	0.2141	-0.0016441	0.007958	-0.003918	0.2140	2.137	2.638	2.634
		2.645e-03	3.356e-05	6.420e-05	3.516e-05	1.840e-03			
tkaft	80,258	0.1943	-0.0009069	0.004923	-0.002941	0.1942	1.433	1.754	1.757
		4.145e-03	3.661e-05	6.452e-05	4.263e-05	3.462e-03			

Table A.5.: *(continued)*

		Step 1	**Step 2**		**Step 3**		**Spread**		
Ticker	N	ρ^{mod}_{MRR}	α_1 $(-\rho^{mod}_{\text{MRR}}\theta^{mod}_{\text{MRR}})$	α_2 $(\theta^{mod}_{\text{MRR}})$	φ_{12}	φ_{22}	$\hat{s}^0_{\text{MRR}}$ [Euro-Cent]	$\hat{s}^{mod}_{\text{MRR}}$ [Euro-Cent]	$\bar{s}$ [Euro-Cent]
tuift	67,646	0.2142	-0.0014879	0.007181	-0.003099	0.2139	1.863	2.273	2.320
		4.723e-03	6.107e-05	1.071e-04	6.375e-05	3.756e-03			
vowft	162,360	0.2274	-0.0013031	0.007023	-0.002842	0.2274	1.751	2.227	2.175
		3.340e-03	4.183e-05	7.897e-05	3.978e-05	2.417e-03			
Mean	**133,835**	**0.2199**	**-0.0020728**	**0.009561**	**-0.003720**	**0.2198**	**2.356**	**2.933**	**2.955**

Table A.6.: **HS basic model.** The table shows estimation results from the model of Huang and Stoll (1997). ϕ_{HS} denotes the transitory component of the spread and θ_{HS} the adverse selection component. All estimations were conducted in `R-3.0.2`.

Ticker	N	ϕ_{HS}	θ_{HS}	$\hat{s}_{\text{HS}}$	$\bar{s}$	Bias	Rel Bias
		Std. err.		[Euro-Cent]	[Euro-Cent]	[Euro-Cent]	[%]
adsft	62,394	0.012280	0.0131824	5.092	6.51	-1.417	-21.8
		3.777e-04	3.035e-04				
altft	69,721	0.008196	0.0075555	3.150	3.87	-0.722	-18.7
		2.310e-04	1.740e-04				
alvft	288,276	0.011062	0.0081262	3.838	4.86	-1.027	-21.1
		1.549e-04	1.162e-04				
basft	164,692	0.004646	0.0042324	1.776	2.18	-0.404	-18.5
		7.721e-05	5.753e-05				
bayft	153,092	0.003873	0.0029494	1.364	1.71	-0.349	-20.4
		5.225e-05	3.926e-05				
bmwft	134,603	0.004519	0.0038961	1.683	2.07	-0.386	-18.7
		7.828e-05	5.722e-05				
cbkft	92,285	0.003557	0.0023784	1.187	1.52	-0.334	-22.0
		5.022e-05	3.903e-05				
contft	63,703	0.005342	0.0059456	2.258	2.89	-0.636	-22.0
		1.822e-04	1.316e-04				
db1ft	62,518	0.007759	0.0066456	2.881	3.51	-0.625	-17.8
		1.948e-04	1.502e-04				
dbkft	252,666	0.006811	0.0053298	2.428	2.96	-0.535	-18.0
		9.462e-05	7.079e-05				
dcxft	211,053	0.004782	0.0033294	1.622	1.98	-0.354	-17.9
		5.521e-05	4.150e-05				
dpwft	83,772	0.004544	0.0025697	1.423	1.76	-0.332	-18.9
		7.577e-05	5.490e-05				
dteft	283,502	0.003889	0.0008461	0.947	1.12	-0.175	-15.6
		1.233e-05	1.216e-05				
eoaft	183,284	0.005622	0.0045905	2.042	2.54	-0.497	-19.6
		8.319e-05	6.241e-05				
fmeft	39,538	0.009539	0.0105000	4.008	5.29	-1.277	-24.2
		3.488e-04	2.758e-04				
hen3ft	44,704	0.009937	0.0104722	4.082	5.05	-0.971	-19.2
		3.337e-04	2.563e-04				
hvmft	123,296	0.004363	0.0028343	1.439	1.82	-0.379	-20.8
		7.936e-05	5.157e-05				
ifxft	178,506	0.003597	0.0011641	0.952	1.20	-0.243	-20.4
		2.234e-05	1.901e-05				
lhaft	85,964	0.003708	0.0024947	1.240	1.55	-0.311	-20.0
		5.083e-05	4.139e-05				
linft	57,135	0.005771	0.0080257	2.759	3.50	-0.737	-21.1
		2.095e-04	1.591e-04				
manft	67,326	0.005656	0.0051081	2.153	2.67	-0.514	-19.3
		1.504e-04	1.105e-04				
meoft	78,735	0.005211	0.0067599	2.394	3.10	-0.710	-22.9
		1.440e-04	1.137e-04				
muv2ft	218,564	0.010630	0.0084203	3.810	4.62	-0.815	-17.6
		1.631e-04	1.120e-04				
rweft	147,573	0.004554	0.0037767	1.666	2.11	-0.439	-20.9
		8.091e-05	5.902e-05				

Table A.6.: *(continued)*

Ticker	N	ϕ_{HS}	θ_{HS}	$\hat{s}_{\text{HS}}$	$\bar{s}$	Bias	Rel Bias
		Std. err.		[Euro-Cent]	[Euro-Cent]	[Euro-Cent]	[%]
sapft	178,952	0.013496	0.0115176	5.003	6.48	-1.474	-22.8
		2.554e-04	1.893e-04				
schft	97,004	0.006365	0.0052438	2.322	2.89	-0.568	-19.7
		1.365e-04	9.889e-05				
sieft	281,927	0.006206	0.0044811	2.137	2.63	-0.497	-18.9
		6.474e-05	5.045e-05				
tkaft	80,258	0.004337	0.0028288	1.433	1.76	-0.324	-18.4
		6.978e-05	5.370e-05				
tuift	67,646	0.005096	0.0042199	1.863	2.32	-0.457	-19.7
		1.077e-04	8.581e-05				
vowft	162,360	0.004839	0.0039164	1.751	2.17	-0.424	-19.5
		6.826e-05	5.336e-05				
Mean	**133,835**	**0.006340**	**0.0054447**	**2.357**	**2.95**	**-0.598**	**-19.9**

Table A.7.: **HS covariances.** The table shows for all stocks in the DAX-30 index the variances of the new public information announcement, u_t, the variance of the trade indicator, q_t, as well as their common covariance and the correlation coefficient. All estimates were performed in the statistical programming language `R-3.0.2`.

Ticker	N	σ_u^2	σ_q^2	$\mathrm{Cov}[u_t, q_t]$	$\mathrm{Corr}[u_t, q_t]$
adsft	62,394	0.00124621	0.9999	-0.006219	-0.17617
altft	69,721	0.00039429	1.0000	-0.003425	-0.17249
alvft	288,276	0.01027869	0.9983	-0.004182	-0.04128
basft	164,692	0.00012235	0.9997	-0.002095	-0.18941
bayft	153,092	0.00006326	0.9994	-0.001696	-0.21329
bmwft	134,603	0.00011382	0.9995	-0.001998	-0.18736
cbkft	92,285	0.00004837	1.0000	-0.001705	-0.24511
contft	63,703	0.00024351	0.9992	-0.002594	-0.16631
db1ft	62,518	0.00032504	0.9999	-0.002751	-0.15262
dbkft	252,666	0.00020771	0.9994	-0.002639	-0.18316
dcxft	211,053	0.00008536	0.9983	-0.001796	-0.19460
dpwft	83,772	0.00005899	0.9999	-0.001521	-0.19809
dteft	283,502	0.00001419	0.9998	-0.000836	-0.22192
eoaft	183,284	0.00015605	1.0000	-0.002380	-0.19051
fmeft	39,538	0.00081994	1.0000	-0.005201	-0.18164
hen3ft	44,704	0.00067751	0.9997	-0.003981	-0.15297
hvmft	123,296	0.00006960	0.9991	-0.001953	-0.23425
ifxft	178,506	0.00002201	0.9975	-0.001195	-0.25491
lhaft	85,964	0.00004965	0.9996	-0.001521	-0.21592
linft	57,135	0.00034056	0.9992	-0.003053	-0.16549
manft	67,326	0.00019291	0.9989	-0.002508	-0.18069
meoft	78,735	0.00028310	0.9999	-0.003292	-0.19568
muv2ft	218,564	0.00052437	0.9989	-0.004159	-0.18173
rweft	147,573	0.00011254	0.9984	-0.002224	-0.20982
sapft	178,952	0.00106215	0.9999	-0.006632	-0.20352
schft	97,004	0.00021102	0.9999	-0.002642	-0.18191
sieft	281,927	0.00014440	1.0000	-0.002396	-0.19937
tkaft	80,258	0.00005840	0.9999	-0.001548	-0.20259
tuift	67,646	0.00011880	0.9980	-0.001963	-0.18028
vowft	162,360	0.00012537	1.0000	-0.002297	-0.20512
Mean	**133,835**	**0.00060567**	**0.9994**	**-0.002747**	**-0.18927**

Table A.8.: **HS VAR model.** The table shows the results for the two-step estimation procedure of the Huang and Stoll (1997) model for each stock of the DAX-30. $\hat{s}^0_{\text{HS}}$ is the spread estimated under the assumption that $\text{Cov}[u_t, q_t] = 0$ and $\hat{s}^{mod}_{\text{HS}}$ is the spread estimated with an estimate of the empirical $\text{Cov}[u_t, q_t]$. All estimations have been performed in the statistical programming language `R-3.0.2` and for the VAR model the `R`-package `dse (version 2013.3.2)` has been used.

		Parameters		**Spreads**		
Ticker	N	θ^{mod}_{HS}	φ_{12}	$\hat{s}^0_{\text{HS}}$	$\hat{s}_2$	$\bar{s}$
		Std. err.	$(-\phi^{mod}_{\text{HS}})$	[Euro-Cent]	[Euro-Cent]	[Euro-Cent]
adsft	62,394	0.01784	-0.01226	5.091	6.33	6.51
		3.006e-04	1.841e-04			
altft	69,721	0.01011	-0.00817	3.148	3.83	3.87
		1.694e-04	1.021e-04			
alvft	288,276	0.01133	-0.01104	3.837	4.67	4.86
		2.001e-04	5.838e-05			
basft	164,692	0.00577	-0.00464	1.775	2.19	2.18
		6.206e-05	3.479e-05			
bayft	153,092	0.00426	-0.00387	1.364	1.70	1.71
		4.403e-05	2.586e-05			
bmwft	134,603	0.00538	-0.00451	1.683	2.08	2.07
		6.639e-05	3.599e-05			
cbkft	92,285	0.00368	-0.00355	1.187	1.53	1.52
		4.612e-05	2.808e-05			
contft	63,703	0.00779	-0.00533	2.257	2.78	2.89
		1.276e-04	8.001e-05			
db1ft	62,518	0.00863	-0.00773	2.879	3.43	3.51
		1.523e-04	9.382e-05			
dbkft	252,666	0.00733	-0.00680	2.428	2.96	2.96
		7.204e-05	3.760e-05			
dcxft	211,053	0.00469	-0.00478	1.622	1.98	1.98
		4.515e-05	2.541e-05			
dpwft	83,772	0.00376	-0.00454	1.423	1.73	1.76
		5.531e-05	3.460e-05			
dteft	283,502	0.00152	-0.00389	0.947	1.11	1.12
		1.441e-05	8.493e-06			
eoaft	183,284	0.00639	-0.00562	2.042	2.52	2.54
		6.743e-05	3.739e-05			
fmeft	39,538	0.01447	-0.00950	4.006	5.05	5.29
		2.615e-04	1.860e-04			
hen3ft	44,704	0.01316	-0.00992	4.080	4.88	5.05
		2.436e-04	1.575e-04			
hvmft	123,296	0.00443	-0.00436	1.439	1.83	1.82
		5.974e-05	3.063e-05			
ifxft	178,506	0.00216	-0.00359	0.952	1.19	1.20
		2.047e-05	1.310e-05			
lhaft	85,964	0.00362	-0.00370	1.240	1.54	1.55
		4.544e-05	2.990e-05			
linft	57,135	0.01020	-0.00574	2.757	3.37	3.50
		1.565e-04	1.005e-04			
manft	67,326	0.00704	-0.00564	2.152	2.65	2.67
		1.171e-04	6.903e-05			

Table A.8.: *(continued)*

		Parameters		**Spreads**		
Ticker	N	θ_{HS}^{mod}	φ_{12}	$\hat{s}_{HS}^{0}$	$\hat{s}_2$	$\bar{s}$
		Std. err.	$(-\phi_{HS}^{mod})$	[Euro-Cent]	[Euro-Cent]	[Euro-Cent]
meoft	78,735	0.00925	-0.00522	2.394	3.05	3.10
		1.254e-04	7.515e-05			
muv2ft	218,564	0.01159	-0.01063	3.810	4.64	4.62
		1.340e-04	6.433e-05			
rweft	147,573	0.00548	-0.00455	1.666	2.11	2.11
		6.663e-05	3.457e-05			
sapft	178,952	0.01671	-0.01346	5.001	6.33	6.48
		1.842e-04	1.008e-04			
schft	97,004	0.00727	-0.00636	2.321	2.85	2.89
		1.134e-04	6.109e-05			
sieft	281,927	0.00631	-0.00620	2.137	2.62	2.63
		4.898e-05	2.978e-05			
tkaft	80,258	0.00403	-0.00433	1.433	1.74	1.76
		5.020e-05	3.514e-05			
tuift	67,646	0.00570	-0.00509	1.863	2.26	2.32
		7.865e-05	5.443e-05			
vowft	162,360	0.00566	-0.00483	1.751	2.21	2.17
		6.041e-05	3.451e-05			
Mean	**133,835**	**0.00752**	**-0.00633**	**2.356**	**2.91**	**2.95**

Bibliography

Ball, C., Chordia, T., 2001. True spreads and equilibrium prices. Journal of Finance 56 (5), 1801–1835.

Bessembinder, H., Maxwell, W., Venkataraman, K., 2006. Market transparency, liquidity externalities, and institutional trading costs in corporate bonds. Journal of Financial Economics 82 (2), 251–288.

Bjønnes, G. H., Rime, D., 2005. Dealer behavior and trading systems in foreign exchange markets. Journal of Financial Economics 75 (3), 571–605.

Brockman, P., Chung, D., Yan, X., 44. Block ownership, trading activity, and market activity. Journal of Financial and Quantitative Analysis 6 (1403-1426).

Brockman, P., Chung, D. Y., 2001. Managerial timing and coprorate liquidity: evidence from actual share repurchases. Journal of Financial Economics 61 (3), 417–448.

Brockwell, P. J., Davis, R. A., 2009. Time series: theory and methods. Springer.

Cao, C., Field, L., Hanka, H., 2004. Does Insider Trading Impair Market Liquidity? Evidence from IPO Lockup Expirations. Journal of Financial and Quantitative Analysis 39 (1), 25–46.

Chakravarty, S., Jaon, P., Upson, J., Robert, W., 2012. Clean sweep: Informed trading through intermarket sweep orders. Journal of Financial and Quantitative Analysis 47 (2), 415–435.

Da, Z., Gao, P., Jagannathan, R., 2011. Impatient trading, liquidity provision, and stock selection by mutual funds. Review of Financial Studies 24 (3), 675–720.

Glosten, L., Harris, L., 1988. Estimating the Components of the Bid-Ask Spread. Journal of Financial Economics 21 (1), 123–142.

Glosten, L. R., Milgrom, P. R., 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. Journal of financial economics 14 (1), 71–100.

Grammig, J., Theissen, E., Wünsche, O., 2006. Zur Schätzung von Geld-Brief-Spannen aus Transaktionsdaten. In: Bessler, W. (Ed.), Börsen, Banken und Kapitalmärkte: Festschrift für Hartmut Schmidt zum 65. Geburtstag. Duncker & Humboldt, pp. 71–83.

Green, J., Smart, S., 1999. Liquidity provision and noise trading: Evidence from the "investment dartboard" column. Journal of Finance 54 (5), 1885–1899.

Green, T. C., 2004. Economic news and the impact of trading on bond prices. Journal of Finance 59 (3), 1201–1233.

Han, S., Zhou, X., 2014. Informed bond trading, corporate yield spreads, and corporate default prediction. Management Science 60 (3), 675–694.

Hasbrouck, J., 2007. Empirical market microstructure: The institutions, economics and econometrics of securities trading. Vol. 4. Oxford University Press New York.

Hatch, B. C., Johnson, S. A., 2002. The impact of specialist firm acquisitions on market quality. Journal of Financial Economics 66 (1), 139–167.

Heflin, F., Shaw, K. W., 2000. Blockholder ownership and market liquidity. Journal of Financial and Quantitative Analysis 35 (4), 621–633.

Huang, R., Stoll, H., 1997. The components of the bid-ask spread: A general approach. Review of Financial Studies 10 (4), 995–1034.

Lee, C. M. C., Ready, M. A., 1991. Inferring trade direction from intraday data. Journal of Finance 46 (2), 733–746.

Lütkepohl, H., 2005. New introduction to multiple time series analysis. Cambridge Univ Press.

Madhavan, A., Richardson, M., Roomans, M., 1997. Why do security prices change? A transaction-level analysis of NYSE stocks. Review of Financial Studies 10 (4), 1035–1064.

Odders-White, E., Ready, M., 2006. Credit ratings and stock liquidity. Review of Financial Studies 19 (1), 119–157.

Petris, G., 2010. An r package for dynamic linear models. Journal of Statistical Software 36 (12), 1–16.

Roll, R., 1984. A simple implicit measure of the effective bid-ask spread in an efficient market. The Journal of Finance 39 (4), 1127–1139.

Weston, J. P., 2000. Competition on the Nasdaq and the impact of recent market reforms. Journal of Finance 55 (6), 2565–2598.

Part II.

Bayesian estimation of the probability of informed trading

We propose a methodology to estimate the probability of informed trading (PIN) that only requires data on the daily number of transactions (but not on the number of buyer-initiated and seller-initiated trades). Because maximum likelihood estimation of the model is problematic we propose a Bayesian estimation approach. We perform extensive simulations to evaluate the performance of our estimator. Our methodology increases the applicability of PIN estimation to situations in which the data necessary for trade classification is unavailable, or in which trade classification is inaccurate.

1. Introduction

The notion of informational asymmetries is central to our understanding of financial markets. Traders spend resources to acquire information, or to get access to information earlier than other traders. Because the presence of informed traders imposes costs on other traders, the amount, or intensity, of informed trading has immediate implications for the distribution of profits from trading and for market liquidity. Against this background it is not surprising that researchers in market microstructure have sought to measure the amount of informed trading.

In a seminal paper, Easley et al. (1996) have proposed a direct measure of the probability of informed trading (PIN). Their approach builds on the theoretical work of Easley and O'Hara (1987) and Easley and O'Hara (1992). In their model it is assumed that there are three types of trading days, good news days (with probability $\alpha(1-\delta)$), bad news days (probability $\alpha\delta$) and no news days (probability $(1-\alpha)$). The market is populated by liquidity traders, informed traders and uninformed market makers. Liquidity traders buy and sell with intensity ϵ. Informed traders buy with intensity μ on good news days, sell with intensity μ on bad news days, and abstain from trading on no-news days. Maximum likelihood estimation yields estimates of the structural parameters α, δ, ϵ and μ. From these estimates, an estimator of PIN is obtained as $\text{PIN} = \alpha\mu/(\alpha\mu + 2\epsilon)$. The data required for estimation are the daily numbers of buyer-initiated trades and seller-initiated trades. This data is typically generated by applying a trade classification algorithm (such as the Lee and Ready (1991) algorithm) to time-stamped intraday data on bid, ask and transaction prices. This kind of data is usually available for developed equity markets and recent sample periods. It is, however, often unavailable for less developed markets, for many derivatives markets, and for sample periods in the more distant past. Therefore, a method to estimate PIN that does not require the availability of intraday data would be of obvious interest.

Even in cases in which intraday data is available, there is an important impediment to accurate PIN estimation. It is well known that the Lee and Ready (1991) trade classification algorithm may yield inaccurate trade classification (e.g. Ellis et al. (2000), Lee and Radhakrishna (2000) and Odders-White (2000)). Boehmer et al. (2007) have shown that inaccurate trade classification results in biased PIN estimates. If the bias is severe, a methodology to estimate PIN that does not rely on trade classification may be superior to a method that uses misclassified data.

In this paper we build on previous research by Kokot (2004) and Jackson (2007) and propose a methodology to estimate PIN that only requires data on the daily number of transactions. This data is obviously much more easily available than the intraday data required to classify trades into buyer-initiated and seller-initiated trades. Further, it is not necessary to apply a trade classification algorithm. Therefore, the estimation results are unaffected by inaccurate trade classification. The intuition of our procedure is simple. From the formula for PIN shown

above it is obvious that the parameter δ is not needed. We consequently classify the trading days into two groups, information days and non-information days. On non-information days only liquidity traders trade (with intensity 2ϵ) while on information days both liquidity traders and informed traders trade (with total intensity $2\epsilon + \mu$). The resulting model is a mixture of two Poisson distributions with intensity parameters 2ϵ and $(2\epsilon + \mu)$ and mixing probability α. We refer to this model as the compressed model.

Unfortunately, even though the compressed model itself is very simple, estimation is not. It is well known that mixtures of Poisson distributions are difficult to estimate because of the presence of local solutions.[1] These difficulties arise particularly in small samples. We therefore propose to estimate the model using a Markov chain Monte Carlo (MCMC) simulation approach.

We evaluate our estimation procedure in extensive simulations. We know the data generating process used in our simulations and therefore know the true parameters α, ϵ and μ, and we know the true PIN. We then apply three versions of our MCMC estimation procedure (to be described in detail in section 2.4) to the simulated data and compare the resulting PIN estimate to the true PIN. In order to assess the accuracy of our estimators we also compare them to the traditional PIN estimator, to the traditional PIN estimator applied to misclassified data, and to a maximum likelihood estimator of the compressed model.

The simulation results indicate that our procedure yields accurate PIN estimates. We pick as an example the simulations in which we choose the true parameters to be equal to the parameters that Easley et al. (1996) estimated for the first, fifth and eighth size decile of NYSE firms. The true PIN is 0.1581 for the first decile, 0.2143 for the fifth and 0.2253 for the eighth. Our best estimation procedure (Bayesian maximum likelihood) yields an average PIN estimate (over 250 simulation runs) of 0.1577 (standard error 0.0162) for the first decile, 0.2184 (0.0417) for the fifth decile and 0.2369 (0.0698) for the eighth decile.

Our paper adds to the literature in two important ways. First, and most importantly, the substantially lower data requirements of our procedure as compared to traditional PIN estimation broaden the applicability of PIN estimation significantly. In particular, our procedure can be used to estimate PIN for markets and sample periods for which intraday data is unavailable and trade classification thus impossible. Further, because our procedure does not require trade classification it is unaffected by trade misclassification. It can thus be applied in cases in which the misclassification bias in traditional PIN estimates documented by Boehmer et al. (2007) is a concern.

The remainder of the paper is organized as follows. In section 2.1 we briefly describe the Easley et al. (1996) model and our modified model. Sections 2.2 to 2.4 describes our estimation

[1] See e.g. Finch et al. (1989), Frühwirth-Schnatter (2006) and the documentation of the `mix2Poisson` command in `R` at `http://hosho.ees.hokudai.ac.jp/~kubo/Rdoc/library/VGAM/html/mix2poisson.html`. We note that Jackson (2007) has proposed a model which is essentially identical to ours. He uses standard ML to estimate his model and finds that the PIN estimates he obtains differ significantly from those of a traditional PIN model. Jackson (2007) argues that this is because his procedure is unaffected by trade misclassification (but does not substantiate this claim). We argue instead that the different results are due to the fact that ML estimation of finite mixture models yields unreliable results.

strategy in detail. Chapter 3 describes the results, chapter 4 concludes.

2. Econometric methodology

2.1. Model

We start with a short description of the Easley et al. (1996) model. The model assumes a market populated by competitive market makers, informed traders, and uninformed (liquidity or noise) traders. Uninformed traders are equally likely to buy or sell shares. The arrival of uninformed buy and sell orders is modeled by two independent Poisson processes with equal arrival rates ϵ. Before the trading day begins an information event may occur with probability α. The information is either bad (with probability δ) or good (with probability $(1-\delta)$). Informed traders buy on good news days, sell on bad news days, and do not trade on no-news days. The arrival of their orders is modeled as a Poisson process with intensity parameter μ (assumed to be identical for informed buy and sell orders). The four structural parameters of the model – the probability that an information event occurs on a given day, α, the probability that the information is bad news, δ, and the order arrival rates of uninformed and informed traders, ϵ and μ – can be estimated by maximum likelihood (ML). The likelihood of observing B buyer-initiated trades and S seller-initiated trades on a given day, conditional on the true parameter vector $\vartheta_{EKOP} = (\alpha,\ \delta,\ \epsilon,\ \mu)$, is

$$
\begin{aligned}
L(B,S \mid \vartheta_{EKOP}) = {} & \alpha\delta(\epsilon T)^B \frac{\exp(-\epsilon T)}{B!}((\epsilon+\mu)T)^S \frac{\exp(-(\epsilon+\mu)T)}{S!} \\
& + (1-\alpha)(\epsilon T)^B \frac{\exp(-\epsilon T)}{B!}(\epsilon T)^S \frac{\exp(-\epsilon T)}{S!} \\
& + \alpha(1-\delta)((\epsilon+\mu)T)^B \frac{\exp(-(\epsilon+\mu)T)}{B!}(\epsilon T)^S \frac{\exp(-\epsilon T)}{S!}\ ,
\end{aligned}
\tag{2.1}
$$

where T denotes the length of the trading day in minutes. Easley et al. (1996) derive from their model the unconditional probability of informed trading (PIN),

$$
\mathrm{PIN} = \frac{\alpha\mu}{\alpha\mu + 2\epsilon}\ . \tag{2.2}
$$

In the following we refer to this model as the EKOP model.

Estimation of PIN requires data on the number of buyer- and seller-initiated trades. This data is typically obtained by applying a trade classification algorithm such as the Lee and Ready (1991) algorithm to intraday data. As noted in the introduction, there are two potential problems with this procedure. First, the intraday data needed for trade classification may simply

be unavailable. Second, trade classification algorithms are known to be inaccurate, and inaccurate trade classification, in turn, is known to result in biased PIN estimates (see Boehmer et al. (2007)). It is thus desirable to develop a PIN estimator that does not require trade classification. Conceptually, such an estimator is easily developed (see also Kokot (2004) and later Jackson (2007)). Note from eq. (4.2), that the PIN does not depend on the direction of the information. Therefore, the distinction between good news days and bad news days is not necessary to estimate PIN. When we collapse the good news days and the bad news days into a new category 'information days', we obtain a compressed likelihood function:

$$L(Q \mid \vartheta_{COMP}) = \alpha((2\epsilon + \mu)\, T)^Q \frac{\exp\left(-(2\epsilon + \mu)\, T\right)}{Q!} + (1-\alpha)(2\epsilon T)^Q \frac{\exp\left(-2\epsilon T\right)}{Q!}\,, \qquad (2.3)$$

where $Q = B + S$ is the number of transactions on a given trading day. From hereon this model is referred to as the 'compressed' model or the 'compressed EKOP' model.

Depending on the type of the day (news or no-news) an observation is generated from one of the two component distributions (either a Poisson distribution with intensity parameter $(2\epsilon + \mu)$ or a Poisson distribution with intensity parameter 2ϵ). Our modified model is thus a finite mixture model. It is well known that models of this type are difficult to estimate by maximum likelihood methods because of the presence of local solutions.[2] We therefore propose the Bayesian estimation approach described in the next section.

2.2. Mixture likelihood and complete-data likelihood representation for the compressed EKOP model

The starting point for our Bayesian PIN estimation procedure is the observation that the probability to observe Q trades during a trading day of T minutes can be described as a mixture of two Poisson distributions, viz

$$p(Q|\vartheta) = \gamma \mathcal{P}(Q|\lambda_1) + (1-\gamma)\mathcal{P}(Q|\lambda_2)\,, \qquad (2.4)$$

where $\vartheta = (\gamma, \lambda_1, \lambda_2)'$, $\lambda_1 = (2\epsilon + \mu)T$ and $\lambda_2 = 2\epsilon T$. $\mathcal{P}(Q|\lambda) = e^{-\lambda}\frac{\lambda^Q}{Q!}$ denotes the Poisson probability density function. Assuming that the information events are independent across days, the joint probability of observing a sequence of the number of daily trades over N days, $\mathbf{Q} = (Q_1, \ldots, Q_N)'$, is given by $p(\mathbf{Q}|\vartheta) = \prod_{i=1}^N p(Q_i|\vartheta)$, which implies the mixture log-likelihood function

$$\mathcal{L}_{\text{mix}}(\vartheta) = \ln p(\mathbf{Q}|\vartheta) = \sum_{i=1}^N \ln p(Q_i|\vartheta)\,. \qquad (2.5)$$

[2] See e.g. Finch et al. (1989), Frühwirth-Schnatter (2006) and the documentation of the `mix2Poisson` command of the `VGAM` R-package in Yee (2013).

The parameters in (2.4) are defined up to a 'label switch', which means that

$$p(Q|\vartheta^*) = \gamma^*\mathcal{P}(Q|\lambda_1^*) + (1-\gamma^*)\mathcal{P}(Q|\lambda_2^*)\,, \tag{2.6}$$

where $\vartheta^* = (\gamma^*, \lambda_1^*, \lambda_2^*)'$, $\gamma^* = 1-\gamma$, $\lambda_1^* = \lambda_2$, and $\lambda_2^* = \lambda_1$ represents an observationally equivalent parametrization. As a consequence, the mixture-likelihood in (2.5) is symmetrically bimodal, a fact that must be accounted for when attempting to maximize it.

While the maximum likelihood method has been the standard approach towards estimating mixture models (for a review see Redner and Walker (1984)), previous literature has identified severe problems that will also affect PIN estimation through the maximization of the mixture log-likelihood (2.5). Frühwirth-Schnatter (2006), among others, points out convergence problems and starting value dependence of the maximization algorithm, poor performance in small samples and weakly separated components of the mixture distribution (here: information and no-information days). Moreover, there are severe caveats to apply asymptotic theory to compute reliable standard errors.[3]

Our Bayesian approach towards PIN estimation avoids these drawbacks. The starting point is the so-called complete-data likelihood, which is based on the joint distribution of the number of trades per day, Q_i and the binary variables $Z_i, i = 1, \ldots, N$, henceforth referred to as allocations. The vector of allocations $\mathbf{Z} = (Z_1, \ldots, Z_N)'$ consists of i.i.d. Bernoulli(γ) random variables that are equal 1 when Q_i is drawn from the Poisson distribution associated with the parameter λ_1 and 0 when Q_i is drawn from a Poisson distribution associated with the parameter λ_2. The joint distribution of $\mathbf{Q}$ and $\mathbf{Z}$ can then be written as:

$$\begin{aligned} p(\mathbf{Q},\mathbf{Z}|\vartheta) =& p(\mathbf{Q}|\mathbf{Z},\vartheta)p(\mathbf{Z}|\vartheta) = \prod_{i=1}^{N} p(Q_i|Z_i,\vartheta)p(Z_i|\vartheta) \\ =& \prod_{i=1}^{N} (\mathcal{P}(Q_i|\lambda_1)\gamma)^{Z_i} \times (\mathcal{P}(Q_i|\lambda_2)(1-\gamma))^{1-Z_i} \\ =& \prod_{i:Z_i=1} \mathcal{P}(Q_i|\lambda_1) \times \prod_{i:Z_i=0} \mathcal{P}(Q_i|\lambda_2) \times \left(\gamma^{N_1(\mathbf{Z})}(1-\gamma)^{N_2(\mathbf{Z})}\right)\,, \end{aligned} \tag{2.7}$$

where $N_1(\mathbf{Z}) = \sum_{i=1}^{N} Z_i$ and $N_2(\mathbf{Z}) = N - N_1(\mathbf{Z})$. The maximization of the complete-data log-likelihood

$$\mathcal{L}_{\text{comp}}(\vartheta) = \ln p(\mathbf{Q},\mathbf{Z}|\vartheta)\,, \tag{2.8}$$

requires knowing the allocations $\mathbf{Z}$, which in the present application will not be available. How-

[3] Chung et al. (2004) report unreliable ML estimates for small sample sizes, Redner and Walker (1984) analyze the *condition number* of Hessian matrices and find that a sample size of 10^6 is needed to ensure reasonable standard errors for mixtures with weakly separated components. Chung et al. (2004) perform simulations using small samples and report severe irregularities of likelihood surfaces, and the failure to apply asymptotic reasoning to compute parameter standard errors. The problem of unreliable standard errors is also addressed by Basford et al. (1997).

ever, we can use

$$p(Z_i = 1|Q_i, \vartheta) = \frac{p(Q_i|Z_i = 1, \vartheta)p(Z_i = 1|\vartheta)}{p(Q_i|Z_i = 1, \vartheta)p(Z_i = 1|\vartheta) + p(Q_i|Z_i = 0, \vartheta)p(Z_i = 0|\vartheta)}$$

$$= \frac{\mathcal{P}(Q_i|\lambda_1)\gamma}{\mathcal{P}(Q_i|\lambda_1)\gamma + \mathcal{P}(Q_i|\lambda_2)(1-\gamma)} \tag{2.9}$$

for a classification of each trading day. As we will describe in detail in section 2.4, we use a Markov chain Monte Carlo (MCMC) algorithm to alternate between N independent Bernoulli draws to generate the allocations vector $\mathbf{Z}$ using the probability in (2.9) (based on values for λ_1, λ_2, and γ), and draws from the posterior distributions of λ_1, λ_2, and γ based on $\mathbf{Z}$. In the next two sections we first describe how to obtain the posterior distributions conditional on $\mathbf{Z}$ and $\mathbf{Q}$ and then explain the MCMC procedure and how to obtain Bayesian point estimates of the PIN.

2.3. Posterior, prior, and hyperprior distributions

The key object of interest for our Bayesian approach towards PIN estimation is the posterior joint distribution of parameters and allocations, $p(\vartheta, \mathbf{Z}|\mathbf{Q}) \propto p(\mathbf{Q}, \mathbf{Z}|\vartheta)p(\vartheta)$, where $p(\vartheta)$ is the prior joint density of the parameters. We can write:

$$p(\vartheta, \mathbf{Z}|\mathbf{Q}) \propto p(\mathbf{Q}, \mathbf{Z}|\vartheta)p(\vartheta)$$

$$= \prod_{i:Z_i=1} \mathcal{P}(Q_i|\lambda_1) \times \prod_{i:Z_i=0} \mathcal{P}(Q_i|\lambda_2) \times \left(\gamma^{N_1(\mathbf{Z})}(1-\gamma)^{N_2(\mathbf{Z})}\right) \times p(\vartheta). \tag{2.10}$$

Assuming prior independence and conditioning on $\mathbf{Z}$ yields the following expressions for the marginal posterior distributions of the model parameters:[4]

$$p(\lambda_1|\mathbf{Q}, \mathbf{Z}) \propto \left(\prod_{i:Z_i=1} \mathcal{P}(Q_i|\lambda_1)\right) p(\lambda_1)\,, \tag{2.11}$$

$$p(\lambda_2|\mathbf{Q}, \mathbf{Z}) \propto \left(\prod_{i:Z_i=0} \mathcal{P}(Q_i|\lambda_2)\right) p(\lambda_2)\,, \tag{2.12}$$

$$p(\gamma|\mathbf{Q}, \mathbf{Z}) \propto \left(\gamma^{N_1(\mathbf{Z})}(1-\gamma)^{N_2(\mathbf{Z})}\right) p(\gamma)\,. \tag{2.13}$$

We choose conjugate prior distributions, which in the present case entails using a Gamma density, $\mathcal{G}(a_0, b_0)$, for $p(\lambda_1)$, and $p(\lambda_2)$, and a Beta density, $\mathcal{B}(e_0, e_0)$, for $p(\gamma)$.[5] The posterior

[4] A detailed derivation can be found in section A.1 of chapter A in the appendix. We also used a dependent prior as suggested by Viallefont et al. (2002). Albeit the EKOP model structure may suggest otherwise, assuming prior independence yields superior results.

[5] Recall that the conjugate prior distribution for a Poisson likelihood is the Gamma distribution and the

densities in (2.11) -(2.13) are then given by:

$$\lambda_1|\mathbf{Q},\mathbf{Z} \sim \mathcal{G}(a_1(\mathbf{Z}), b_1(\mathbf{Z}))\ , \tag{2.14}$$

where $a_1(\mathbf{Z}) = a_0 + \sum_{i=1}^{N} Q_i \cdot Z_i$ and $b_1(\mathbf{Z}) = b_0 + N_1(\mathbf{Z})$, and

$$\lambda_2|\mathbf{Q},\mathbf{Z} \sim \mathcal{G}(a_2(\mathbf{Z}), b_2(\mathbf{Z}))\ , \tag{2.15}$$

where $a_2(\mathbf{Z}) = a_0 + \sum_{i=1}^{N} Q_i \cdot (1 - Z_i)$ and $b_2(\mathbf{Z}) = b_0 + N_2(\mathbf{Z})$, and

$$\gamma|\mathbf{Z} \sim \mathcal{B}(e_1(\mathbf{Z}), e_2(\mathbf{Z}))\ , \tag{2.16}$$

where $e_1(\mathbf{Z}) = e_0 + N_1(\mathbf{Z})$ and $e_2(\mathbf{Z}) = e_0 + N_2(\mathbf{Z})$.[6]

We follow the recommendations of previous literature and use a Gamma hyperprior for the parameter b_0 in the prior distributions of λ_1 and λ_2, $b_0 \sim \mathcal{G}(g_0, G_0)$, (see Viallefont et al. (2002)). The posterior density of the hyperparameter b_0 given by

$$\begin{aligned} p(b_0|\vartheta,\mathbf{Q},\mathbf{Z}) = p(b_0|\vartheta) &\propto p(\lambda_1|b_0)p(\lambda_2|b_0)p(b_0) \\ &\propto b_0^{g_0+2a_0-1} \exp\left(-\left(G_0 + (\lambda_1+\lambda_2)\right) b_0\right) \end{aligned} \tag{2.17}$$

is then again a Gamma density,[7]

$$p(b_0|\vartheta,\mathbf{Q},\mathbf{Z}) = p(b_0|\vartheta) \sim \mathcal{G}(g_0 + 2a_0, G_0 + (\lambda_1 + \lambda_2). \tag{2.18}$$

Regarding the other hyperparameters, we follow Viallefont et al. (2002) who suggest to use $a_0 = \overline{Q}^2/(s_Q^2 - \overline{Q})$, where $\overline{Q}$ and s_Q^2 denote the sample mean and variance of the data vector $\mathbf{Q}$, and Frühwirth-Schnatter (2006) who advocates using $g_0 = 0.5$ and $G_0 = g_0\overline{Q}/a_0$. Accordingly, $\mathbb{E}(b_0) = g_0/G_0 = a_0/\overline{Q}$, which implies that $\mathbb{E}(\lambda_1) = \mathbb{E}(\lambda_2) = \overline{Q}$ (the sample mean) and $\mathrm{Var}(\lambda_1) = \mathrm{Var}(\lambda_2) = s_Q^2 - \overline{Q}$ (the sample overdispersion) when evaluated at $b_0 = \mathbb{E}(b_0)$. The only subjective choice is the hyperparameter for the Beta prior, e_0. We want to ensure that γ is bounded away from zero, which is accomplished by setting $e_0 = 4$.[8]

conjugate prior distribution for a Bernoulli likelihood is the Beta distribution. We prefer conjugate priors because as pointed out by Robert (2007), the information conveyed in $\mathbf{Q}$ about ϑ should not lead to a modification of the whole structure of $p(\vartheta)$, only of its parameters. In the present case conjugate priors result in posteriors from standard distribution families, which facilitates sampling from these distributions.

[6] A detailed derivation of the posterior Gamma distribution is shown in section A.2 of chapter A in the appendix.

[7] This use of hierarchical priors should compensate for modeling errors at the lower levels and reinforces the non-informative perspective, while providing a well-defined posterior distribution of the parameters (Robert, 2007). We do not follow the approach taken in previous Bayesian market microstructure analysis to use improper priors. We thereby pay heed to the warning of Frühwirth-Schnatter (2006), who points out that improper priors may lead to improper mixture posteriors.

[8] We also used smaller hyperparameter values and the results remain robust.

2.4. Gibbs sampling and Bayesian point estimates

In order to draw from the posterior distributions (2.14)-(2.16) and (2.18), we use a Gibbs sampling algorithm that works as follows. We start with an initial allocations vector $\mathbf{Z}^{(0)}$ that results from a two-means cluster analysis of the data $\mathbf{Q}$, the vector of the number of daily trades, and an initial value for b_0, for which we use the midrange of $\mathbf{Q}$, the average of the smallest and largest number of daily trades in the sample. The next steps consist of $M_0 + M$ alternate draws from the posterior distributions (2.14)-(2.16), based on the vector of allocations from the previous step, and an update of the allocations vector along with a draw from the posterior density (2.18), for which the current draw from the posterior distributions for γ, λ_1 and λ_2 are taken as given. This procedure yields the Gibbs-sampled sequences $\{\gamma^{(m)}\}_{m=1}^{M_0+M}$, $\{\lambda_1^{(m)}\}_{m=1}^{M_0+M}$, $\{\lambda_2^{(m)}\}_{m=1}^{M_0+M}$, $\{\mathbf{Z}^{(m)}\}_{m=1}^{M_0+M}$, as well as $\{b_0^{(m)}\}_{m=1}^{M_0+M}$. The classification of the observations (trading days), which is required to obtain the allocations $\mathbf{Z}^{(m)}$, is performed using Equation (2.9). M_0 are the number of steps of the Gibbs sampler in the so-called burn-in phase, which is not used for post-processing.

The Gibbs sampling algorithm must account for the aforementioned label switching problem, that is, the fact that (2.4) and (2.6) are observationally equivalent parametrizations. For that purpose, we use the random permutation Gibbs sampler proposed by Frühwirth-Schnatter (2001). Random permutation means allowing for the probability of a label switch after each step, such that one of the two observationally equivalent parameter combinations is randomly selected. If a label switch is indicated, then the original draws $\lambda_1^{(m)}$ $\lambda_2^{(m)}$, $\gamma^{(m)}$ are replaced by the permuted values $\lambda_1^{(m)*} = \lambda_2^{(m)}$, $\lambda_2^{(m)*} = \lambda_1^{(m)}$, and $\gamma^{(m)*} = 1 - \gamma^{(m)}$. Moreover, all ones in $\mathbf{Z}^{(m)}$ are replaced by zeros, and all zeros are replaced by ones, so the allocations are also permuted.[9] All steps of our Gibbs sampler are summarized by algorithm B.1.1 in section B.1 of the appendix. An illustration of the resulting empirical marginal distributions of the Gibbs-sampled data $\{\lambda_1^{(m)}\}_{m=M_0+1}^{M}$, $\{\lambda_2^{(m)}\}_{m=M_0+1}^{M}$, and $\{\gamma^{(m)}\}_{m=M_0+1}^{M}$ is depicted in Figure 2.1(a). The bimodal kernel densities reflect the purposeful label switching of the random permutation Gibbs sampler.

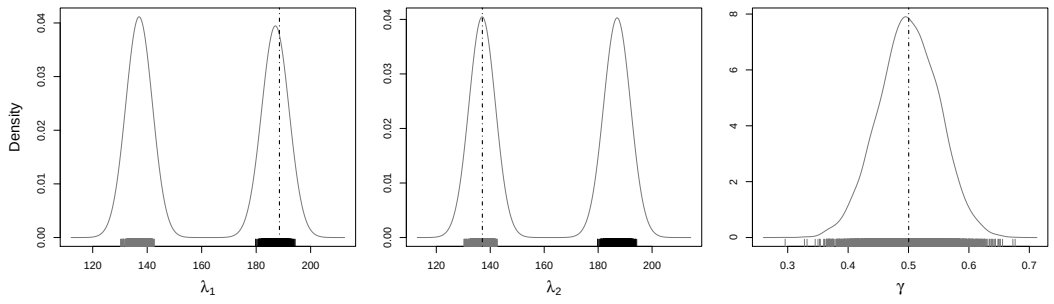
We consider three Bayesian point estimates for the PIN, which are based on the Gibbs-sampled data. The first is the maximum a-posteriori (MAP) estimate, which is the parameter combination $\hat{\boldsymbol{\vartheta}}_{\text{MAP}} = (\hat{\gamma}_{\text{MAP}}, \hat{\lambda}_{1,\text{MAP}}, \hat{\lambda}_{2,\text{MAP}})'$ that provides the maximum value of the posterior joint density,

$$\hat{\boldsymbol{\vartheta}}_{\text{MAP}} = \max_{\boldsymbol{\vartheta}^{(m)}} \{ p(\boldsymbol{\vartheta}^{(m)} | \mathbf{Q}, \mathbf{Z}^{(m-1)}) | m = M_0 + 1, \ldots, M_0 + M \} , \tag{2.19}$$

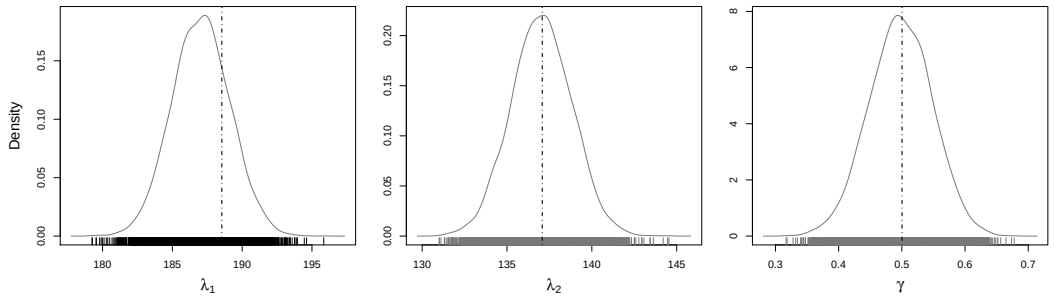
where the posterior joint density is given by the product of the three marginal densities in Equations (2.14)-(2.16).

To provide point estimates of the structural parameters of the EKOP model (α, ϵ, μ, and the

[9] As noted by Jasra et al. (2005), label switching is a desirable property, as it indicates that the Gibbs sampler is able to freely move around the parameter space.



(a)



(b)

Figure 2.1.: **Marginal posterior densities.** The figure depicts kernel density plots of the marginal posterior densities before (2.1(a)) and after relabeling (2.1(b)). Dashed lines indicate the true parameters. For the data generating process we used the first EKOP decile with $\alpha = 0.5003$, $\epsilon = 0.1757$, $\delta = 0.3491$, $\mu = 0.1320$, and $T = 390$, i.e. $\lambda_1 = 137.046$ trades per day and $\lambda_2 = 180.526$ trades per day and $\gamma = 0.5003$. The data set contained $N = 100$ observations and the Gibbs sampler used a burn-in of $M_0 = 1000$ iterations and a simulation size of $M = 10000$. We used the two-means clustering for classification of the resulting MCMC sample and applied a kernel density with a Gaussian kernel to the data for the plots. All plots were generated in `R-3.0.1` using the package `KernSmooth` (version `2.23.10`, see Wand and Ripley (2006) for the `KernSmooth`-package). The bandwidth is chosen by the 'oversized bandwidth selector', of Wand and Jones (1994, p.61). Original sample values in fig. 2.1(a) show a bimodal pattern for component parameters, λ_1, and λ_2, as label switching mixed sampled values from both components. Values from the first component are displayed by black lines and respective values from the second component by gray lines. Relabeled parameters in fig. 2.1(b) then show, that relabeling is successful in separating values from the two components. Sample values for γ scatter around the true value 0.5003 and therefore both component weights, γ and $(1 - \gamma)$, have very similar sample values and relabeling does not show significant changes in the posterior densities.

implied PIN) we use

$$\max\{\hat{\lambda}_{1,\text{MAP}}, \hat{\lambda}_{2,\text{MAP}}\} = (2\hat{\epsilon}_{\text{MAP}} + \hat{\mu}_{\text{MAP}})T \tag{2.20}$$

$$\min\{\hat{\lambda}_{1,\text{MAP}}, \hat{\lambda}_{2,\text{MAP}}\} = (2\hat{\epsilon}_{\text{MAP}})T \tag{2.21}$$

The point estimate of the probability of an information day $\hat{\alpha}_{\text{MAP}}$ is then either $\hat{\gamma}_{\text{MAP}}$, if $\hat{\lambda}_{1,\text{MAP}} > \hat{\lambda}_{2,\text{MAP}}$, or $1 - \hat{\gamma}_{\text{MAP}}$, if $\hat{\lambda}_{1,\text{MAP}} < \hat{\lambda}_{2,\text{MAP}}$. The MAP-based Bayesian PIN estimate is then $\widehat{\text{PIN}}_{\text{MAP}} = \frac{\hat{\alpha}_{\text{MAP}} \times \hat{\mu}_{\text{MAP}}}{\hat{\alpha}_{\text{MAP}} \times \hat{\mu}_{\text{MAP}} + 2\hat{\epsilon}_{\text{MAP}}}$.

The second vector of Bayesian point estimates $\hat{\vartheta}_{\text{BML}} = (\hat{\gamma}_{\text{BML}}, \hat{\lambda}_{1,\text{BML}}, \hat{\lambda}_{2,\text{BML}})'$ emerges from

$$\hat{\vartheta}_{\text{BML}} = \max_{\vartheta^{(m)}} \{p(\mathbf{Q}|\vartheta^{(m)})|m = M_0 + 1, \ldots, M_0 + M\}, \tag{2.22}$$

and is, for obvious reasons, referred to as Bayesian maximum likelihood (BML) estimate. The point estimates of the structural parameters of the EKOP model and the implied PIN are identified analogously to the MAP case, yielding $\hat{\alpha}_{\text{BML}}$, $\hat{\epsilon}_{\text{BML}}$, $\hat{\mu}_{\text{BML}}$, and the BML-implied Bayesian PIN estimate $\widehat{\text{PIN}}_{\text{BML}} = \frac{\hat{\alpha}_{\text{BML}} \times \hat{\mu}_{\text{BML}}}{\hat{\alpha}_{\text{BML}} \times \hat{\mu}_{\text{BML}} + 2\hat{\epsilon}_{\text{BML}}}$.

Because of this identification scheme, the bimodality of the posterior distributions used in (2.19) and the mixture likelihood function used in (2.22) do not pose a problem when computing the MAP- and BML-based PIN estimates. However, the third Bayesian point estimate that we consider requires a unique labeling. For that purpose we follow Celeux (1998) who proposes to apply a two-means cluster analysis on the Gibbs-sampled data, which assigns each of the $M - M_0$ parameter vectors to one of two clusters. Once the grouping is accomplished, the parameters in one of the two clusters are re-labeled, such that $\lambda_1^{(m)}$ $\lambda_2^{(m)}$, $\gamma^{(m)}$ are replaced by the permuted values $\lambda_1^{*(m)} = \lambda_2^{(m)}$, $\lambda_2^{*(m)} = \lambda_1^{(m)}$, and $\gamma^{*(m)} = 1 - \gamma^{(m)}$. Figure 2.1(b) shows an example of the resulting empirical distributions after clustering and relabeling the data in Figure 2.1(a). The kernel densities have become unimodal.[10] The third Bayesian point estimate, denoted $\hat{\vartheta}_{\text{EIA}} = (\hat{\gamma}_{\text{EIA}}, \hat{\lambda}_{1,\text{EIA}}, \hat{\lambda}_{2,\text{EIA}})'$, is then computed as the sample average of these relabeled draws produced by the Gibbs sampler. In line with previous literature, we refer to the estimate as the ergodic identified average (EIA) estimate. Using the relabeled Gibbs sample, we obtain the EIA estimate as:

$$\hat{\vartheta}_{\text{EIA}} = \frac{1}{M - M_0} \sum_{m=M_0+1}^{M_0+M} \vartheta^{(m)}. \tag{2.23}$$

Similar to the two other point estimates, the EIA-based PIN estimate results from equating the greater of the two estimates $\hat{\lambda}_{1,\text{EIA}}$ or $\hat{\lambda}_{2,\text{EIA}}$ with $(2\hat{\epsilon}_{\text{EIA}} + \hat{\mu}_{\text{EIA}})T$ and the smaller of the two point estimates with $2\hat{\epsilon}_{\text{EIA}}T$. The point estimate of the probability of an information day $\hat{\alpha}_{\text{EIA}}$ is then either $\hat{\gamma}_{\text{EIA}}$, if $\hat{\lambda}_{1,\text{EIA}} > \hat{\lambda}_{2,\text{EIA}}$, or $1 - \hat{\gamma}_{\text{EIA}}$, if $\hat{\lambda}_{1,\text{EIA}} < \hat{\lambda}_{2,\text{EIA}}$. The EIA-based Bayesian PIN estimate is then $\widehat{\text{PIN}}_{\text{EIA}} = \frac{\hat{\alpha}_{\text{EIA}} \times \hat{\mu}_{\text{EIA}}}{\hat{\alpha}_{\text{EIA}} \times \hat{\mu}_{\text{EIA}} + 2\hat{\epsilon}_{\text{EIA}}}$.

[10] We also applied the relabeling algorithms of Stephens (1997a,b). Results were quite similar and we decided to use the simplest method.

3. Results

3.1. Simulation Results

In order to evaluate the three Bayesian estimators and compare them to the traditional maximum likelihood estimators we perform extensive simulations. The simulations were implemented with the statistical software `R-3.0.1` using a self-coded package.[11] Our choice of the parameters α, ϵ and μ for the data generating process follows Boehmer et al. (2007). We choose 11 different parameter constellations. For each constellation we run 250 simulations with 100 data points each (corresponding to 100 days or slightly more than four months) and 250 simulations with 250 data points each (corresponding to approximately one year). We sort the parameter constellations into three groups, A, B and C. Group A (parameter ids 1-3 and 12-14) contains 'realistic' parameter constellations. They correspond to the parameters estimated in Easley et al. (1996) for stocks in the first, fifth and eighth size decile of NYSE stocks. In group B (ids 4-7 and 15-18) the intensity of informed trading, μ, is small relative to the intensity of uninformed trading, ϵ. We therefore expect that our algorithm (as well as the traditional ML estimator) will have difficulties to separate information and non-information days, particularly in cases in which the probability of an information event, α, is also small. In group C (ids 8-11 and 19-22) the intensity of informed trading, μ, is high relative to the intensity of uninformed trading, ϵ, and we therefore expect well separated components. All parameter combinations are summarized in table C.1.

In our simulation study we implement six different estimation methods. MCMC with a burn-in of $M_0 = 1000$ and a sample of $M = 10000$ using (1) the maximum likelihood estimator (BML), (2) the maximum a-posteriori estimator (MAP), and (3) the ergodic identified average estimator (EIA). Further, we implement three maximum likelihood estimators, namely, (4) the original EKOP model (EKOP) using accurately classified data, (5) the original EKOP model with misclassified data (EKOPM), and (6) maximum likelihood estimation of the compressed EKOP model (COMPML). The three maximum likelihood estimators (EKOP, EKOPM, COMPML) serve as benchmarks against which to evaluate our Bayesian estimation approach. If the Bayesian approaches perform better than the compressed EKOP model, a Bayesian approach should be used in all cases in which only data on the number of trades (but not on trade direction) is available. If the Bayesian approaches perform better than the EKOP model on misclassified data, then a Bayesian approach should be used whenever accurate trade classification

[11] `R` is a language and environment for statistical computing and graphics and it is open-source. This language shows the fastest growing user community in the area of statistical computing. For further information see R Core Team (2013). Our self-coded package relies heavily on `C++` extensions using the API `RcppArmadillo` (Eddelbuettel and Sanderson (2013)) that wraps the `C++` high-performance linear algebra library `Armadillo` (Sanderson et al. (2010)).

is infeasible. We do not expect the Bayesian estimators to perform better than the EKOP model applied to accurately classified data because the EKOP model uses more information (namely, information on trade direction).

We also tested the performance of the EM-based REBMIX algorithm. Although the EM algorithm is the most commonly applied method to find the ML estimator nowadays, it has several drawbacks as reported by Reddy and Rajaratnam (2010). The REBMIX algorithm origins in Nagode and Fajdiga (2011) and avoids these drawbacks. It is an iterative numerical procedure relying on the following steps: First, an empirical density is assigned to the dataset and the global mode position is identified. Next, rough component parameters are estimated from the empirical density and its global mode. Based on the rough component parameters the dataset is then clustered successively into the classes linked to the predictive component density and the residual. Enhanced component parameters and the component weights are then assessed for all classes and the remaining observations are distributed between the existing components by the Bayes decision rule. Finally, the parameters of the finite mixture are fine-tuned.[12] The results of the REBMIX algorithm applied to our simulated datasets are similar to the results of the ML estimation, though with greater bias and RMSE. In addition, the algorithm exhibited significantly more convergence problems than the ML estimator and offered, for some datasets in group B, solutions with one component only. For these reasons we omit the results. They are available from the authors upon request.

As noted above, each of the six estimation methods was applied to 250 samples of 100 data points (corresponding to 100 trading days or roughly four months) and another 250 samples of 250 trading days (roughly a year). We produced the data samples by drawing values from Poisson mixture distributions defined by the input parameters shown in table C.1 using the `Mersenne-Twister` random number generator in `R-3.0.1` with a seed of zero. We generate misclassified data by randomly changing the trade direction.[13] For each trade in the simulated data the trade direction was reversed with a probability of 15%. This corresponds to a misclassification rate of 15%, consistent with Odders-White (2000).

For the ML procedures we adopted the `optim` function in `R-3.0.1` with an `L-BFGS-B` algorithm using appropriate restrictions for the parameters and data-dependent starting values.[14] More precisely, we used a logit transformation for the parameter α (and also for the parameter δ which is estimated in the EKOP and EKOPM models) and starting values $\alpha = 0$, $\epsilon = 0.75/2 \cdot \overline{Q}$, ($\delta = 0$) and $\mu = 0.25/2 \cdot \overline{Q}$. In cases in which the `L-BFGS-B` algorithm did not converge or produced an error we applied as a second approach a derivate-free optimization, namely, a bounded Nelder-Mead algorithm via the `nmkb`-function from the `R`-package `dfoptim` (version `2011.8.1`, Varadhan et al. (2011)).[15]

[12] For a more detailed description of the REBMIX algorithm we refer to the vignette of the `R`-package `REBMIX` that we used, written by Nagode (2014).

[13] More precisely, we start at a random seed of zero and generate for each of the 11 parameter constellations in table C.1 (ids 1-11) 100 data points and then for each of them another 250 observations (ids 12-22). All data sets contain misclassified data though the original trade direction is recorded.

[14] See Zhu et al. (1997) for the `L-BFGS-B` algorithm.

[15] See Kelley (1999) for the bounded `Nelder-Mead` method.

All simulations were run on a node of the high-performance cluster of the RWTH Aachen University Computing Center.[16] A run with 250 simulations, applied to all parameter constellations and the two sample sizes of 100 and 250 days required approximately 90 minutes. Running a single MCMC simulation with a burn-in of 1000 iterations and a sample of 10000 iterations on a MacBook Pro with 2 Intel i7 cores (2.8 Ghz) takes about 0.85 seconds. Another 1.35 seconds are consumed for relabelling the sample and computing the three estimators, namely MAP, BML and EIA. These numbers demonstrate that the MCMC procedure is fast and is applicable with standard hardware.

As noted in chapter 2 we used a random permutation Gibbs sampler (see algorithm B.1.1 in the appendix) to improve mixing. Random permutation Gibbs samplers converged quickly as can be seen in the example traceplots shown in fig. 3.1 and the corresponding plots of the posterior density of the PIN in fig. 3.2. Figure 3.1 shows traceplots for parameter constellation id 1 (which corresponds to the parameters estimated for the first NYSE size decile by Easley et al. (1996)) before (panel a) and after (panel b) component classification. The traces fluctuate stationary around their mean values, thereby exploring the whole parameter space. For the component parameters in $\boldsymbol{\lambda}$ random permutation is obviously necessary.[17] From the identified traceplot (fig. 3.1(b)) we can infer that the two-means clustering is successful in separating the two components of the mixture.

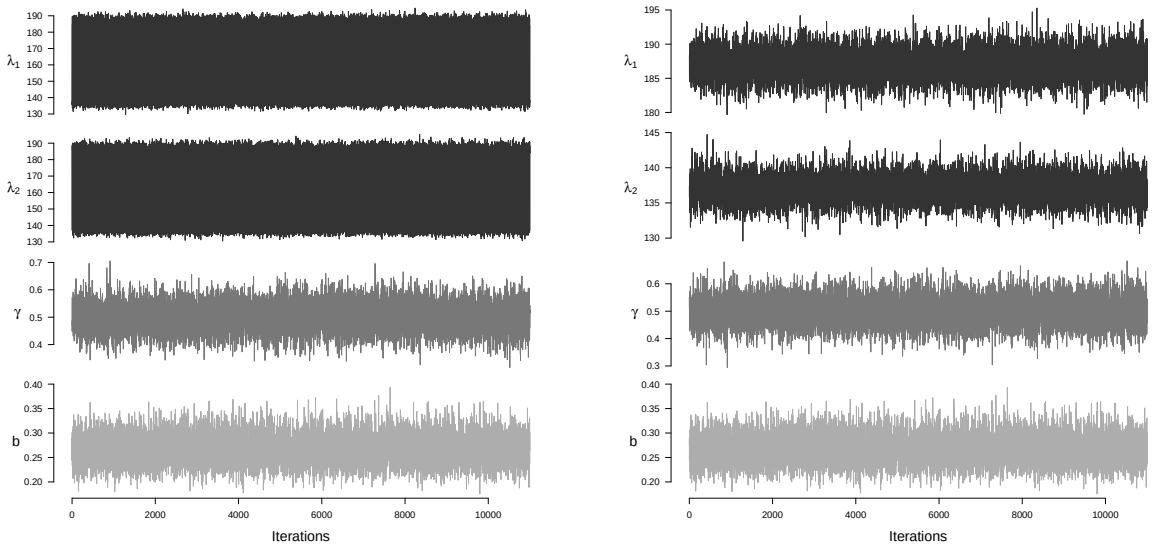
The posterior distributions of the PINs are shown in fig. 3.2 for the first four parameter constellations. Each plot shows the posterior estimated from the corresponding parameter draws of the Gibbs sampler (solid line) and the true PIN (dotted vertical line). The shaded areas mark the PIN values corresponding to the 95% highest posterior density (HPD).[18] In all cases the HPD interval includes the true PIN. This is evidence of the reliability and adequacy of the MCMC approach for finite mixtures. Note that the posteriors are dependent on the generated data and therefore simulated posteriors do inherently not disperse symmetrically around the true input parameter values.

The simulation results for all six estimators (the three MCMC point estimators and the three ML estimators) are presented in table C.2 (simulations with 100 observations) and table C.3 (250 observations). The tables show Monte Carlo averages of the PINs for each parameter combination together with corresponding Monte Carlo standard errors, calculated over all simulations. The PIN estimates for most parameter constellations are reasonably close to the true PIN. There is one notable exception, though. The EKOPM estimator (the EKOP estimator applied to misclassified data) tends to severly underestimate the true PIN. This confirms the findings of Boehmer et al. (2007). PIN estimates are the least accurate for parameter constellations 4 and

[16] This node is based on a Bullx S6010 board with 32 Intel Xeon (Nehalem) X7550 processors (2 Ghz). See for further information `http://www.itc.rwth-aachen.de/cms/IT-Center/Forschung-Projekte/~eubj/High-Perfomance-Computing/lidx/1/`

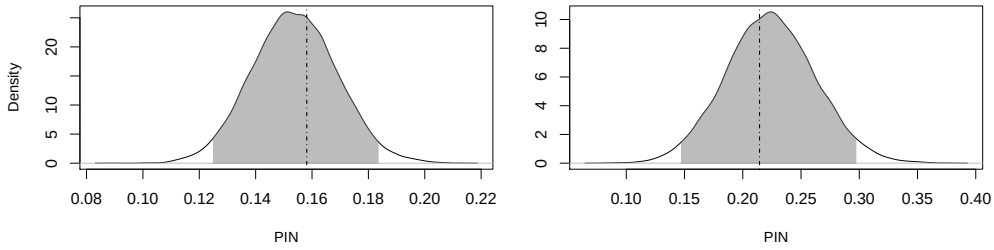
[17] The parameter constellation shown in fig. 3.1 is characterized by α close to 0.5. For parameter constellations with small α (ids 4-7 and ids 15-18), random permutation is also necessary for the parameter γ.

[18] We used the `LaplacesDemon` package (version `14.6.23`) in `R` to compute the highest posterior density intervals.

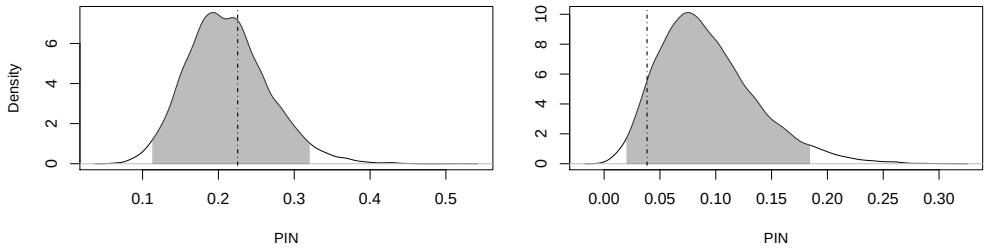


(a) (b)

Figure 3.1.: **Trace plots.** The figures depicts trace plots of the Gibbs-sampled sequences $\{\gamma^{(m)}\}_{m=1}^{M_0+M}$, $\{\lambda_1^{(m)}\}_{m=1}^{M_0+M}$, $\{\lambda_2^{(m)}\}_{m=1}^{M_0+M}$, and $\{b_0^{(m)}\}_{m=1}^{M_0+M}$, of a simulation with input parameters id 1 (first EKOP decile). All values are obtained by running a Gibbs sampler on $N = 100$ data points for $M = 10000$ iterations with a burn-in of $M_0 = 1000$. Results for fig. 3.1(a) come from the non-identified samples and fig. 3.1(b) plots the values identified through two-means clustering.



(a)



(b)

Figure 3.2.: **Posterior PINs.** The figures depict posterior densities for the PIN of randomly chosen simulations with input parameters id 1 (corresponding to the first NYSE size decile), id 2 (left and right in fig. 3.2(a)), id 3, and id 4 (left and right in fig. 3.2(b)) with $N = 100$ observations. The true PIN values are indicated by the dotted vertical lines. Shaded areas define the highest 95% posterior density intervals. All values were obtained by random permutation Gibbs sampling with a burn-in of $M_0 = 1000$ iterations and a sample of $M = 10000$ iterations. HPD intervals were computed using the `R`-package `LaplacesDemon` (version `14.6.23`).

5 in table C.2 and 15 and 16 in table C.3. This comes as no surprise, though, because these parameter constellations are characterized by a low intensity of informed trading and a low probability of an information event.

For a more rigorous statistical evaluation of the six estimators we used the two common loss functions *bias* and *root mean squared error* (RMSE):

$$\text{Bias}(\hat{\vartheta}_E) = \frac{1}{\text{Sim}} \sum_{i=1}^{\text{Sim}} (\hat{\vartheta}_{E,i} - \vartheta'), \quad \text{RMSE}(\hat{\vartheta}_E) = \left(\frac{1}{\text{Sim}} \sum_{i=1}^{\text{Sim}} (\hat{\vartheta}_{E,i} - \vartheta')^2 \right)^{0.5}, \tag{3.1}$$

where $\hat{\vartheta}_E$ is the parameter estimate obtained with estimator E, ϑ' is the true parameter and 'Sim' the number of Monte Carlo iterations. The results for the bias are shown in table C.4, those for the RMSE in table C.5.

Of all six estimators, the traditional EKOP estimator performs best. The average bias is (with the notable exceptions of parameter constellations 3, 4, 14 and 15) low, and the RMSE is, with few exceptions, the lowest of all six estimators. The good performance of the EKOP estimator comes as no surprise, though. It uses all available information (including trade direction), and it is based on the assumption that there are no trade classification errors. The effect of trade misclassification is clearly visible when comparing the EKOP estimator (which is based on correctly classified data) to the EKOPM estimator (which is based on misclassified data). The EKOPM estimator exhibits a significant, and often large, downward bias. This bias has already been documented by Boehmer et al. (2007). There are some parameter constellations in which the EKOPM estimator has low bias (ids 3, 4, 5, 14, 15, 16). Interestingly, these are exactly those constellations in which the EKOP estimator has a large positive bias. In these cases, the downward bias induced by trade misclassification counterbalances the positive bias of the EKOP estimator, resulting in low overall bias. This characteristic is likely to be of limited practical use, though, because the researcher does not typically know whether the EKOP estimator is biased in a particular application and will therefore not be able to identify those circumstances under which EKOPM performs well.

An important question in the presence of trade misclassification is whether it is better to use the inaccurate data, or to resort to one of the concentrated likelihood estimators which does not use the trade classification data. In cases in which the data necessary to classify trades is unavailable, the application of one of the four concentrated likelihood estimators is without alternative anyway. Therefore, we now turn to the evaluation of the COMPML, BML, MAP and EIA estimators. We will first discuss the results for parameter constellation groups A and C before turning to the more challenging conditions of group B.

The estimated PINs from the compressed model in tables C.2 and C.3 are reasonably close to the true values in the parameter constellation groups A and C. The standard errors of the estimators do not differ much, however, the BML estimator has the largest range of standard deviations, namely between 0.0100 for simulations with parameter combination id 12 and 0.0698 for id 3.

The biases shown in table C.4 range from -5.2964×10^{-2} for the EKOPM estimator on id 13

to 4.0392×10^{-2} for the EKOP estimator on id 3. Averaging the bias across different parameter constellations is inappropriate because some parameter constellations result in a negative bias for all estimators while other constellations predominantly result in positive biases. We therefore consider the average across parameter constellations of the absolute bias. In parameter constellation groups A and C the Bayesian estimators perform very well. They have lower average absolute bias than the three ML estimators. The BML estimator performs best with an average absolute bias of 0.001852.

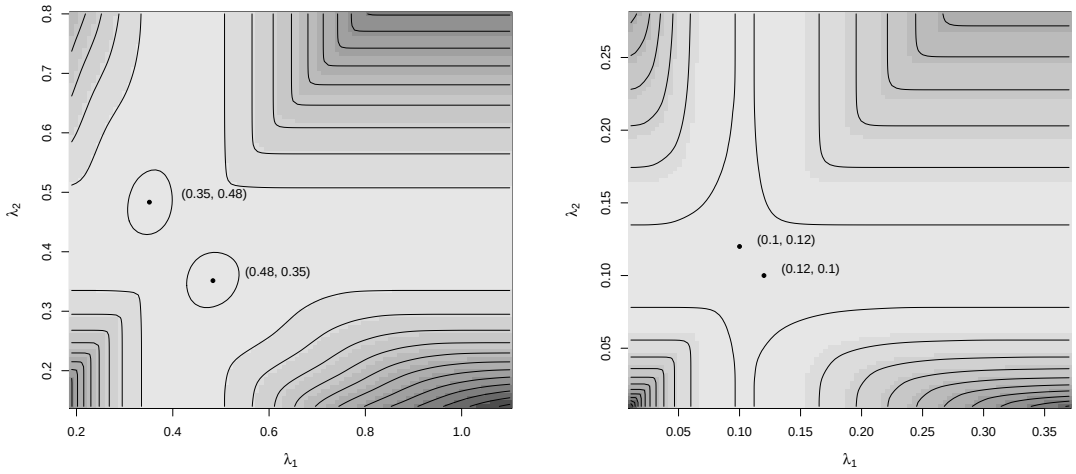
The EKOP estimator has the lowest RMSE (see table C.5. The average across the parameter constellations in groups A and C is 0.020. The three Bayesian estimators are ranked second to fourth and thus have a lower RMSE than the COMPML and the EKOPM estimators. Of the three Bayesian estimators the EIA estimator has the lowest average RMSE, followed by the MAP estimator.

We now turn to the results for the parameter constellations in group B. Here, the intensity of informed trading is low relative to the intensity of uninformed trading. Consequently, it is more difficult for the estimators to differentiate between information days and non-information days. This is particularly true if the probability of an information event is also low. The plot of the log-likelihood contours in fig. 3.3 illustrate this point. The first simulation (id 1, corresponding to the first NYSE decile) has well separated components and shows two well separated modes in the log-lieklihood surface (see fig. 3.3(a)). The second simulation (id 4, characterized by a low value of μ and a low probability α) appears almost unimodal (see fig. 3.3(b)). Correspondingly, the PIN estimates in tables C.2 and C.3 reveal that the accuracy of the point estimates is rather low when, in addition to a low intensity of informed trading, the probability of an information event, α, is also low (as in parameter ids 4 and 15). We further note that increasing the number of observations from 100 to 250 does not significantly increase the accuracy of the PIN point estimates (tables C.2 and C.3) but does reduce bias and RMSE by a substantial amount (tables C.4 and C.5).

In terms of bias the EKOP estimator dominates. The average absolute bias across the eight parameter constellations in group B is 0.0068×10^{-2}. Two of the Bayesian estimators, the MAP and EIA estimators, follow with average absolute biases of 0.0107×10^{-2} and 0.0117×10^{-2}, respectively. The BML estimator (which has the lowest average absolute bias in groups A and C) performs worst with an average absolute bias of 0.0183×10^{-2}.

The EKOP estimator also has the lowest average RMSE (0.209×10^{-2}). The EKOPM estimator (i.e. the EKOP estimator estimated on misclassified data) has a low average RMSE as well (0.0222×10^{-2}). Of the Bayesian estimators the EIA and MAP estimators perform reasonably well with average RMSEs of 0.0290×10^{-2} and 0.0331×10^{-2}, respectively. In particular, the two estimators have a considerably lower RMSE than the COMPML estimator (0.0496×10^{-2}). The BML estimator again performs worst; its average RMSE amounts to 0.0576×10^{-2}.

The performance of the BML point estimator deserves a discussion. It performs very well in parameter constellations of groups A and C but exhibits the largest RMSE and bias in parameter constellations of group B. If the information provided by the data is inconclusive, i.e. insufficient for the two modes to clearly become apparent in the likelihood, the region around



(a) (b)

Figure 3.3.: **Log-likelihood contours.** The figures depict log-likelihood contours for simulated data sets with parameters from id 1 (fig. 3.3(a)) and id 4 (fig. 3.3(b)). The contour plots cover a sufficient range around the true component parameters $\lambda_1 = 2\epsilon + \mu$ and $\lambda_2 = 2\epsilon$ and show the two possible optima due to label switching as labeled points.

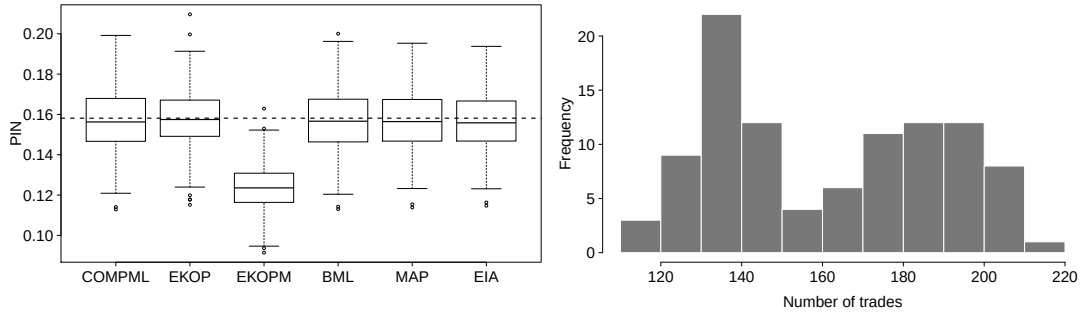
the two modes forms rather a plateau and the optimization algorithm encounters difficulties in finding the maximum. Consequently, estimated PINs suffer from higher bias.[19] It is further possible that the algorithm converges to the saddlepoint right in the middle of the line connecting the two modes in fig. 3.3(b). This will obviously result in a biased PIN estimate. It is further known that saddlepoints may result in a positive definite or an indefinite hessian. The two other Bayesian point estimators (MAP and EIA) are less affected by these difficulties because they include information about the prior distribution. Our simulation results indicate that this feature enables these estimators to reduce the bias significantly and to easily outperform the BML estimator (and also the EKOPM and COMPML estimators) in parametrical environments characterized by components that are poorly separated.

Figure 3.4 provides further insights into the simulation results. It shows boxplots (constructed from the results of all 250 simulations) for the same two parameter constellations (id 1, corresponding to the first NYSE decile and id 4, a parameter set with poorly separated components) already shown in fig. 3.3 above. The figure further shows histograms of a representative data sample. Parameter id 1 is characterized by well separated mixture components. The data sample shown in fig. 3.4(a) exhibits two clearly separated modes. All estimators (with the exception of EKOPM) are hardly biased. The PIN estimates of the 250 simulations are distributed symmetrically around the true PIN (see fig. 3.4(a)). The EKOPM estimator exhibits a pronounced downward bias. The bias is a consequence of trade misclassification. Because of this bias we exclude the EKOPM estimator from the second plot. This plot shows a parameter constellation with poorly separated components and a low news probability (id 4, $\alpha = 0.2$, fig. 3.4(b)). Correspondingly, the data sample in fig. 3.4(b) suggests a unimodal distribution. Consequently, the traditional maximum likelihood estimator (COMPML) and the BML estimator both produce a high number of outliers. The MAP and the EIA estimators, on the other hand, show a reasonably good performance because they also use the information contained in the prior distribution of the Bayesian framework. This advantage is of particular importance in 'difficult' settings characterized by small and/or poorly separated components.

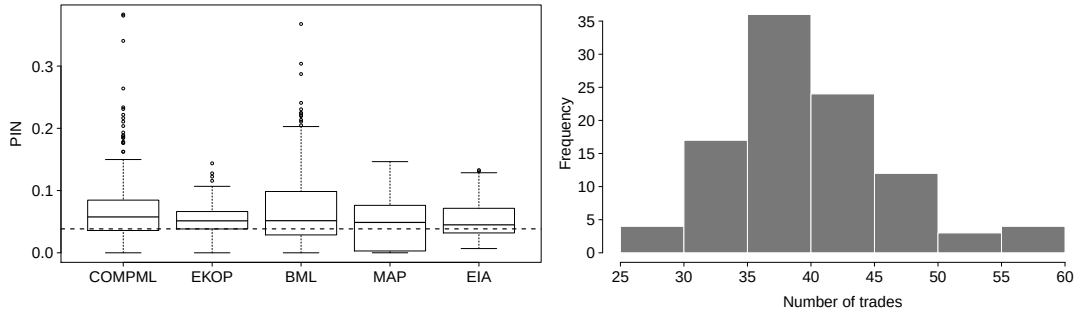
Figure 3.5 provides a different perspective on the simulation results. It shows the bias (Panel a) and the RMSE (Panel b) of all four estimators of the compressed model (COMPML, BML, MAP and EIA) for all 22 parameter constellations. The parameter constellations are sorted by the amount of bias (Panel a) and the RMSE (Panel b) of the maximum likelihood estimator COMPML. The shaded areas correspond to the parameter constellations belonging to group B. Figure 3.5 confirms the finding that in these difficult environments, the MAP and EIA estimators (which use information from the prior distribution) outperform the COMPML and BML estimators. In less difficult environments, the differences between the four estimators are much less pronounced. The RMSEs are hardly distinguishable. In terms of bias, the BML estimator performs best.

So far we have discussed the average performance of our estimators across all simulation runs. For practical applications it is also of interest how the estimators behave in individual

[19] In addition, if the likelihood forms a plateau around its modes, the hessian matrix could turn singular.

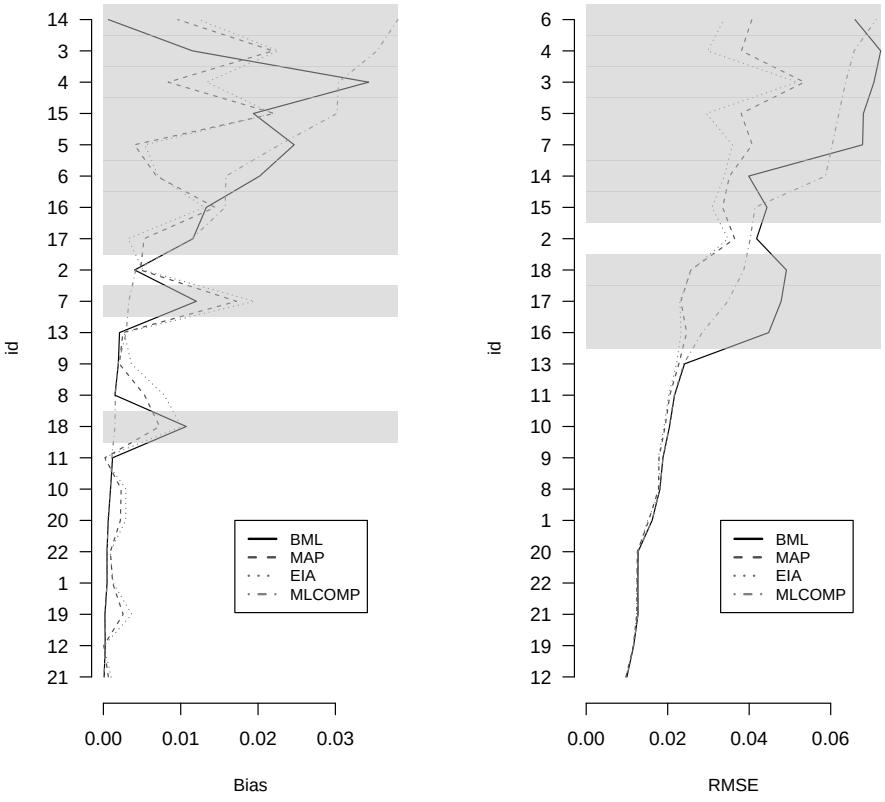


(a)



(b)

Figure 3.4.: **Boxplots.** The figures depict boxplots and data samples of PIN estimation routines from simulations with input parameters id 1 (first EKOP decile, fig. 3.4(a)) and id 4 (fig. 3.4(b)). The boxplots are constructed with all 250 estimators of the PIN from each simulation using the `boxplot`-function in `R-3.0.1`. The whiskers are determined by multiplying the interquartile range by 1.5.



(a) (b)

Figure 3.5.: **Bias and RMSE.** The figures depict plots of the bias (fig. 3.5(a)) and the RMSE (fig. 3.5(b)) for all estimators of the compressed EKOP model along all parameter combinations. Shaded areas identify parameter settings from group B with weakly separated components in the mixture.

data sets. Table C.6 provides some insights. The first three columns report the number of cases (out of a total of 250) in which the traditional maximum likelihood estimators (EKOP, EKOPM and COMPML) did not converge. On average slightly more than 7% of the estimations do not converge. However, in some parameter settings (among them ids 3 and 14 which correspond to the parameter values for the eighth NYSE decile) the percentage increases to almost 20%. In these cases the ML estimators do not provide results while the Bayesian estimators do.[20]

The remaining columns of table C.6 show the percentage of cases in which the Bayesian estimators produced results which were closer to the true PIN than the estimates produced by the ML estimators. These figures suggest that the Bayesian estimators are less accurate than the EKOP estimator and more accurate than the COMPML estimator. Consequently, the EKOP estimator should be chosen when accurate trade classification data is available while one of the Bayesian estimators should be chosen when no trade classification data is available. The comparison of the Bayesian estimators to the EKOPM estimator (i.e. the EKOP estimator applied to misclassified data) yields ambiguous results. The Bayesian estimators appear to be more accurate for parameter constellations in groups A and C while the EKOPM estimator is more accurate for parameter constellations in group B. The relatively good performance of the EKOPM estimator in group B was already documented earlier and is due to the fact that the inherent negative bias of the EKOPM estimator essentially corrects for a positive bias which the EKOP estimator exhibits in the parameter constellations of group B.

Of the three Bayesian estimators the BML estimator performs worst. Consider, for example, the comparison of the Bayesian estimators to the COMPML estimator in the last three columns of table C.6. For 17 out of 22 parameter constellations the MAP and the EIA estimator yield PIN estimates that are closer to the true value than the COMPML estimator in more than 50% of the cases. The BML estimator achieves a 50+% result only in 10 parameter constellations. The poor performance of the BML estimator documented here contrasts with the low average bias reported previously. These findings are easily reconciled, though. The BML estimator has a low bias but has high variance (as illustrated by the large RMSE documented in table C.5). Therefore, individual estimates are often inaccurate.

3.2. An algorithm to choose a best estimator

The discussion of the results in the previous section has revealed important differences between the three Bayesian estimators. Which estimator is best depends on (a) whether the bias or the RMSE is used as a criterion and (b) whether the parameter constellation is characterized by well separated components (constellations in groups A and C) or by poorly separated

[20] All results presented in this paper are based on all simulated datasets for which convergence was achieved. Thus, when a ML estimator converges in 240 out of 250 cases, the ML results for these 240 cases are compared to the Bayesian estimators obtained for the full set of 250 simulated datasets. It is conceivable that ML estimation fails in 'difficult' datasets. If this is the case our procedure puts the Bayesian estimators at a disadvantage. To control for this possibility we compared the ML estimates to the Bayesian estimates obtained only for those cases in which ML estimation converged. The results are similar to those presented in the paper.

components (group B). When judged by the RMSE the EIA estimator performs best in all environments. When judged by the bias, the BML estimator performs best in environments with well separated components while the MAP and EIA estimators are less biased in environments with poorly separated components.

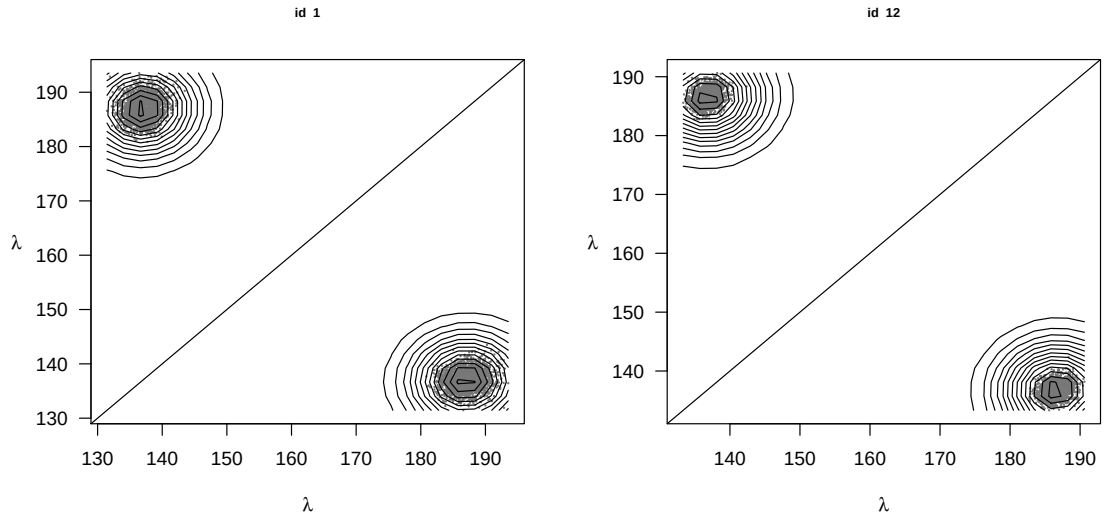
Thus, when the researcher is interested in an estimator with low RMSE she should choose the EIA estimator. However, when the focus is on reducing bias, the optimal choice of estimator depends on the environment. In empirical applications the true parameter values are obviously unknown. What is thus needed is a method that uses the data to differentiate between environments with well separated components and those with poorly separated components. In the cases of MCMC simulations such a differentiation can be achieved by considering sampling representations. Sampling representations were explored in Bayesian finite mixture estimation by Celeux et al. (2000); Frühwirth-Schnatter (2001) and Hurn et al. (2003) as a tool to visualize the posterior density in case of higher dimensionality. Sampled component parameters are stacked together and plotted against each other in a scatter plot, thereby revealing the modes of the posterior density. Figure 3.6 shows sampling representations of randomly chosen data sets for four of our parameter combinations in table C.1, namely parameter ids 1, 4, 12 and 15. When components in the underlying model are well separated the clusters in the sampling representation are also well separated (see fig. 3.6(a)). However, if the intensity of informed trading, μ, is small, the sampling representations of MCMC component parameter samples appear to be almost unimodal (id 4 and id 15 in fig. 3.6(b)).

If the number of estimations is small (e.g. because the cross-section is small), choosing an estimator based on manual inspection of the sampling representations is feasible. This procedure will be cumbersome, however, when the number of estimations is large. In these cases it will be preferable to conduct a formal test and decide upon the choice of estimator based on the result of that test. A test-based procedure also has the advantage of being unaffected by personal judgement and thus more reliable.

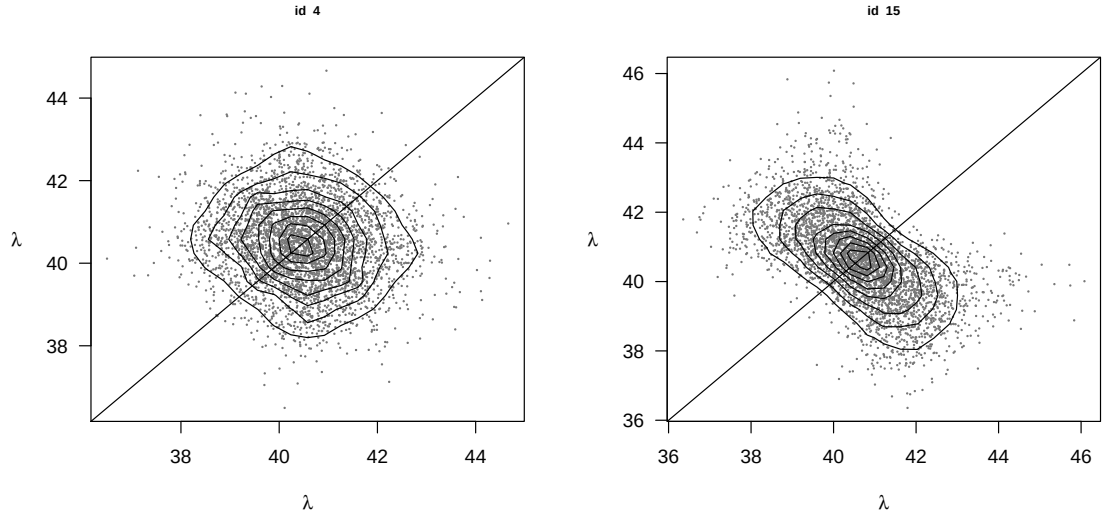
As illustrated above (see fig. 3.6), well separated components result in a sampling representation which is clearly bimodal while sampling representations for data sets with poorly separated components appear to be almost unimodal. We therefore propose an algorithm (presented as algorithm B.2.1 in section B.2 of the appendix) that uses the Hartigan and Hartigan (1985) test for unimodality. In the first step of the algorithm an MCMC simulation as described previously is run. In the second step the Hartigan and Hartigan (1985) test is applied to the sample of component parameters, $\boldsymbol{\lambda}$. If the test rejects the null hypothesis of unimodality at the 5% level of significance the BML point estimator is chosen.[21] If the null hypothesis is not rejected the MAP point estimator is chosen instead.

We tested this algorithm using all data samples of our simulation study. The performance of the algorithm is summarized in fig. 3.7. The figure shows the absolute average bias of all four estimators of the compressed EKOP model (COMPML, BML, EIA and MAP) together with the point estimator proposed by our algorithm (denoted ALGBM, the solid line in the figure). The

[21] Tests with other significance levels gave similar results.



(a)



(b)

Figure 3.6.: **Sampling representations.** The figure shows sampling representations for the parameter combinations 1 and 12 (fig. 3.6(a), 100 observations), as well as 4 and 15 (fig. 3.6(b), 250 observations) from table C.1 on randomly chosen data sets. Each plot displays the clusters of sampled component parameters resulting from the two ways of labelling these parameters.

algorithm consistently delivers a small bias. We therefore conclude that it can be successfully applied in cases in which the main objective is to obtain an estimator that delivers a small bias.

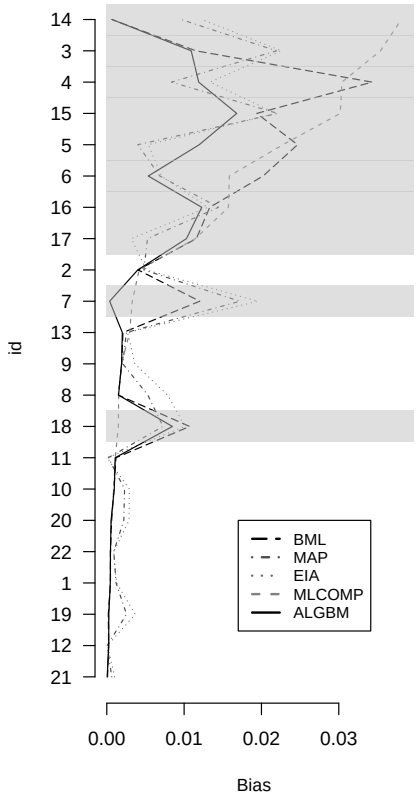


Figure 3.7.: **Bias-reducing algorithm.** The figure shows a plot of the bias for all estimators of the compressed EKOP model and our bias-reducing algorithm (ALGBM) along all parameter combinations. Shaded areas identify parameter settings from group B with weakly separated components in the mixture.

The results in fig. 3.7 are based on averages over the 250 simulations for each parameter constellation. Table C.7 considers the results from individual simulations instead. It reveals that the unimodality test always rejects the null hypothesis in the parameter constellations of groups A and C. However, it also tends to reject the null hypothesis in the parameter constellations of group B. In fact, there is only one parameter constellation (id 4) in which the rejection rate is below 50%. Thus, the ability of the Hartigan and Hartigan (1985) test to discriminate between environments with well separated and poorly separated components is limited. Consequently, our algorithm recommends the BML estimator most of the time. This is reflected by the results shown in the last three columns of table C.7. They show how frequently our algorithm deliv-

ers PIN estimates that are closer to the true value than the estimates obtained from the EKOP, EKOPM and COMPML estimators. These frequencies can be compared to the corresponding values for the BML, MAP and EIA estimators (shown in table C.6 above). In parameter constellations A and C the algorithm is equivalent to the BML algorithm. In parameter constellation B it improves slightly upon the BML estimator but does not achieve the performance of the MAP and EIA estimators. This is a reflection of the fact that (as outlined above) the algorithm sometimes selects the BML estimator when the MAP estimator would have been superior. In future research we will try to improve further our algorithm by using a test of the null hypothesis of bimodality.

4. Conclusion

In this paper we explore procedures to estimate the probability of informed trading (PIN) in settings in which data on trade classification is unavailable, or in which trade classification is inaccurate. The starting point of our investigation is the compressed maximum likelihood estimator first proposed by Kokot (2004). This estimator only requires data on the daily number of trades (but not on the daily number of buyer- and seller-initiated trades as does the original Easley et al. (1996) (EKOP) estimator).

The compressed maximum likelihood estimator estimates the parameters of a univariate mixture of two Poisson distributions. It is well known that maximum likelihood estimation of mixture distributions encounters difficulties. We therefore propose Bayesian estimation of the compressed model. Specifically, we apply a random permutation Gibbs sampler and evaluate in a simulation study the performance of three Bayesian point estimators, the Bayesian maximum likelihood (BML), the maximum a-posteriori (MAP), and the ergodic identified average (EIA). To put the performance of these estimators into perspective we compare them to the traditional EKOP estimator, the EKOP estimator applied to misclassified data, and an ML estimator of the compressed likelihood model (COMPML).

In our simulation exercise we define different parameter constellations. If the probability of an information event is high and the intensity of informed trading is high relative to the intensity of uninformed trading the two components (information days and non-information days) are well separated and accurate estimation of the probability of informed trading is 'easy'. If, on the other hand, the probability of an information event is low and/or the intensity of informed trading is low, the two components are poorly separated and accurate estimation is 'difficult'.

Our simulation results yield the following recommendations. If accurate trade classification data is available the traditional EKOP model should be used. It uses all information in the data and yields the best PIN estimates in terms of bias and RMSE. If trade classification is inaccurate it may be advisable to discard the information on trade direction and estimate the compressed

model instead. Finally, if data on trade classification is unavailable, there is obviously no alternative to the compressed model.

Of the four estimators of the compressed model the Bayesian estimators are generally superior to the COMPML estimator. The EIA estimator has the lowest RMSE. If the focus is on low bias rather than low RMSE, the BML estimator should be used in environments with well separated components and the MAP or EIA estimator in environments with poorly separated components. We recommend to use sampling representations or a unimodality test to differentiate between these cases.

Our results have important implications for empirical research in market microstructure. The compressed model extends the applicability of PIN estimation considerably. All that is required is data on the daily number of trades. Intraday data (which, in the context of the traditional EKOP model is required to classify trades) is not needed. Therefore, the compressed model can be applied to historical data, or to data from markets that do not make available intraday data. The results of our simulation study provide valuable advice to researchers on how to implement the compressed likelihood model.

A. Derivations

A.1. Derivation of posterior distributions

We assume the parameters to be independent a priori, i.e.

$$p(\vartheta) = p(\lambda_1)p(\lambda_2)p(\gamma)\,. \tag{A.1}$$

For the joint probability of ϑ and $\mathbf{Z}$ in the Markov chain holds

$$p(\vartheta, \mathbf{Z}|\mathbf{Q}) = \frac{p(\mathbf{Q}|\mathbf{Z}, \vartheta)p(\mathbf{Z}, \vartheta)}{p(\mathbf{Q})} = \frac{p(\mathbf{Q}|\mathbf{Z}, \vartheta)p(\mathbf{Z}|\vartheta)p(\vartheta)}{p(\mathbf{Q})}$$

$$= \frac{p(\mathbf{Q}, \mathbf{Z}|\vartheta)p(\vartheta)}{p(\mathbf{Q})}\,, \tag{A.2}$$

where $p(\mathbf{Q}, \mathbf{Z}|\vartheta)$ is the complete-data likelihood defined by

$$p(\mathbf{Q}, \mathbf{Z}|\vartheta) = \left(\prod_{i:Z_i=1} \mathcal{P}(Q_i|\lambda_1)\right)\left(\prod_{i:Z_i=0} \mathcal{P}(Q_i|\lambda_2)\right)$$

$$(\gamma^{N_1(\mathbf{Z})}(1-\gamma)^{N_2(\mathbf{Z})}) \tag{A.3}$$

$$= p(\mathbf{Q}|\mathbf{Z}, \vartheta)p(\mathbf{Z}|\vartheta)\,. \tag{A.4}$$

Furthermore, it holds via Bayes' rule

$$p(\vartheta|\mathbf{Q}, \mathbf{Z}) = \frac{p(\mathbf{Q}, \mathbf{Z}|\vartheta)p(\vartheta)}{p(\mathbf{Q}, \mathbf{Z})}\,. \tag{A.5}$$

Let Λ^j with $j \in \{1, 2\}$ be the parameter space for λ_j, and let $\mathcal{E}_2$ be the unit simplex. Then,

$$p(\lambda_1|\mathbf{Q}, \mathbf{Z}) = \int_{\Lambda^2 \times \mathcal{E}_2} p(\vartheta|\mathbf{Q}, \mathbf{Z})d(\lambda_2, \gamma)$$

$$\overset{\text{eq. (A.5)}}{=} \int_{\Lambda^2 \times \mathcal{E}_2} \frac{p(\mathbf{Q}, \mathbf{Z}|\vartheta)p(\vartheta)}{p(\mathbf{Q}, \mathbf{Z})}d(\lambda_2, \gamma)$$

$$\overset{\text{eq. (A.1)}}{=} \int_{\Lambda^2 \times \mathcal{E}_2} \frac{p(\mathbf{Q}, \mathbf{Z}|\vartheta)p(\lambda_1)p(\lambda_2)p(\gamma)}{p(\mathbf{Q}, \mathbf{Z})}d(\lambda_2, \gamma) \tag{A.6}$$

It remains to identify the joint unconditional distribution $p(\mathbf{Q}, \mathbf{Z})$:

$$\begin{aligned}
p(\mathbf{Q}, \mathbf{Z}) &= \int_{\Lambda^1 \times \Lambda^2 \times \mathcal{E}_2} p(\mathbf{Q}, \mathbf{Z}, \vartheta) d(\vartheta) \\
&= \int_{\Lambda^1 \times \Lambda^2 \times \mathcal{E}_2} p(\mathbf{Q}, \mathbf{Z} | \vartheta) p(\vartheta) d(\vartheta) \\
&\overset{eq.(A.1)}{=} \int_{\Lambda^1 \times \Lambda^2 \times \mathcal{E}_2} p(\mathbf{Q}, \mathbf{Z} | \vartheta) p(\lambda_1) p(\lambda_2) p(\gamma) d(\lambda_1, \lambda_2, \gamma) \\
&\overset{eq.(A.3)}{=} \int_{\Lambda^1 \times \Lambda^2 \times \mathcal{E}_2} \left(\prod_{i:Z_i=1} \mathcal{P}(Q_i|\lambda_1) \right) \left(\prod_{i:Z_i=0} \mathcal{P}(Q_i|\lambda_2) \right) \\
&\quad (\gamma^{N_1(\mathbf{Z})} (1-\gamma)^{N_2(\mathbf{Z})}) p(\lambda_1) p(\lambda_2) p(\gamma) d(\lambda_1, \lambda_2, \gamma)
\end{aligned} \tag{A.7}$$

Using the derived equation (A.7) in equation (A.6) and considering the independence of the single terms in equation (A.3) gives us:

$$\begin{aligned}
p(\lambda_1 | \mathbf{Q}, \mathbf{Z}) = \int_{\Lambda^2 \times \mathcal{E}_2} & \frac{\left(\prod_{i:Z_i=0} \mathcal{P}(Q_i|\lambda_2) \right) p(\lambda_2)}{\int_{\Lambda^2 \times \mathcal{E}_2} \left(\prod_{i:Z_i=0} \mathcal{P}(Q_i|\lambda_2) \right) p(\lambda_2)} \\
& \frac{(\gamma^{N_1(\mathbf{Z})} (1-\gamma)^{N_2(\mathbf{Z})}) p(\gamma)}{(\gamma^{N_1(\mathbf{Z})} (1-\gamma)^{N_2(\mathbf{Z})}) p(\gamma) d(\lambda_2, \gamma)} d(\lambda_2, \gamma) \\
& \frac{\left(\prod_{i:Z_i=1} \mathcal{P}(Q_i|\lambda_1) \right) p(\lambda_1)}{\int_{\Lambda^1} \left(\prod_{i:Z_i=1} \mathcal{P}(Q_i|\lambda_1) \right) p(\lambda_1) d\lambda_1}
\end{aligned} \tag{A.8}$$

and as the denominator defines a constant,

$$\begin{aligned}
p(\lambda_1 | \mathbf{Q}, \mathbf{Z}) = & \frac{\int_{\Lambda^2 \times \mathcal{E}_2} \left(\prod_{i:Z_i=0} \mathcal{P}(Q_i|\lambda_2) \right) p(\lambda_2)}{\int_{\Lambda^2 \times \mathcal{E}_2} \left(\prod_{i:Z_i=0} \mathcal{P}(Q_i|\lambda_2) \right) p(\lambda_2)} \\
& \frac{(\gamma^{N_1(\mathbf{Z})} (1-\gamma)^{N_2(\mathbf{Z})}) p(\gamma) d(\lambda_2, \gamma)}{(\gamma^{N_1(\mathbf{Z})} (1-\gamma)^{N_2(\mathbf{Z})}) p(\gamma) d(\lambda_2, \gamma)} \\
& \frac{\left(\prod_{i:Z_i=1} \mathcal{P}(Q_i|\lambda_1) \right) p(\lambda_1)}{\int_{\Lambda^1} \left(\prod_{i:Z_i=1} \mathcal{P}(Q_i|\lambda_1) \right) p(\lambda_1) d\lambda_1} \\
= & \frac{\left(\prod_{i:Z_i=1} \mathcal{P}(Q_i|\lambda_1) \right) p(\lambda_1)}{\int_{\Lambda^1} \left(\prod_{i:Z_i=1} \mathcal{P}(Q_i|\lambda_1) \right) p(\lambda_1) d\lambda_1}
\end{aligned} \tag{A.9}$$

The derivation of the marginal parameter distributions $p(\lambda_2)$ and $p(\gamma)$ works analogously.

A.2. Derivation of the posterior Gamma distribution

We start with the notation of the posterior distribution of λ_1 for given observations, $\mathbf{Q}$, and allocations, $\mathbf{Z}$:

$$p(\lambda_1|\mathbf{Q},\mathbf{Z}) = \frac{p(\mathbf{Q},\mathbf{Z}|\lambda_1)p(\lambda_1)}{p(\mathbf{Q},\mathbf{Z})} \propto p(\mathbf{Q},\mathbf{Z}|\lambda_1)p(\lambda_1) \tag{A.10}$$

$$= \underbrace{\prod_{i:Z_i=1} e^{-\lambda_1}\frac{\lambda_1^{Q_i}}{Q_i!}}_{p(\mathbf{Q},\mathbf{Z}|\lambda_1)} \underbrace{\frac{b_0^{a_0}}{\Gamma(a_0)}\lambda_1^{a_0-1}e^{-b_0\lambda_1}}_{p(\lambda_1),\ \lambda_1\sim\Gamma(a_0,b_0)}$$

$$\propto \prod_{i:Z_i=1} e^{-\lambda_1}\lambda_1^{Q_i}\lambda_1^{a_0-1}e^{-b_0\lambda_1}$$

$$\propto e^{-\lambda_1 N_1(\mathbf{Z})}\lambda_1^{\sum_{i=1}^N Q_i Z_i}\lambda_1^{a_0-1}e^{-b_0\lambda_1}$$

$$\propto e^{-\lambda_1 N_1(\mathbf{Z})-b_0\lambda_1}\lambda_1^{\sum_{i=1}^N Q_i Z_i+a_0-1}$$

$$\propto e^{-\lambda_1 (N_1(\mathbf{Z})+b_0)}\lambda_1^{\sum_{i=1}^N Q_i Z_i+a_0-1}$$

$$\propto e^{-\lambda_1 b_1}\lambda_1^{a_1-1}$$

$$\propto \frac{b_1^{a_1}}{\Gamma(a_1)}e^{-\lambda_1 b_1}\lambda^{a_1-1}$$

$$\propto \Gamma(a_1,b_1)\ ,$$

where $a_1 = a_0 + \sum_{i=1}^N Q_i Z_i$ and $b_1 = b_0 + N_1(\mathbf{Z})$. The Gamma posterior for λ_2 is derived analogously.

B. Algorithms

B.1. MCMC algorithm

The algorithm of the Gibbs sampler consists of the following steps:

Algorithm B.1.1. *Start with some classification* $\mathbf{Z}^{(0)}$ *and a hyperparameter* $b_0^{(0)}$ *and repeat the following steps for* $m = 1, \ldots, M_0, \ldots, M + M_0$.

1. *Parameter simulation conditional on the classification* $\mathbf{Z}^{(m-1)}$*:*
   a) *Sample* $\gamma, (1-\gamma)$ *from a* $\mathcal{B}(e_1(\mathbf{Z}^{(m-1)}), e_2(\mathbf{Z}^{(m-1)}))$*-distribution, given by eq.* (2.16).
   b) *Sample* λ_1 *from a* $\mathcal{G}(a_1(\mathbf{Z}^{(m-1)}), b_1(\mathbf{Z}^{(m-1)}))$*-distribution, given by eq.* (2.14).
   c) *Sample* λ_2 *from a* $\mathcal{G}(a_2(\mathbf{Z}^{(m-1)}), b_2(\mathbf{Z}^{(m-1)}))$*-distribution, given by eq.* (2.15).
2. *Classification of each observation* Q_i *conditional on knowing* $\boldsymbol{\vartheta}^{(m)}$*: sample* Z_i *independently for each* $i = 1, \ldots, N$ *from the conditional posterior distribution* $p(Z_i|Q_i, \boldsymbol{\vartheta}^{(m)})$*, which is given by*

$$p(Z_i = 1|Q_i, \boldsymbol{\vartheta}^{(m)}) \propto (\lambda_1^{(m)} T)^{Q_i} \exp(-\lambda_1^{(m)} T)\gamma^{(m)} , \tag{B.1}$$

$$p(Z_i = 0|Q_i, \boldsymbol{\vartheta}^{(m)}) \propto (\lambda_2^{(m)} T)^{Q_i} \exp(-\lambda_2^{(m)} T)(1-\gamma)^{(m)} . \tag{B.2}$$

3. *Sample* $b_0^{(m)}$ *from* $p(b_0|\boldsymbol{\lambda})$ *given by eq.* (2.18)*:*

$$b_0|\boldsymbol{\lambda} \sim \mathcal{G}\left(g_0 + 2a_0, G_0 + (\lambda_1^{(m)} + \lambda_2^{(m)})T\right) . \tag{B.3}$$

4. *Conclude each draw by selecting randomly one of the 2 possible permutations, denoted by* $\{\rho_s(1), \rho_s(2)\}$*,* $s \in \{1,2\}$ *(namely* $\{1,2\}$ *and* $\{2,1\}$*), of the current labeling. This permutation is applied to* $\boldsymbol{\gamma}^{(m)}$*, the component parameters* $\lambda_1^{(m)}$*,* $\lambda_2^{(m)}$*, and the classifications* $\mathbf{Z}^{(m)}$*:*
   a) *The group weight* $\gamma^{(m)}$ *is substituted by* $\gamma^{(m)*} = (1-\gamma^{(m)})$*, if the permutation* $\{\rho_2(1), \rho_2(2)\} = \{2,1\}$ *is drawn and remains equal for* $\{\rho_1(1), \rho_1(2)\} = \{1,2\}$.
   b) *The component parameters* $\lambda_1^{(m)}$ *and* $\lambda_2^{(m)}$ *are substituted by* $\lambda_{\rho_s(1)}^{(m)}$ *and* $\lambda_{\rho_s(2)}^{(m)}$ *respectively.*
   c) *The classifications* $Z_i^{(m)}$*,* $i = 1, \ldots, N$*, are substituted by* $\rho_s(Z_i^{(m)} + 1) - 1$*,* $i = 1, \ldots, N$.

M_0 is called the *burn-in* of the sampler and denotes the number of iterations that are descarded due to convergence criteria. M is the actual number of iterations after discarding the burn-in.

B.2. Bias-reducing algorithm

Algorithm B.2.1.

1. *Run the MCMC simulation defined in algorithm B.1.1.*
2. *Apply the unimodality test of Hartigan and Hartigan (1985) with a certain confidence level, κ, to the sampled component parameters, $\boldsymbol{\lambda}$.*
3. *If the unimodality test in the second step accepts the null of unimodality, calculate the MAP from simulation results. Otherwise, calculate the BML.*

C. Tables

Table C.1.: **Input parameters.** The table shows all parameter combinations used for the Monte Carlo simulation of the EKOP model. In addition the resulting Probability of informed trading (PIN) is shown in the last column. 'id' indicates the simulation ID and 'N' is the number of trades simulated per dataset.

			EKOP Parameters				
	id	N	α	ϵ	δ	μ	PIN
			News probability	Uninformed trading intensity	Probability of good news	Informed trading intensity	Probability of Informed trading
Group A							
	1	100	0.5003	0.1757	0.3491	0.1320	0.1581
	2	100	0.4340	0.0240	0.4444	0.0301	0.2144
	3	100	0.3563	0.0096	0.5018	0.0157	0.2254
Group B							
	4	100	0.2000	0.0500	0.5000	0.0200	0.0385
	5	100	0.3000	0.0500	0.5000	0.0200	0.0566
	6	100	0.4000	0.0500	0.5000	0.0200	0.0741
	7	100	0.5000	0.0500	0.5000	0.0200	0.0909
Group C							
	8	100	0.2000	0.1000	0.5000	0.1000	0.0909
	9	100	0.3000	0.1000	0.5000	0.1000	0.1304
	10	100	0.4000	0.1000	0.5000	0.1000	0.1667
	11	100	0.5000	0.1000	0.5000	0.1000	0.2000
Group A							
	12	250	0.5003	0.1757	0.3491	0.1320	0.1581
	13	250	0.4340	0.0240	0.4444	0.0301	0.2144
	14	250	0.3563	0.0096	0.5018	0.0157	0.2254
Group B							
	15	250	0.2000	0.0500	0.5000	0.0200	0.0385
	16	250	0.3000	0.0500	0.5000	0.0200	0.0566
	17	250	0.4000	0.0500	0.5000	0.0200	0.0741
	18	250	0.5000	0.0500	0.5000	0.0200	0.0909
Group C							
	19	250	0.2000	0.1000	0.5000	0.1000	0.0909
	20	250	0.3000	0.1000	0.5000	0.1000	0.1304
	21	250	0.4000	0.1000	0.5000	0.1000	0.1667
	22	250	0.5000	0.1000	0.5000	0.1000	0.2000

Table C.2.: **Estimation results (100 observations).** The table shows the estimation results from Monte Carlo simulations of the EKOP model and the compressed model on $N = 100$ data points generated with the `Mersenne-Twister` RNG in `R`. All parameter values are Monte Carlo means with corresponding Monte Carlo standard errors. For each parameter combination 250 simulations were conducted. Each parameter set has been estimated using `R`. For the MCMC approach we used a burn-in of $M_0 = 1000$ iterations and ran the Gibbs sampler for $M = 10000$ iterations.

			MCMC PIN Estimates			**ML PIN Estimates**		
	id	True PIN	BML	MAP	EIA	EKOP	EKOPM	COMPML
			Std. err.					
Group A								
	1	0.1581	0.1577	0.1569	0.1569	0.1576	0.1239	0.1575
			0.0162	0.0153	0.0151	0.0151	0.0115	0.0161
	2	0.2144	0.2184	0.2192	0.2195	0.2173	0.1653	0.2168
			0.0417	0.0362	0.0345	0.0279	0.0231	0.0397
	3	0.2254	0.2369	0.2473	0.2480	0.2658	0.2133	0.2626
			0.0698	0.0489	0.0467	0.0286	0.0287	0.0519
Group B								
	4	0.0385	0.0728	0.0468	0.0518	0.0521	0.0410	0.0694
			0.0637	0.0372	0.0267	0.0222	0.0191	0.0584
	5	0.0566	0.0813	0.0606	0.0620	0.0649	0.0483	0.0778
			0.0635	0.0379	0.0291	0.0228	0.0192	0.0531
	6	0.0741	0.0943	0.0672	0.0670	0.0778	0.0558	0.0902
			0.0629	0.0402	0.0331	0.0263	0.0201	0.0706
	7	0.0909	0.1029	0.0737	0.0714	0.0904	0.0635	0.0933
			0.0669	0.0370	0.0302	0.0219	0.0197	0.0605
Group C								
	8	0.0909	0.0894	0.0962	0.0988	0.0895	0.0735	0.0892
			0.0181	0.0170	0.0165	0.0176	0.0145	0.0181
	9	0.1304	0.1285	0.1325	0.1341	0.1290	0.1051	0.1277
			0.0188	0.0178	0.0173	0.0177	0.0143	0.0184
	10	0.1667	0.1676	0.1690	0.1696	0.1671	0.1337	0.1675
			0.0205	0.0193	0.0189	0.0182	0.0142	0.0201
	11	0.2000	0.2012	0.2002	0.2000	0.2013	0.1581	0.2007
			0.0217	0.0206	0.0203	0.0178	0.0140	0.0212

Table C.3.: **Estimation results (250 observations).** The table shows the estimation results from Monte Carlo simulations of the EKOP model and the compressed model on $N = 250$ data points generated with the `Mersenne-Twister` RNG in `R`. All parameter values are Monte Carlo means with corresponding Monte Carlo standard errors. For each parameter combination 250 simulations were conducted. Each parameter set has been estimated using `R`. For the MCMC approach we used a burn-in of $M_0 = 1000$ iterations and ran the Gibbs sampler for $M = 10000$ iterations.

			MCMC PIN Estimates			**ML PIN Estimates**		
	id	True PIN	BML	MAP	EIA	EKOP	EKOPM	COMPML
			Std. err.					
Group A								
	12	0.1581	0.1584	0.1581	0.1580	0.1583	0.1250	0.1582
			0.0100	0.0098	0.0097	0.0093	0.0072	0.0099
	13	0.2144	0.2165	0.2169	0.2171	0.2148	0.1614	0.2177
			0.0241	0.0226	0.0221	0.0174	0.0147	0.0239
	14	0.2254	0.2260	0.2350	0.2377	0.2613	0.2194	0.2628
			0.0400	0.0339	0.0317	0.0261	0.0164	0.0455
Group B								
	15	0.0385	0.0579	0.0606	0.0604	0.0555	0.0486	0.0687
			0.0400	0.0252	0.0218	0.0134	0.0114	0.0299
	16	0.0566	0.0699	0.0710	0.0697	0.0625	0.0533	0.0716
			0.0429	0.0200	0.0193	0.0151	0.0129	0.0237
	17	0.0741	0.0857	0.0793	0.0773	0.0776	0.0594	0.0842
			0.0465	0.0227	0.0228	0.0154	0.0119	0.0297
	18	0.0909	0.1016	0.0836	0.0808	0.0928	0.0643	0.0924
			0.0480	0.0247	0.0239	0.0153	0.0119	0.0393
Group C								
	19	0.0909	0.0907	0.0935	0.0946	0.0908	0.0749	0.0907
			0.0117	0.0113	0.0112	0.0111	0.0089	0.0117
	20	0.1304	0.1311	0.1326	0.1333	0.1311	0.1067	0.1309
			0.0127	0.0124	0.0123	0.0118	0.0094	0.0128
	21	0.1667	0.1668	0.1673	0.1677	0.1662	0.1336	0.1667
			0.0127	0.0124	0.0123	0.0111	0.0087	0.0128
	22	0.2000	0.1995	0.1991	0.1991	0.1997	0.1573	0.1992
			0.0127	0.0124	0.0124	0.0109	0.0084	0.0125

Table C.4.: **Bias.** The table shows the bias for the Markov chain Monte Carlo and the maximum likelihood estimators. The bias is computed over all 250 simulations corresponding to one combination of true parameters.

		MCMC			**ML**		
	id	BML $[10^{-2}]$	MAP $[10^{-2}]$	EIA $[10^{-2}]$	EKOP $[10^{-2}]$	EKOPM $[10^{-2}]$	COMPML $[10^{-2}]$
Group A							
	1	−0.0477	−0.1198	−0.1265	−0.0556	−3.4190	−0.0600
	2	0.4041	0.4796	0.5106	0.2951	−4.9056	0.2439
	3	1.1525	2.1935	2.2594	4.0392	−1.2085	3.7194
Group B							
	4	3.4300	0.8372	1.3357	1.3684	0.2580	3.0950
	5	2.4670	0.4002	0.5375	0.8310	−0.8337	2.1189
	6	2.0209	−0.6858	−0.7074	0.3774	−1.8267	1.6146
	7	1.2025	−1.7202	−1.9537	−0.0549	−2.7418	0.2396
Group C							
	8	−0.1494	0.5319	0.7872	−0.1369	−1.7376	−0.1757
	9	−0.1933	0.2071	0.3625	−0.1392	−2.5312	−0.2723
	10	0.0923	0.2299	0.2917	0.0443	−3.2953	0.0856
	11	0.1186	0.0199	0.0032	0.1296	−4.1938	0.0690
Group A							
	12	0.0236	−0.0045	−0.0092	0.0155	−3.3109	0.0101
	13	0.2086	0.2530	0.2688	0.0416	−5.2964	0.3346
	14	0.0641	0.9619	1.2311	3.5861	−0.5952	3.7377
Group B							
	15	1.9396	2.2108	2.1895	1.7074	1.0185	3.0228
	16	1.3310	1.4440	1.3071	0.5946	−0.3338	1.5027
	17	1.1582	0.5259	0.3185	0.3537	−1.4699	1.0079
	18	1.0718	−0.7266	−1.0095	0.1918	−2.6581	0.1442
Group C							
	19	−0.0198	0.2549	0.3672	−0.0148	−1.6019	−0.0198
	20	0.0628	0.2200	0.2882	0.0629	−2.3761	0.0484
	21	0.0094	0.0643	0.0997	−0.0449	−3.3102	0.0031
	22	−0.0467	−0.0903	−0.0884	−0.0339	−4.2671	−0.0796

Table C.5.: **RMSE.** The table shows the root mean squared error (RMSE) for the Markov chain Monte Carlo and the maximum likelihood estimators. The RMSE is computed over all 250 simulations corresponding to one combination of true parameters.

		MCMC			**ML**		
	id	BML $[10^{-2}]$	MAP $[10^{-2}]$	EIA $[10^{-2}]$	EKOP $[10^{-2}]$	EKOPM $[10^{-2}]$	COMPML $[10^{-2}]$
Group A							
	1	1.6181	1.5355	1.5075	1.5036	3.6059	1.6107
	2	4.1824	3.6434	3.4774	2.8041	5.4224	3.9691
	3	7.0619	5.3535	5.1755	4.9478	3.1126	6.3727
Group B							
	4	7.2277	3.8106	2.9826	2.6003	1.9240	6.5989
	5	6.8029	3.8037	2.9500	2.4228	2.0913	5.7056
	6	6.5974	4.0660	3.3788	2.6491	2.7124	7.2290
	7	6.7831	4.0744	3.5878	2.1828	3.3718	6.0374
Group C							
	8	1.8092	1.7735	1.8234	1.7599	2.2582	1.8111
	9	1.8835	1.7870	1.7684	1.7761	2.9047	1.8561
	10	2.0449	1.9377	1.9106	1.8132	3.5879	2.0106
	11	2.1683	2.0520	2.0231	1.7822	4.4216	2.1170
Group A							
	12	0.9990	0.9759	0.9687	0.9277	3.3879	0.9917
	13	2.4121	2.2740	2.2178	1.7415	5.4949	2.4054
	14	3.9893	3.5171	3.3920	4.4299	1.7373	5.8770
Group B							
	15	4.4370	3.3470	3.0876	2.1679	1.5298	4.2442
	16	4.4796	2.4656	2.3284	1.6244	1.3347	2.8029
	17	4.7844	2.3242	2.2975	1.5728	1.8893	3.1273
	18	4.9136	2.5663	2.5870	1.5403	2.9095	3.9272
Group C							
	19	1.1663	1.1571	1.1780	1.1040	1.8324	1.1686
	20	1.2734	1.2545	1.2612	1.1765	2.5550	1.2747
	21	1.2714	1.2405	1.2348	1.1128	3.4215	1.2742
	22	1.2688	1.2447	1.2384	1.0926	4.3492	1.2537

Table C.6.: **Convergence errors and frequencies.** The table shows the frequencies of Bayesian estimations with lower absolute bias than the corresponding MLE from either EKOP, EKOPM, or COMPML. In addition the number of convergence errors in the gradient method for the maximum likelihood estimations are displayed. In case the `L-BFGS-B` algorithm did not converge a bounded `Nelder-Mead` routine was applied.

		Conv. Errors			**MCMC vs. EKOP**			**MCMC vs. EKOPM**			**MCMC vs. COMPML**		
	id	EKOP	EKOPM	COMPML	BML	MAP	EIA	BML	MAP	EIA	BML	MAP	EIA
					[%]	[%]	[%]	[%]	[%]	[%]	[%]	[%]	[%]
Group A													
	1	5	6	4	42	42	43	88	91	91	52	64	73
	2	21	29	17	34	39	39	72	77	78	48	71	72
	3	48	52	40	45	54	53	26	32	34	51	62	64
Group B													
	4	11	26	17	34	28	50	28	26	45	54	56	66
	5	18	24	16	29	36	44	30	29	40	46	58	62
	6	16	18	19	21	33	36	27	40	40	41	60	58
	7	20	15	20	18	34	27	38	54	49	41	53	52
Group C													
	8	10	5	17	46	51	47	72	66	64	49	50	46
	9	12	10	14	42	47	50	78	76	76	47	51	48
	10	18	7	10	40	46	49	84	84	83	41	64	60
	11	8	11	8	34	42	43	88	90	90	50	72	72
Group A													
	12	8	2	8	42	44	46	98	99	99	44	58	67
	13	33	28	19	39	39	40	92	92	92	54	66	72
	14	36	41	39	57	63	68	26	25	25	72	77	79
Group B													
	15	25	31	28	50	34	38	34	14	22	62	68	71
	16	19	28	24	28	41	39	24	31	28	46	59	59
	17	14	16	18	21	36	32	29	46	44	38	56	57
	18	27	30	32	18	32	33	40	68	65	28	52	48
Group C													
	19	13	4	8	44	44	41	81	78	75	44	46	43
	20	11	13	6	44	52	51	85	84	84	51	47	45
	21	8	7	11	39	41	43	94	95	95	46	62	51
	22	6	9	12	38	40	41	100	100	100	50	60	65

Table C.7.: **Frequencies (bias-reducing algorithm).** The table shows the frequencies of Bayesian estimations with lower absolute bias than the corresponding MLE from either EKOP, EKOPM, or COMPML when using the ALGBM. In addition the number of rejected unimodality tests are presented.

	id	Rejected unimodality tests	**ALGBM vs. MLE** EKOP [%]	EKOPM [%]	COMPML [%]
Group A					
	1	250	42	88	52
	2	250	34	72	48
	3	250	45	26	51
Group B					
	4	98	37	33	57
	5	148	34	33	46
	6	154	20	28	39
	7	177	17	34	36
Group C					
	8	250	46	72	49
	9	250	42	78	47
	10	250	40	84	41
	11	250	34	88	50
Group A					
	12	250	42	98	44
	13	250	39	92	54
	14	250	57	26	72
Group B					
	15	207	50	37	63
	16	234	29	25	45
	17	233	21	29	37
	18	241	17	40	27
Group C					
	19	250	44	81	44
	20	250	44	85	51
	21	250	39	94	46
	22	250	38	100	50

Bibliography

Basford, K., Greenway, D., McLachlan, G., Peel, D., 1997. Standard errors of fitted component means of normal mixtures. Computational Statistics 12 (1), 1–18.

Boehmer, E., Grammig, J., Theissen, E., 2007. Estimating the probability of informed trading—does trade misclassification matter? Journal of Financial Markets 10 (1), 26–47.

Celeux, G., 1998. Bayesian inference for mixture: The label switching problem. In: Green, P., Rayne, R. (Eds.), COMPSTAT 98. pp. 227–232.

Celeux, G., Hurn, M., Robert, C. P., 2000. Computational and inferential difficulties with mixture posterior distributions. Journal of the American Statistical Association 95 (451), 957–970.

Chung, H., Loken, E., Schafer, J. L., 2004. Difficulties in drawing inferences with finite-mixture models. The American Statistician 58 (2).

Easley, D., Kiefer, N., O'Hara, M., Paperman, J., 1996. Liquidity, information, and infrequently traded stocks. Journal of Finance 51, 1405–1436.

Easley, D., O'Hara, M., 1987. Price, trade size, and information in securities markets. Journal of Financial economics 19 (1), 69–90.

Easley, D., O'Hara, M., 1992. Time and the process of security price adjustment. The Journal of finance 47 (2), 577–605.

Eddelbuettel, D., Sanderson, C., 2013. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. Computational Statistics and Data Analysis in press.
URL `http://dx.doi.org/10.1016/j.csda.2013.02.005`

Ellis, K., Michaely, R., O'Hara, M., 2000. The accuracy of trade classification rules: Evidence from Nasdaq. The Journal of Financial and Quantitative Analysis 35 (4), 529–551.

Finch, S. J., Mendell, N. R., Henry Jr, C., 1989. Probabilistic measures of adequacy of a numerical search for a global maximum. Journal of the American Statistical Association 84 (408), 1020–1023.

Frühwirth-Schnatter, S., 2001. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association 96 (453), 194–209.

Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer Series in Statistics. Springer.

Hartigan, J. A., Hartigan, P., 1985. The dip test of unimodality. The Annals of Statistics, 70–84.

Hurn, M., Justel, A., Robert, C. P., 2003. Estimating mixtures of regressions. Journal of Computational and Graphical Statistics 12 (1), 55–79.

Jackson, D., 2007. Inferring trader behavior from transaction data: A trade count model. Journal of Economics and Finance 31 (3), 283–301.

Jasra, A., Holmes, C. C., Stephens, D. A., 2005. Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. Statistical Science 20 (1), 50–67.

Kelley, C. T., 1999. Iterative methods for optimization. Vol. 18. Siam.

Kokot, S., 2004. The Econometrics of Sequential Trade Models: Theory and Applications Using High Frequency Data. Springer.

Lee, C., Ready, M. J., 1991. Inferring trade direction from intraday data. The Journal of Finance 46 (2), 733–746.

Lee, C. M. C., Radhakrishna, B., 2000. Inferring investor behavior: Evidence from TORQ data. Journal of Financial Markets 3 (2), 83–111.

Nagode, M., 2014. rebmix: The Rebmix package.

Nagode, M., Fajdiga, M., 2011. The REBMIX algorithm and the univariate finite mixture estimation. Communications in Statistics—Theory and Methods 40 (5), 876–892.

Odders-White, E. R., 2000. On the occurrence and consequences of inaccurate trade classification. Journal of Financial Markets 3 (3), 259–286.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org`

Reddy, C. K., Rajaratnam, B., 2010. Learning mixture models via component-wise parameter smoothing. Computational Statistics & Data Analysis 54 (3), 732–749.

Redner, R. A., Walker, H. F., 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM review 26 (2), 195–239.

Robert, C., 2007. The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer.

Sanderson, C., et al., 2010. Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments. Tech. rep., Technical report, NICTA.

Stephens, M., 1997a. Bayesian methods for mixtures of normal distributions. Ph.D. thesis, University of Oxford.

Stephens, M., 1997b. Discussion of the paper by Richardson and Green. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59, 768–769.

Varadhan, R., Johns Hopkins University, Borchers, H. W., ABB Corporate Research, 2011. dfoptim: Derivative-free Optimization. R package version 2011.8-1.
URL http://CRAN.R-project.org/package=dfoptim

Viallefont, V., Richardson, S., Green, P. J., 2002. Bayesian analysis of Poisson mixtures. Journal of Nonparametric Statistics 14 (1-2), 181–202.

Wand, M., Ripley, B., 2006. Kernsmooth: Functions for kernel smoothing for Wand & Jones (1995). R package version, 2–22.

Wand, M. P., Jones, M. C., 1994. Kernel smoothing. Vol. 60. Crc Press.

Yee, T. W., 2013. VGAM: Vector Generalized Linear and Additive Models. R package version 0.9-2.
URL http://CRAN.R-project.org/package=VGAM

Zhu, C., Byrd, R. H., Lu, P., Nocedal, J., 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Transactions on Mathematical Software (TOMS) 23 (4), 550–560.

Part III.

The probability of informed trading on U.S. corporate bond markets: Conclusions from a fragmented market

This article investigates information asymmetries in U.S. corporate bond markets using transaction data from the Trade Reporting and Compliance Engine (TRACE) for constituents of the S&P 500 in the first half-year of 2011. As a measurement of information asymmetry I employ the probability of informed trading (PIN) proposed by Easley, Kiefer, O'Hara, and Paperman (1996). In a cross-sectional regression of 4,155 fixed-income securities on bond characteristics, market variables, and stock statistics, I find that nearly 50% of the variation in PINs is explained. All estimated coefficients conform to expectations. While a comparison of PINs in bond and corresponding equity markets confirms prior findings of lower PINs on more active stock markets, it indicates the reverse for fixed-income securities: Less-frequently traded bonds exhibit lower PINs. These findings accord with there being lower transaction costs on less active bond markets as found by Goldstein, Hotchkiss, and Sirri (2007). However, as news probabilities for bonds from the same issuer and bonds and corresponding stocks differ significantly I question the appropriateness of traditional models for measuring information asymmetries. The probability of informed trading might not be a suitable measure for highly fragmented markets such as the U.S. corporate bond market.

1. Introduction

The secondary market for U.S. corporate bonds is organized as an 'over-the-counter' market with fragmented liquidity pools. The unavailability of transaction data left empirical research in the field of market microstructure to fortunate authors in possession of proprietary data sets directly provided by market participants. The field changed entirely after the introduction of the Trade Reporting and Compliance Engine (TRACE) in 2002 through efforts of the SEC. TRACE collects trade reports from nearly all corporate bonds traded in the United States and makes these data accessible to researchers. Since its introduction, empirical research on bond trading has proliferated and offered important insights into trading costs and market efficiency in general.

Although the TRACE data have been available since 2002, papers concerned with information asymmetry in corporate bond markets are rare. This paper begins to close this gap in the literature by investigating the probability of informed trading (PIN), as proposed in a seminal paper by Easley et al. (1996). To the best of my knowledge, I am the first to provide a thorough study of information asymmetries in bond markets.

Cai et al. (2012) provide PIN estimates in their herding study. In contrast to their study, I exclusively focus on information asymmetry. Two questions motivated this paper: (1) What are the causes of information asymmetries and how do other market variables relate to information asymmetry? (2) Are information asymmetries in bond and corresponding stock markets comparable? I begin my analysis with cross-sectional regressions to explain the causes of information asymmetries in U.S. corporate bond markets. For this purpose I select from issuers listed in the S&P 500 in the first half-year of 2011 all available bonds from the enhanced TRACE database with complete information in the FISD Mergent and complete stock data in the Compustat Security Daily databases and estimate the PIN measure for each of the resulting 4,155 bonds. To estimate the PIN I employ a slightly altered version of the Easley et al. (1996) (EKOP) model that is based on a compressed likelihood and only requires data on the number of trades. I apply a Bayesian approach first proposed in Grammig et al. (2014) and considered to be a robust procedure with reliable convergence. I then regress the PIN measure on several bond-, stock- and market-specific variables. The results indicate that nearly 50% of the variation in estimated PINs can be explained by bond characteristics and market measures and certain issuer characteristics.

In the second part of my analysis I compare PIN measures from U.S. corporate bond and corresponding equity markets. Similar to Easley et al. (1996), I divide my sample by trading volume and investigate information asymmetries among different volume deciles. My bond sample comprises 108 companies from the S&P 500 composite index and market data from 1 January 2011 to 30 June 2011. My results are the following: While PIN estimates for the stock

markets reveal a structure reported in previous studies of higher information asymmetry in less active markets, measures for bond markets indicate the reverse. Quite infrequently traded bonds appear to suffer less from information asymmetries than bonds with high trading volumes. Furthermore, news probabilities for stocks and bonds of the same company differ significantly and this could be considered conflicting evidence. Statistical tests reinforce my findings and allow me to assess whether the EKOP and the compressed model can be appropriately applied to fragmented markets.

My findings may be relevant for regulators and issuers. Referring to the recently published agenda for the roundtable of fixed income markets of the SEC many yet unanswered questions relate to information asymmetries and their drivers.[1] For debt issuers information asymmetries appear to be important, as reported in the recent paper of Han and Zhou (2013). The authors demonstrate that information asymmetries get priced and this in turn increases the issuer's cost of debt. My results indicate that there exist factors that there are factors under the issuer's direct control and could therefore be used to design a debt obligation that accomodates the influences issuer's characteristics on information asymmetries.

The remainder of this paper is structured as follows. Section 2 describes the market structure of the U.S. corporate debt market and related literature. Section 3 contains details on the data sets used in my analyses and provides some summary statistics. In section 4 estimation methods and results are discussed and section 5 concludes.

2. Market structure and literature

Corporate bonds represent in addition to equity an important capital resource for a company. They are contractual obligations to make a series of 'coupon' payments and repay the 'face value' at maturity. Coupons are either fixed or variable, where in the latter case the coupon is tied to prevailing interest rates. There are also 'zero-coupon' bonds that only pay a face value after a pre-fixed time horizon. The typical coupon payment frequency is twice per year, however, monthly or yearly payments are not unusual. Bonds are issued on the so called 'primary market' and afterwards trade on the 'secondary market' investigated in this paper. The market values of bonds depend on general economic variables, most notably the interest rate and on company-related factors such as the default risk. In contrast to equity, bonds are issued in a series of blocks where each block defines a distinct contract between the issuer and all buy-side investors. This contract also occasionally involves special options that are also priced, such as a rebuy option for the issuer.

If intended for public sale, the bond has to be registered with the SEC. An important aspect of bond issuance and subsequent trading on secondary markets is a publicly available rating

[1] For the agenda of the roundtable for fixed income markets see `http://www.sec.gov/spotlight/fixed-income-markets/fixed-income-markets-agenda.htm` (last visited on 11 November 2014).

issued by one of the main rating agencies, Fitch, Standard & Poor's, Moody's or Duff & Phelps.[2] Issuers directly hire a rating agency to signal their credit worthiness and improve investors' trust in the payment obligation.

The annual value of newly issued bonds in the United States is enormous. In 2011, included in my data set, companies issued a volume of $1,005.5 billions new debt in U.S. bond markets with $8,324.7 billions in totals debt outstanding.[3] As former SEC chairman Arthur Levitt noted in a speech at the Media Studies Center in New York 1998:[4]

> *There seems to be a common misconception that the bond market is somehow less important than the equity market. But history shows, if anything, that the opposite may be true.*
> (Arthur Levitt, New York, 1998)

Levitt mentions in relation to this comment that the daily transaction volume in stocks on the NYSE is approximately $28 billion whereas the total corresponding volume on all secondary bond markets is over $380 billion per day.[5]

2.1. Trading on secondary markets

After bonds are issued they are traded on secondary markets. The secondary market for U.S. corporate bonds is organized as an 'over-the-counter' (OTC) market where broker-dealers execute the majority of customer transactions in a principal capacity.[6] Biais and Green (2007) provide a comprehensive overview of bond trading history and the development of contemporary markets.

Following the SEC's definition, the OTC corporate bond market does not have a centralized interdealer quotation system and can therefore not be described as either a quote-driven dealer market or an order-driven auction market.[7] Instead, this market is most often referred to as a fragmented market. The way in which trade agreements are arranged does not appear to be standardized: Saunders et al. (2002) find that the trading mechanism closely resembles a first-price sealed-bid auction in which dealers collect bids and offers to match a client's order at the best price. Goldstein et al. (2007), Gu and Zhao (2006), Randall (2013), Harris (2002) and the Securities Industry and Financial Markets Association (2006) mention a bargaining mechanism. As requests for quotations often have to be conducted over the telephone or in person, bargaining can be expected to determine an important share of a price agreement; this, then, particularly holds for large block trades.

[2] These four rating agencies were the only rating agencies assigning ratings to the fixed-income securities in my sample from Mergent FISD.

[3] This information was collected from the Securities Industry and Financial Markets' (SIFMA) website (`http://www.sifma.org/research/statistics.aspx`, February 21, 2014).

[4] See Levitt (1998) for the entire speech.

[5] However, one has to account for the large markets for government bonds, treasury bills and municipal bonds.

[6] 'Over-the-counter' is the established term used by the SEC to describe markets not being legally organized like an 'exchange'.

[7] See Edwards (2006).

2.1.1. Electronic quotation systems

Early attempts to introduce centralized electronic quotation systems into the highly fragmented U.S. bond markets were a costly failure.[8] Edwards (2006) also comments on the failure in mid-1998 to embrace many of the technological innovations in trading that had swept other secondary markets in the United States. Although the attempts to adopt centralized systems were unsuccessful, several dealers and brokers had already implemented proprietary electronic information systems to serve their clients. Some of the large systems currently in use were introduced in these times: Tradeweb by Thomson Reuters and ten leading dealers began operating in 1996 and reported a trading volume of $1 trillion per year in 2000.[9] BondDesk Trading by the BondDesk Group (acquired by Tradeweb in 2013), was founded in 1995 and then developed from thereon services to access liquidity pools, quotations, post-trade information and research support.[10] BGC Partners introduced the eSpeed system in 1996 (acquired by NASDAQ OMX Group Inc. in 2013) and one of the largest players in the market, MarketAxess from ETP, began operating in 2002 with a system for requests of executable bids from multiple broker-dealers.[11] Most of these systems are hybrid voice/electronic systems.

In 2005 the Bond Market Association conducted a study of trading systems among 54 firms that offered electronic trade execution systems in fixed-income securities.[12] The study notes that trading systems in fixed-income markets are characterized by broad variety and can be categorized by the participants trading on them (dealer-customer or dealer-dealer) and the methodologies and technologies used for price discovery and trade execution. The largest trading platform in 2005 registered 7,650 buy-side participants and four other platforms reported over 1,000 registered buy-side firms.

2.1.2. Market participants

Generally, market participants desiring to trade a certain bond contact dealers to obtain quotes and then select the best bid or offer for execution. However, this seemingly simple process involves a series of costly operations in terms of human resources and especially time:

[8] The case of InterVest Financial Services made headlines when its chairman Larry E. Fondren offered an electronic service that attempted to provide information on the bond market - at the time occasionally called a souk. His business was boycotted by bond dealers and closed in 1997 with a loss of more than $10 million. Fondres himself believes that the dealers' rejection of his system was related to an informal investigation by the Justice Department and the SEC in 1996 into whether dealers were colluding to keep InterVest from growing. His story was published in an article in BloombergBusinessweek by Laura Duffie (`http://www.businessweek.com/stories/1998-03-08/never-cross-a-bond-dealer`, last visited on 17 February 2014) and appeared in SEC publications (Edwards (2006)).

[9] For detailed information see `http://www.tradeweb.com`.

[10] Additional information can be gathered from the company's website at: `http://www.bonddeskgroup.com`.

[11] For eSpeed see `http://www.espeed.com`. MarketAxess offers services specifically designed for U.S. and European high-grade corporate bonds, high-yield and emerging market bonds (`http://www.marketaxess.com`).

[12] Bond Market Association (2005) also found in a speech by an SEC commissioner in May 2006 (`http://www.sec.gov/news/speech/2006/spch051906cag.htm`, 17 February 2014).

To obtain the best price several dealers have to be contacted and prices have to be collected and evaluated. As Bessembinder and Maxwell (2008) note, an executable quote by a dealer is typically only valid 'as long as the breath is warm', which limits the ability to obtain multiple quotations before committing to trade. This problem is exacerbated, if the bond is traded infrequently or the dealer has no inventory. In the latter case, dealers have to either reject services to the market participant or engage in a similar search process, although they profit from having far better market insight and substantially more contacts than a trader. If the dealer ultimately executes a client's order, she faces an inventory adjustment and after a number of trades in a certain bond, inventory must be rebalanced to reduce risks. Inventory management may represent 'the' most important aspect of a dealer's business. A formal analysis of this dimension of the dealer's problem is offered in the influential papers of Stoll (1978) and Ho and Stoll (1981) and has been the subject of intensive discussion since. Note however, that these models assume compensation for inventory-based risks through higher spreads, whereas interdealer markets reduce this inventory-based risk by simplifying the management of large positions.

To manage their inventory dealers trade with other dealers. This interdealer business can be considered delicate, especially, if the rebalancing of inventory involves substantial trade sizes. When a dealer approaches another dealer, she reveals her identity and likely sustains negative market impacts due to other dealers running in front of her order or earning significant markups as counterparties. Revealing her identity and needs to trade reduces her bargaining power.

Regarding the interdealer market, Saunders et al. (2002) mention the perhaps least known and least understood but likely most important market participants: The interdealer brokers (IDBs). Interdealer brokers are middlemen in fixed-income trades. It is difficult to find information describing the role of these special market entities in the fixed income markets, despite their importance. This void in the literature is due to their highly confidential services. The Securities Industry and Financial Markets Association (2006) provides an overview of the interdealer brokers' business and cites the enhancement of price discovery and transparency, the provision of anonymity and confidentiality, facilitating of information flow, liquidity enhancement, improving market efficiency, and reducing costs as the general values added by these middlemen to the fixed-income markets. Further, the Securities Industry and Financial Markets Association (2006) stresses that the corporate bond market in particular makes extensive use of IDBs. As mentioned above, the information flow on this market is viscous and transparency is low due to its highly fragmented structure. Search costs can be expected to be very high, in contrast to markets with a centralized structure such as the stock market, and the transaction of large blocks becomes especially difficult in such an environment. As a result, interdealer brokers developed naturally as there was a need for intermediaries to make markets efficient and it can be suspected that their special role might have significant effects on information asymmetries in this market.

Before the SEC introduced TRACE, interdealers also constituted the main source for price transparency in the market. Through proprietary information systems IDBs provided pre-trade transparency by aggregating multiple orders and dealer quotations either on an indicative or an executable basis. These systems were voice, electronic or hybrid brokerage models. Their

assurance of anonymity encouraged dealers to supply IDBs with their quotations, typically in the form of calls received seeking bonds or seeking a buyer for their bonds. IDBs' aggregation of trade interest and the subsequent dissemination of this information to participants enhanced price formation, price discovery and trade frequency. Post-trade transparency was facilitated by information on whether the bonds were reoffered or concerning the 'cover bid', the next best bid after the level at which a bond traded. The Securities Industry and Financial Markets Association (2006) considers this information flow to be important for dealers to quote markets with certainty, thereby reducing spreads.

In interdealer trades IDBs preserve the anonymity of their clients and reduce the market impact of inventory rebalancing orders. In addition, their broad range of contacts with other dealers and brokers or large institutional traders enables them to divide a large block among several potential buyers and sellers; this keeps markets more stable. Moreover, the business model of IDBs allows dealers to better serve their clients' needs: If a client requests a bond that is not in the dealer's inventory at a given point in time, the dealer can simply contact an IDB and request a quote for the security. The network between various market players often forms chains in which securities pass through more than three hands: From a client to a dealer through an IDB to another dealer who in turn transfers the bond to her client (if the client acts via a broker the chain becomes even longer).[13] Serving as the key party in a large number of trades in U.S. bond markets, IDBs should be expected to learn permanently from orderflow and from the revealed identities of dealers and traders that are contacted directly.

As mentioned by the Securities Industry and Financial Markets Association (2006), the interdealer broker business is considered a pure agency function and therefore as predominantly riskless. There are a few large IDBs in the corporate bond market, among them ICAP, BGC Partners, Tradition, GFI Group and Mint Partners which means that competition is questionable.[14] Recently, IDBs emerged as key enablers of the Libor scandal in 2010. They are considered to have played the role of central middleman in Libor-rigging.[15] It is assumed that the specific market environment in which traders know one another made the assumed collusion possible.

However, interdealer brokers are not the only actors that possess delicate private information: Former SEC chairman Arthur Levitt reported in his speech at the Media Studies Center in New York, September 1998, that the SEC found anecdotal evidence of the possible misuse of insider information in the high-yield bond market. Unidentified sources informed the SEC, that investment banks and institutional investors that purchase high-yield corporate bonds occasionally participate in loan syndicates for those same companies issuing high-yield bonds. At the time the chairman was in contact with the FINRA (formerly NASD) concerning the introduction of a standardized reporting system for the corporate bond market: TRACE.

[13] See Li and Schürhoff (2012) for an explicit overview over dealer networks.

[14] See `http://www.icap.com`, `http://www.bgcpartners.com`, `http://www.gfigroup.com` and `http://www.mintpartners.com/default.aspx` for further information.

[15] Article published in BloombergBusinesweek in 2013 by Gavin Finch and Liam Vaughan (`http://www.bloomberg.com/news/2013-02-08/interdealer-brokers-emerge-as-key-enablers-in-libor-scandal.html`, last visited on February 17th, 2014).

2.1.3. Introduction of TRACE

After the introduction of the world-wide electronic reporting system GovX in 1991 for government securities and the SEC's encouragement of the Municipal Securities Rulemaking Board (MSRB) to collect the details of dealer-to-dealer transactions in the municipal bond market, Arthur Levitt believed that it was time for the corporate bond market to, in his words, 'step up to the plate' as he said. In the following years, the Trade Reporting And Compliance Engine (TRACE) was introduced in three large phases and replaced the prior system FIPS.[16] TRACE requires brokers and dealers to report virtually all trades in the fixed-income securitized product market, stores these reports and makes them available. Asquith et al. (2013) describe the introduction of TRACE in detail.

In its news release, FINRA (2005) cites the positive effects of the introduction of TRACE. While FINRA emphasizes its role to be the leading private-sector provider of financial regulatory services, it notes increased interest among individual investors in the bond market following the sharp decline in the equity market at the time. Following the report, "[t]wo out of three corporate bond transactions are carried out at the retail leve[l]".[17] Edwards (2006) mentions the testimony of Doug Shulman (FINRA) before the US Senate Committee, in which he declared, "[c]ontrary to popular belief [...] the bond market has a substantial retail participatio[n]". Table 1 in Edwards (2006) demonstrates, that while the average trade size in corporate bonds is 788 bonds in 2003 and 2004, median trade size is only 32. Institutional investors determine transaction volumes, while the number of trades is dominated by the retail sector. Retail trading is generally considered to be uninformative and hence this development in the U.S. corporate bond market should have an effect on information asymmetries.

While TRACE has improved post-trade transparency no effects are to be observed for pre-trade transparency. Indeed, as Edwards (2006) reports, the pre-trade transparency formerly provided by FIPS had been deliminated.[18] The Securities Industry and Financial Markets Association (2006) mentions that the benefits of market data distribution - long championed by IDBs - have now institutionalized by regulators. However, IDBs still provide pre-trade transparency to their clients; therefore it is unlikely that retail investors are able to afford registration (they could only access such services via their broker). While post-trade transparency was substantially improved through TRACE, the pre-trade transparency and fragmented market structure of the corporate bond market remain unchanged. The market microstructure and the initiation of trades are particularly important determinants of the distribution of information among market participants. With middlemen masking the identities of their clients and maintaining personal contacts with them and dealers and large institutional investors enjoying long-time profitable relationships, information asymmetries could become crucial for market efficiency in

[16] The Fixed Income Reporting System (FIPS) had been used predominantly for high-yield bonds and operated from 1994 to 2002. This reporting system resulted from a regulatory intervention to better monitor insider trading in this security segment; see Edwards (2006).

[17] It should be mentioned here that the literature classifies trades below a $100,000 par as 'retail' and the average par of a bond is $1,000.

[18] FIPS also reported one-sided quotations for 50 bonds.

an environment with a growing retail segment.

This paper analyzes post-TRACE information asymmetry in the U.S. corporate bond market and concludes, that information asymmetries are highly volitile in comparison to the equity sector though primarily driven by an insufficient amount of uninformed trading. Additionally, this paper challenges the traditional models for measuring information asymmetries when investigating a highly fragmented market and poses the question of whether research should then continue to rely on these models' results.

2.2. Related literature

Market microstructure models including information asymmetries in securities trading were first developed in the 1990s by Copeland and Galai (1983), Glosten and Milgrom (1985) and Easley and O'Hara (1987) but assume a centralized dealer-market underlying. Copeland and Galai (1983) analyze a one-period model of a dealer's pricing problem under the assumption that certain traders possess superior information. As, on average, the market maker looses to those traders, she compensates for her risk of losing by setting a spread. While the market maker's loss in the model of Copeland and Galai (1983) depends on the demand elasticities of traders (and thus on the current bid and ask prices), Glosten and Milgrom (1985) relate this loss to the speed of information inbound into prices and set thereby the foundation for the understanding of how markets learn from information in order flow. Although this study does not focus on information asymmetry reflected in spreads, this important strand of the literature must be mentioned for the sake of completeness. Easley and O'Hara (1987) describe a model similar in the spirit to Glosten and Milgrom (1985) and consider trade size as a second dimension of order flow information. They conclude that informed traders have an incentive to trade in large order sizes. My analysis relates to their results in that many large orders submitted to the market appear to have a positive effect on information asymmetry.

Empirical models to measure the impact of information asymmetry on transaction costs have been published for example by Glosten and Harris (1988), Hasbrouck (1988), Huang and Stoll (1997), and Madhavan et al. (1997), and are classified as the term 'trade indicator models'. These empirical studies obtain results consistent with theory in that information asymmetries result in higher spreads. Hasbrouck (1988) also finds evidence in his NYSE data set that large trades contain more information.

In their seminal paper Easley et al. (1996) propose a new technique to measure information asymmetry in financial markets. Rather than using prices to obtain indirect evidence of informed trading, the authors directly measure the effect of informed trading by estimating the market maker's beliefs (see Easley et al. (1996, p.1407)). The most important empirical result of Easley et al. (1996) is that the probability of informed trading is lower for high volume stocks. My study follows their approach and concludes that this pattern may not hold when considering debt markets.

The important segment of interdealer markets with their middlemen, the IDBs, is studied in many papers, although for the government bond market, the commodities market and the

FX market rather than for the U.S. corporate bond market. Theoretical models that involve interdealer markets are developed in Lyons (1997) and Calcagno and Lovo (2006). Lyons (1997) develops a model for the foreign exchange market in which dealers trade with customers and then among themselves. The basic premise of his model is that when a dealer engages in inventory balancing via the interdealer market, not only the customer's trade but also any information contained therein is passed from dealer to dealer. This information is subsequently revealed in prices depending on the information content in interdealer markets. Lyons (1997) terms the passing of customer trades among dealers 'hot-potato' trading and demonstrates that it reduces the information in interdealer trades, thus making prices less informative. Calcagno and Lovo (2006) develop a theoretical model with a particular focus on information asymmetry in fragmented markets. In their model, a market maker with superior information competes with market makers that are uninformed. To summarize the results of their study, the informed market maker has positive expected pay-offs due to price leadership. These findings illustrate that adverse selection does not exist solely in dealer-customer relationships but that there are also profitable opportunities from using private information in interdealer markets.

Huang et al. (2002) analyze the U.S. Treasury note interdealer market. The authors examine the trading behavior of primary dealers in the 5-year Treasury note interdealer broker market. They find that trading frequency is determined by public information and private information, represented by the dealers' inventory and order flow information. Additionally, they mention that trade size appears to have no information content in this market. As mentioned above, my results contrast those of Huang et al. (2002) in that trade size is important.

Evans and Lyons (1999) and Payne (2003) observe information content in FX order flow data from direct interdealer trading platforms and broker platforms, respectively. Marsh and O'Rourke (2005) also observe private information in the FX customer order flow by analyzing the order flow of a leading European commercial bank's foreign exchange desk. In addition, the order flows of different exchange rates contain information relevant for other exchange rates. As an interesting byproduct of their study, they mention that the dealer is also able to extract information from the trading accounts of its customers: Citibank clients appear to be better informed on average. Thus, even if the clients themself remain anonymous dealers distinguish between order flows from different brokers (accounts). Furthermore, Bjønnes and Rime (2005) analyze the proprietary order flow data of four interbank dealers in the foreign exchange spot market and find strong support for an information effect in incoming trades. They report that the direction of trade is the most important factor but that trade size is also relevant.

Menkhoff and Schmeling published a series of papers investigating the behavior and characteristics of informed orderflow. Menkhoff and Schmeling (2008) demonstrate that there is only a permanent price impact for orders from certain geographical regions, typically the regions where political and financial decision making occurs. In Menkhoff and Schmeling (2010b) the authors report that informed traders are characterized by simultaneously having large trading volumes, medium-sized orders, trading early during the trading session and at times with a thin order book plus wide spreads, and finally by being located in a financial center. Menkhoff et al. (2010) analyze the limit-order submission strategies of informed traders and conclude that

informed traders are sensitive to spreads, volatility, momentum and depth. Furthermore, they find that the informed treat aggressive limit-orders as a substitute for market-orders. Finally, Menkhoff and Schmeling (2010a) focus on information about the counterparty in FX trades and observe significant effects on the future trading decisions of individual traders. The effect is such that traders tend to reverse their order flows to align them with those of better-informed counterparties.

Studies specifically concerning with the U.S. corporate bond market have proliferated following the increased data availability offered by TRACE. Important studies concerning transaction costs and price efficiency are Edwards et al. (2006), Bessembinder et al. (2006), Goldstein et al. (2007), and most recently, Asquith et al. (2013). All studies report a significant decrease in transaction costs and price dispersion due to the increase in transparency. In addition, transaction costs in more active markets are found to be significantly higher.[19] Asquith et al. (2013) also mention a negative effect of TRACE: The trading activity for certain categories of bonds decreased.

Han and Zhou (2013) measure information asymmetry in bond markets using a trade indicator model and explain U.S. corporate bond yield spreads and default rates using bond characteristics, general economic variables and their information asymmetry measure. In contrast to their study my paper does not use a trade indicator model but a more general measure of information asymmetry. I extend their study by describing the causes of information asymmetry, which might in turn be priced following their explanation of yield spreads.

Cai et al. (2012) also calculate the probability of informed trading in their study, although their focus is on institutional herding in U.S. corporate bond trading. The authors estimate the model of Easley et al. (1996) using data on sell and buy trades and the range of their estimates is similar to my results. In contrast to their study I exclusively focus on the probability of informed trading, cover a broad range of stylized facts associated with this measure, and compare the probabilities of informed trading in fixed-income and equity markets.

Studies concerned with explaining information asymmetries in debt markets are rare. Wittenberg-Moerman (2008) considers the secondary loan market and measures the effects of loan characteristics on information asymmetry. As a measure of information asymmetry, the author adopts the bid-ask spread, a debatable measure with respect to the above mentioned empirical findings in Madhavan et al. (1997), Glosten and Harris (1988) and Huang and Stoll (1997). Information asymmetries determine only one of several components contained in the spread, and it remains questionable whether variation in spreads resemble variations in information asymmetries.

[19] Goldstein et al. (2007, Section 3.3) find their estimated spreads to increase in the level of trading activity. The authors offer the explanation that dealers might trade less active bonds differently than more active bonds.

3. Data

I used the enhanced TRACE data set for bond order flow data, the Mergent FISD data set for information on ratings and issuance, the former online database of the Financial Markets Research Center of Vanderbilt University for stock orderflow data, and the Compustat Daily database for information on stock trading volumes.[20] For my analysis of the probability of informed trading, I considered a time range from 1 January 2011 to 30 June 2011. I selected all bonds available for the constituents of the S&P 500 during this period. Two different datasets were constructed: For the cross-sectional analysis, from issuers listed in the S&P 500, I obtained all available bonds from the enhanced TRACE database with more than 29 trades and complete information in the Mergent FISD database and the Compustat Security Daily database during the period between 1 January 2011 and 30 June 2011. This left 4,155 bonds for my analysis. For the comparison between bond and stock markets I used a different dataset. I followed Easley et al. (1996) and divided bonds into ten deciles by their transaction volume observed during the first half of 2011. I then selected the three deciles, the 3rd, the 5th and 10th (the 10th being the decile with highest transaction volume). For each company, the bond with highest transaction volume and at least 30 trades in the period considered was chosen. This procedure yielded a total of 9,728 observations for 108 bonds over the three deciles. Finally, I assigned to each of these 108 bonds its corresponding equity, i.e., the deciles for equities are determined by the bonds.

As Dick-Nielsen (2013) notes, the enhanced TRACE data set is a tremendous improvement for most applications using TRACE and enjoys several advantages over the standard TRACE data. While the standard TRACE database caps large transaction volumes in reports to '1MM+' and '5MM+', the enhanced TRACE set reports these transaction volumes uncapped. Furthermore, enhanced TRACE includes buy-sell side information and in addition to trade execution dates and times the corresponding report filing dates and times. The enhanced TRACE database contains transaction reports for all transactions dating back to the introduction of TRACE in July 2002, although this comes at the cost of an 18-month lag in availability.

Because TRACE includes raw reports the data must be cleaned to eliminate duplicate entries due to bilateral reporting of counterparties in trades involving two registered members of the FINRA. Furthermore, trades are often cancelled or corrected ex post on the same day or on later days. Cancelled and corrected reports and duplicates from agency and interdealer trades would distort arrival rates in the EKOP or compressed model, respectively, and could have strong effects on the measured PIN. In cleaning the data set I followed suggestions from Dick-

[20] All bond data were retrieved from Wharton Research Data Services (WRDS). For stock order flow data see the website of the Financial Markets Research Center of Vanderbilt University, `http://www.vanderbiltfmrc.org/databases/market-microstructure-database/`.

Nielsen (2009, 2013) and Asquith et al. (2013).

A detailed description of my cleaning algorithm is provided in appendix A, and a summary of the cleaning procedure is presented in table B.1: Of the 507 companies included in the S&P 500 during the first half-year of 2011, 387 issued publicly traded debt eligible for TRACE.[21] For these 387 companies the enhanced TRACE sample included 10,818 bonds prior to cleaning with a total of 4,483,549 reports filed. Of this amount, 31,842 reports indicated same-day corrections resulting in a deletion of the related original report filings and 41,329 trades were cancelled, which resulted in a deletion of both, the original and cancelling reports. Cancellations or corrections on later days defined 43,900 more reports as deleted from the database and unmatched reversals from this step were matched with 'as-of' reports and resulted in 3,085 further deletions.[22] Additionally, remaining 10,205 reversals that could not be matched to prior trade reports were excluded.

Agency trades are trades in which a so-called 'introducing dealer' acts as an agent for its client and transfers its client's order to another dealer that in turn executes the order. This chain results in three reports filed in TRACE, although only one trade was executed. I identified agency trades by three matching trade reports and deleted two of them. In total, 626,632 reports were deleted due to agency trades. In interdealer trades both FINRA members involved in the trade have a report obligation. I matched these reports based on a POSIX timestamp (date and time) and in a second step by date alone and retained only the sell side of such trades. In total, 656,026 filings could be identified as part of an interdealer trade by matching date-times, and further 901,328 reports were additionally found by matching only dates. Remaining unmatched interdealer buy-side reports were also deleted. Using this approach, I followed the conservative method of Asquith et al. (2013). I checked for special trade or settlement agreements and for delayed reporting but no such reports were contained in my enhanced TRACE data sample. In sum, 2,594,567 reports were deleted and 1,888,982 reports remained in my database.

Bond order flow data from the enhanced TRACE database and information on issuance and ratings from Mergent FISD were matched based on the securities' complete CUSIPs. The stock and bond data were matched based on the company ticker.

Table B.2 reports summaries for daily contract volumes in the bond markets and for daily share volumes in stock markets during the considered time interval. The first block summarizes the trading volume statistics of intraday bond data in enhanced TRACE. The volumes were collected from individual trades and aggregated over each trading day for a bond. I report means and medians from the deciles three, five and ten. Contract volumes differ substantially across deciles ranging in mean from approximately 200,000 contracts in the lowest decile to 4.3 million contracts in the highest. With a maximum of over 1.7 million contracts per day and a minimum of nearly 209 contracts, daily volumes in the 10th decile vary significantly. Similar variations can be observed for deciles 3 and 5. Mean standard deviations resemble these high

[21] Constituents of the S&P 500 index are selected using guidelines and an index committee that meets once a month. Over the period from January 2011 to June 2011 some companies in the index were replaced by others resulting in 507 companies over the full period.

[22] 'As-of' reports are filed on subsequent days and replace the original reports.

variations. The median values on the right side of the table indicate that the daily contract volumes in all deciles are driven by some large volumes.

For the stock markets I collected the daily share volumes from Compustat Security Daily to calculate the statistics.[23] The summary of the stock market volumes reveals similar distributions for all share volumes in the considered period. Beginning with a mean share volume of over 449 million for the first half-year of 2011 in the lowest decile, the highest decile has an average of 2.9 billion; these are between 500 to 2000 times the mean contract volumes in the bond markets for this period. The daily share volume in stocks varies between 97 million and 10 million in decile 10 and approximately 14 million and 1.5 million in decile 3. Note, that the minimum daily share volume in the highest stock decile is higher than the maximum contract volume in the highest bond decile. All median volumes are smaller than the mean volumes for the stock deciles and this points to a right-skewed distribution. Interestingly, the bond market behaves accordingly.

Summaries of numbers of trades are contained in table B.3 and reveal a greater difference between fixed-income and equity markets. While on the equity markets trades occurred on approximately 35% more days than on bond markets, the number of trades per day on these markets is also substantially higher: The average in fixed-income markets in the highest decile is approximately 14 trades per day, whereas even in its lowest decile, the stock market records over 15,000 trades daily. Maximum numbers of trades per day were on average 163 in the highest decile of the bond markets and over 54,000 in the lowest decile in the equity sector with values increasing along deciles. The minimum number of daily trades in fixed-income markets is only between 1 and 3 trades on the average bond. Comparing these numbers to the stock markets I arrive at values that are more than 7,000 times higher in magnitude. Overall, the equity markets report between 1.9 and 7.2 million trades between 1 January 2011 and 30 June 2011 whereas the bond markets range in total trades from between 320 in their lowest and 1,314 in their highest decile. This clearly demonstrates that even if the structure of the bond market enables very large transaction volumes, trades are rare compared to the equity markets. There are several hypotheses for these large differences in trade numbers between the markets for corporate bonds and the associated markets for equity shares. One is the substantially lower trading activity of the retail segment on bond markets. A further explanation would be higher transaction costs on fixed-income markets, as demonstrated by Edwards et al. (2006), Bessembinder et al. (2006) and Goldstein et al. (2007). Yet another potential reason is the decentralized market structure in which trading mechanisms continue to significantly rely on communication via phone, email and even face-to-face encounters and the preference for buy-and-hold strategies in fixed-income markets. Furthermore, in equity markets, the algorithmic trading segment enjoyed substantial growth rates over the last seven years, whereas in fixed-income markets this segment is negligible.[24]

[23] The Compustat Security Daily data were collected via Wharton Research Data Services.
[24] See Kishore (2013) for a brief overview of algorithmic trading in fixed-income securities.

4. Estimation and results

This section reports the methodology and results of the two analyses conducted in this paper, namely a cross-sectional regression to explain the causes of information asymmetries in bond markets, and a comparison of PINs between bonds and their corresponding equities. I begin with the cross-sectional analysis.

4.1. Cross-sectional regressions

To explain the causes of information asymmetries in U.S. corporate bond markets I conducted a cross-sectional regression using 4,155 bonds traded in fixed-income markets in the period between 1 January 2011 and 30 June 2011. The probability of informed trading for each bond was estimated using a procedure described in the following paragraph.

Easley et al. (1996) propose a model (henceforth denoted as the EKOP model) to estimate the information asymmetry in a security market using the probability of informed trades in the order flow. Their model considers a market with competitive market makers, informed traders and uninformed (liquidity or noise) traders. Buyer- and seller-initiated trades are counted on each trading day and a mixed Poisson process is fitted to the data. Uninformed buyers and uninformed sellers each arrive at a rate ϵ. Information events occur with probability α and are either good (with probability $(1-\delta)$) or bad news (with probability δ). On good- and bad-news days informed traders are presumed to trade on the demand and supply side of the market, respectively, thereby intensifying the arrival rates on these days by μ. Easley et al. (1996) assume a centralized market in their model, however, even if the market for corporate bonds is decentralized with highly fragmented liquidity pools, TRACE represents a central collection of the market's trading activity.

The enhanced TRACE database contains buy- and sell-flags and enables estimation of the EKOP model, although I used a slightly modified version of this model first described by Kokot (2004) and recently reanalyzed by Grammig et al. (2014) (denoted as the 'compressed' model).[25] The compressed model forgoes the need for buy-sell flags and solely considers the number of trades in a market. It does so by collapsing good- and bad-news days and therefore requires one fewer parameter than the EKOP model, i.e., δ is no longer needed:

$$L(Q|\vartheta) = \alpha((2\epsilon+\mu)T)^Q \frac{\exp\left(-(2\epsilon+\mu)T\right)}{Q!} + (1-\alpha)(2\epsilon T)^Q \frac{\exp\left(-2\epsilon T\right)}{Q!}\,, \tag{4.1}$$

where $\vartheta = (\alpha, \epsilon, \mu)'$, denotes the parameter vector, containing α, the probability of news, ϵ,

[25] Moreover, Wuensche (2007) uses a Poisson mixture and presents stylized facts concerning the empirical count data for trades.

the intensity parameter of uninformed (noise or liquidity) trading, and μ, the arrival rate of informed trading. Q denotes the number of trades per day and T is the length of the trading day, set to 390 minutes (6.5 hours) in my analysis.

The measure for information asymmetry (the probability of informed trading) is then defined by the three parameters contained in ϑ:

$$PIN = \frac{\alpha\mu}{\alpha\mu + 2\epsilon} \,. \tag{4.2}$$

The reason for my choice of the compressed model was, first, the very low trading activities on the sell and/or the buy side of the fixed-income markets (which are occasionally even zero) and, second, the prevailing instabilities of the maximum likelihood function in estimating the EKOP model. These instabilities are based on the nature of the EKOP likelihood function as a finite mixture of three bivariate Poisson distributions.[26]

The prevailing instabilities of the maximum likelihood procedure led me to employ the Markov Chain Monte Carlo (MCMC) approach proposed by Grammig et al. (2014). I ran the MCMC sampler with a burn-in of 1,000 iterations and collected parameter samples from the following 10,000 replications for all 4,155 bonds in my dataset.[27] The resulting probabilities of informed trading in my bond sample vary between a minimum of 0.31×10^{-5} and a maximum of 0.975 with an overall mean of 0.11 and a median of 6.25×10^{-2}. The 1st and 3rd quartiles are 4.17×10^{-3} and 0.179, respectively. In summary, the distribution of PINs in my data set is highly skewed to the right, with many bonds exhibiting small probabilities of informed trading and a few bonds having very high probabilities.

The estimated PINs were then regressed on several variables describing either the bond and its market or the corresponding company's stock. Table B.4 summarizes the variables used in the multiple regression. Maturity and age are considered important variables for the evaluation of a bond and the bonds in my sample are on average 5 years old and have a mean of 10 years until maturity. The coupons are distributed around \$5, with 404 bonds possessing a variable coupon and 137 being zero-coupon bonds (all remaining bonds have a fixed coupon). Coupons are paid with an approximate frequency of twice annually. The offering amount is on average \$400 million with a standard deviation of more than \$600 million. The mean denomination

[26] The contour of this likelihood function contains six modes and in the case of overlapping components these modes form a plateau that impedes the optimization algorithm to determine the maximum. For instance, in estimations of the EKOP model using a Nelder-Mead simplex on stock order flow, 103 of 108 estimations did not depart from the initial values and all hessian matrices were singular. I employed two different likelihood functions: The original likelihood function reported in Easley et al. (1996) and a modified version accounting for large trade amounts proposed in Easley et al. (2008). A maximum likelihood approach with a compressed likelihood function and applied to bond trades yielded better estimation results, although in 24 of 108 estimations, the hessian matrices were singular and had to be inverted via the Moore-Penrose approach and 10 covariance matrices were not positive definite. Results from these maximum likelihood estimations are available upon request.

[27] All estimations were performed by using a self-coded `R`-package based on `C++`-extensions that rely heavily on the high-performance library `RcppArmadillo (0.3.920.1)` (see Eddelbuettel and Sanderson (2013)).

(minimum trade size) of the bonds in my sample is approximately four bonds per trade but with a standard deviation of more than 18 contracts. The high standard deviation results from several bonds with denominations of between 100 and 250 bonds per trade.

Only a small number of bonds are convertible (65) and fewer bonds are exchangeable.[28] Five companies were in bankruptcy during my sample period and for 443 securities their issuer had made a debt tender offer since issuance. During the period from 1 January 2011 to 30 June 2011 203 companies were (re)rated and over 900 companies were not (any longer) monitored by a ratings agency (off-watch). Ratings are predominantly between A and BBB, and only 84 bonds were top-rated (AAA). Three rating agencies analyzed the creditworthiness of issuers, namely Standard & Poor's, Fitch, and Moody's. Of 4,155 bonds, Fitch rated the most, 2,524, followed by Moody's, with 1,087, and Standard & Poors, with 544.

Bond markets possess high volumes, as mentioned in prior sections: The average bond in the cross-sectional sample had a volume of over $300 million traded in the first half-year of 2011. The variables 'MM1' and 'MM5' determine the number of large trades with trading volumes between $1 and $5 million and above $5 million, respectively. The distributions of these variables are highly skewed, with a few bonds showing very large trades quite frequently, but a median at 13 and 4 for 'MM1' and 'MM5', respectively. Traditional market microstructure measures were added to my analysis, among them the Amihud (2002) (il)liquidity measure, the Roll (1984) measure of transaction costs, and dispersion as a second measure of transaction costs. The Amihud liquidity measure was measured as in Dick-Nielsen (2009):

$$\text{Amihud}_t = \frac{1}{N} \sum^{N} \frac{|r_{ij,t}|}{V_{i,t}} \,, \tag{4.3}$$

where N is the number of trades of a bond on day t, r_{ij} is the return between two consecutive trades j and i and V_i is the dollar par volume for trade i (measured in $million). I then took the median over all daily Amihud measures of a bond.

The Roll measure was computed using daily price differences:

$$\text{Roll}_t = 2\sqrt{-\text{Cov}(\Delta P_{t,i}, \Delta P_{t,i-1})} \,, \tag{4.4}$$

where t is the trading day and i denotes intraday transaction times. $\Delta P_{t,i}$ and $\Delta P_{t,i-1}$ are the intraday price difference and its lag, respectively. The single Roll measure for a bond in my sample is the median over all trading days with negative autocovariance in price differences.

Price dispersion was proposed by Feldhütter (2012) and Dick-Nielsen et al. (2012) to capture

[28] A convertible bond includes the option to convert the bond into the equity of its issuer. An exchangeable bond also carries a convertibility option but in contrast to the convertible bond a bondholder has the right to exchange her bond to stock in a subsidiary or a company in which the issuer owns a stake.

roundtrip trading costs:[29]

$$\text{Dispersion}_t = \frac{1}{N}\sum^{N}\frac{P_{t,max} - P_{t,min}}{P_{t,last}}\,, \tag{4.5}$$

where N is the number of days, $P_{t,max}$ and $P_{t,min}$ are the maximum and minimum price on trading day t, and $P_{t,last}$ denotes the closing price on day t.

The Amihud (2002) measures are very small indicating sufficient liquidity to avoid large price jumps between consecutive trades. Further, the dispersion measure proposed by Feldhütter (2012) and Dick-Nielsen et al. (2012) is quite small.

Finally, I included the median of daily standard deviations of bond prices, the mean issuer's stock return (measured in %) for the period between 1 January 2011 and 30 June 2011, the issuer's stock return volatility for the same period (measured by the standard deviation of daily percentage stock returns), and an indicator variable for affiliation with the highly regulated financial or utility industries in the analysis. Prices on bond markets do not exhibit high daily volatility during the first half-year of 2011. The average bond issuer has daily average returns of approximately 0.065%, with a median slightly lower at 0.052%. Stock return volatilities were in an acceptable range, approximately 1.8% for the average bond issuer. Interestingly, over half of the bonds were issued by a company in the financial or utility industries. These industries are highly regulated and one could expect, that the bonds of such firms behave differently.

For my examination of information asymmetries estimated from aggregated trading activity I specified my model as follows:[30]

$$\begin{aligned} PIN_{ij} =& \beta_0 + \beta_1 Maturity_i + \beta_2 Age_i + \beta_3 Coupon_i + \beta_4 VariableCoupon_i \\ &+ \beta_5 \log OfferingAmount_i + \beta_6 Convertible_i + \beta_7 InBankruptcy_i \\ &+ \beta_8 TenderOffer_i + \beta_9 RatedInRange_i + \beta_{10} AAAdum_i \\ &+ \beta_{11} AAdum_i + \beta_{12} Adum_i + \beta_{13} S\&P_i + \beta_{14} Moody's_i \\ &+ \beta_{15} \log Volume_i + \beta_{16} MM1_i + \beta_{17} MM5_i \\ &+ \beta_{18} Dispersion_i + \beta_{19} PriceStd_i + \beta_{20} IssuerStockReturn_{ij} \\ &+ \beta_{21} IssuerStockVolatilty_{ij}\,, \end{aligned} \tag{4.6}$$

[29] Feldhütter (2012) and Dick-Nielsen et al. (2012) demonstrate, that their measure is as robust as the Amihud (2002) liquidity measure.

[30] Beginning with an even larger set of variables, I used the Akaike information criterion (AIC, see Akaike (1974)) to exclude variables from the estimation. Specifically, the `step`-function in `R-3.0.2` was used to exclude and include variables in a stepwise manner. A list of excluded variables is available upon request.

where i indicates a certain bond i in the sample and j the corresponding stock j, $Maturity_i$ is the number of days until bond i matures, Age_i is defined as the number of days since bond i was issued, *Coupon* is the $-valued coupon of the bond, $VariableCoupon_i$ is a dummy, indicating whether bond i has a variable coupon, $OfferingAmt_i$ is the amount of debt issued, $Convertible_i$ indicates whether a bond is convertible to common stock, $InBankruptcy_i$ is equal to one, if bankruptcy proceedings were initiated for an issuer, $TenderOffer_i$ indicates whether a debt tender offer had been made for bond i by its issuer since issuance, $RatedInRange_i$ is a dummy reporting that bond i was (re)rated during the considered time period, $AAAdum_i$ to $Adum_i$ are dummies denoting the rating of bond i (always the most recent rating has been chosen), $S\&P_i$ and $Moody's_i$ are dummies indicating whether a bond has been rated by the two rating agencies. $MM1_i$ and $MM5_i$ count the number of large trades (between 1 and 5 million and more than 5 million contracts per trade) in bond i, $Dispersion_i$ is the price dispersion measure from eq. (4.5), $PriceStd_i$ is the median daily standard deviation of security's i prices, $IssuerStockReturn_{ij}$ is the mean of daily stock returns of bond i's issuer j in the first half-year of 2011, and $IssuerStockVolatility_{ij}$ is the corresponding issuer's volatility of daily stock returns from January to June 2011.

Equation (4.6) was estimated by least squares using heteroskedasticity-consistent standard errors and the results are reported in the first block of table B.5.[31] The variables considered explain over 47% of the variation in the probability of informed trading with all variables being significant at the 10% level. Other than *Age* all variables are also significant at the 5% level and most of them as well at the 1% level. The Wald test supports the choice of the model and rejects the null at a confidence level above 1%. A (re)rating (*RatedInRange*) substantially increases information asymmetry. This is unsurprising, as a (re)rating typically indicates a case of new information arrival. Conversely, a company in bankruptcy exhibits, on average, a lower probability of informed trading as reflected in the negative coefficient of *InBankruptcy*.

The large coefficient of the price dispersion is due to the very low values of this variable. However, large price movements increase information asymmetries in the corporate bond market. In the EKOP model large differences in maximum and minimum transaction prices in a market can only be caused by informed trading, as otherwise the market maker does not alter her quotes. The regression results indicate that price dispersion increases the PIN and thus conform to the model.

Higher coupons positively affect the probability of informed trading. Whether this is because higher coupons could be considered a proxy for lower ratings remains ambiguous. However, a variable coupon also increases the PIN: Such a coupon is coupled with the interest rate and introduces a further source of volatility, which increases information asymmetry in the market.

If an issuer has ever made a tender offer to repurchase its debt, investors regard a second offer as more likely, creating uncertainty concerning the value, thereby increasing information asymmetry in the bond. A rating between AAA and A decreases the PIN: All coefficients for the included rating dummies take a negative sign. This is unsurprising as a high rating makes

[31] In detail, I use the heteroscedasticity-consistent variance-covariance matrix proposed by Cribari-Neto and da Silva (2011). This matrix is implemented in the `sandwich (2.3.0)`-package for `R` (type `HC4m`).

unexpected losses less likely. Further, investment-grade bonds might enjoy higher trading activity in the retail sector, and as Easley et al. (1996) note, uninformed trading reduces the PIN. In the comparison analysis below, I demonstrate that probabilities of informed trading in different deciles of corporate bond markets are primarily driven by the relative rates of uninformed traders.

Interestingly, a rating by S&P or Moody's appears to increase information asymmetries, in contrast to the results for Fitch (which rated most of the bonds in the sample).

Whereas the trading volume of a bond increases information asymmetry, the offering amount at issuance has a negative impact. The former result conforms to the findings from the comparison analysis below, as bonds in higher volume deciles exhibit also higher PINs.

It is intriguing that having a number of trades with between \$1 million and \$5 million in trading volume increases the probability of informed trading, whereas having a high number of trades with more than \$5 million in trading volume decreases this measure to a substantially greater extend. The model of Easley and O'Hara (1987) predicts that informed traders will, on average, trade higher volumes to extract the largest rents possible, although the model does not differentiate between certain large trade volumes. It is possible that trades with a dollar volume in excess of \$5 million are considered liquidity trades by institutional investors such as pension funds and do not rank among informed trades. Menkhoff and Schmeling (2010b) find that informed traders predominantly use medium-sized orders following a trading strategy called stealth trading.

Edwards et al. (2006) regress estimated percentage transactions costs on various bond characteristics. Some of their bond characteristics are identical to the variables used in my cross-sectional regression. However, under the assumption of the commonly accepted adverse selection component in transaction costs, not all coefficient estimates coincide. Ratings of BBB to C increase transaction costs, and this agrees with the finding that ratings of A to AAA decrease information asymmetries which in turn could be assumed to decrease transaction costs. The estimated coefficient of the coupon rate in the analysis of Edwards et al. (2006) is positive as is the estimated coefficient in my cross-sectional analysis. The issue size (in my equation *OfferingAmount*) is positively and a variable coupon negatively related to transaction costs. Both variables affect the probability of informed trading similarly in that a variable coupon increases information asymmetries in a market and the offering amount decreases it accordingly. However, there are estimated coefficients that do not coincide with the findings in Edwards et al. (2006): *InBankruptcy*, *Age*, and *Maturity*. Indirect effects could account for these discrepancies in estimated coefficients.

There could be numerous covariance structures in the residuals that vary by the issuer, although the individual residuals are homoskedastic within each cluster. To allow for greater flexibility in the variance-covariance matrix of the regression I applied cluster-robust standard errors and the results are reported in the second block of table B.5.[32] Accounting for clustering changes the significance levels of the *Age* variable from below 10% to above 15%, of the

[32]For a discussion of robust inference under within-group correlated errors, see for example Cameron et al. (2011) and the references therein.

dummy *AAAdum* from 1.4% to 0.06%, of the dummy *AAdum* from 2.03% to 6.34%, and of the dummy *InBankruptcy* from above 1% to approximately zero, however, most other variables remain highly significant and the Wald-statistic increases from 112.8 to 9,300.

4.2. Comparison of PINs between corporate bond and corresponding equity markets

As described in the data section above I divided bonds with more than 30 trades in the first half-year of 2011 into deciles based on their transaction volume and selected the deciles 3, 5, and 10 for my comparison analysis. For each bond the corresponding stock was then assigned to the same decile. This procedure led to 39 bonds with an average of 87 observations and 39 stocks with an average of 122 observations in the 3rd decile, 35 bonds with an average of 92 observations and 35 corresponding stocks with an average of 121 observations in the 5th decile, and 34 bonds with an average of 90 observations and 34 stocks with an average of 122 observations in the 10th decile.

For all bonds and stocks, I estimated parameters using the random permutation Gibbs sampler with a burn-in of 1,000 iterations and collecting the following 10,000 replications. Parameter and PIN estimates from the comparison analysis (Bonds and Stocks) are reported in table B.6. Estimates in each decile are aggregated by the mean and median. The results reflect the findings from the number of trades in table B.3: Arrival rates of Poisson processes on the stock markets are much higher than corresponding rates on the fixed-income markets. The log-likelihoods accord with these findings. While the stock market results reflect the expected pattern observed in earlier studies, e.g., by Easley et al. (1996), with greater information asymmetries on less active markets, the structure of PIN estimates for the bond markets is the reverse of prior findings. Estimates for the stock market vary between 16.1% in the 3rd and 18.9% in the 5th decile. Easley et al. (1996) obtained similar results with 22% in decile 3, 20.7% in decile 5 and 16.3% in decile 10. Corresponding PIN estimates for bonds in my study vary between 9.9% in low-activity markets and 31.2% in markets with high trading activity. This structure is also observed with reduced values, for the median in the third block of the table, indicating that mean estimations are dominated by a few large values. Standard errors for estimates in the fixed-income sector are much higher than their stock counterparts. Relative standard errors in bond markets are between 105% in the 3rd and 61% in the 10th decile, whereas average variation in stock markets' PINs is between 20% in decile 3 and approximately 28% in deciles 5 and 10.

The anomalous structure of information asymmetries in the bond markets cannot be easily explained by prevailing theories. Typically, more severe information asymmetries on less active markets are assumed to result from there being fewer uninformed traders in these markets (see e.g., Easley et al. (1996)). Comparing the differences in informed trading in bond and stock markets between deciles indicates that informed trading, measured by μ, is distributed similarly across deciles in the bond and stock markets. This also holds for the rates of uninformed trading, measured by ϵ. However, fixed-income and equity markets, differ significantly in the ratios between the intensity rates of informed and uninformed trading. Informed trading rates

are between 4.2 (decile 3) and 20.4 (decile 10) times larger than the arrival rates of uninformed trading. The corresponding figures for the stock markets range between 1.8 (decile 3) and 1.9 (decile 10). This indicates a much less uninformed trading activity *relative* to informed trading intensities in fixed-income segments. The growing portion of retail trading in the fixed-income segment, presented in FINRA (2005) and Edwards (2006), does not appear to generate sufficient uninformed trading, especially for liquid bond markets in which retail traders can be assumed to be the most active.[33] In contrast, this could create an incentive for informed traders to enter markets in which retail trading is more extensive, enabling them to extract greater profits at the expense of the uninformed. Furthermore, it can be assumed that in less active bond markets more trade agreements are concluded on a personal basis. This environment could make informed trading costly relative to markets with more anonymous trading. In addition, a substantial share of the information in corporate bond markets could be substitutional, e.g., information on future interest rates, facilitating shifting informed trading between different bond markets. In turn, informed traders favor bond markets with frequent trading for their business, and as a result, the intensity of informed trading in markets with high trading volumes will increase disproportionally. Marsh and O'Rourke (2005) report such substitution effects in information for FX markets.

Goldstein et al. (2007) and recently, Han and Zhou (2011) report that trading costs are higher in corporate bond markets with high transaction volumes than on markets with low transaction volumes. Differences in transaction costs between low-volume markets and high-volume markets are assumed to arise because of inventory holding costs, order processing costs, and adverse selection costs.[34] Holding inventory in markets with infrequent trading is risky, as rebalancing could become difficult due to the limited number of counterparties. Particularly in decentralized markets the order processing costs could determine a much larger share of trading costs, as concluding trade agreements in this context requires more time and resources. However, the findings in this paper would suggest, that trading costs are primarily driven by the adverse selection component as both transaction costs and information asymmetries appear to increase in transaction volume. My estimates agree with the regression results in the analysis of estimated spreads in Goldstein et al. (2007, Section 3.3).

I followed Easley et al. (1996) and conducted Kruskal-Wallis and Mann-Whitney tests for the parameters of the compressed model and the PIN.[35] The Kruskal-Wallis test allows me to compare the three samples of PINs from deciles 3 to 10 and determine whether at least one sample does not stem from the same distribution as the two others. The Mann-Whitney test is similar but allows for bilateral comparison.

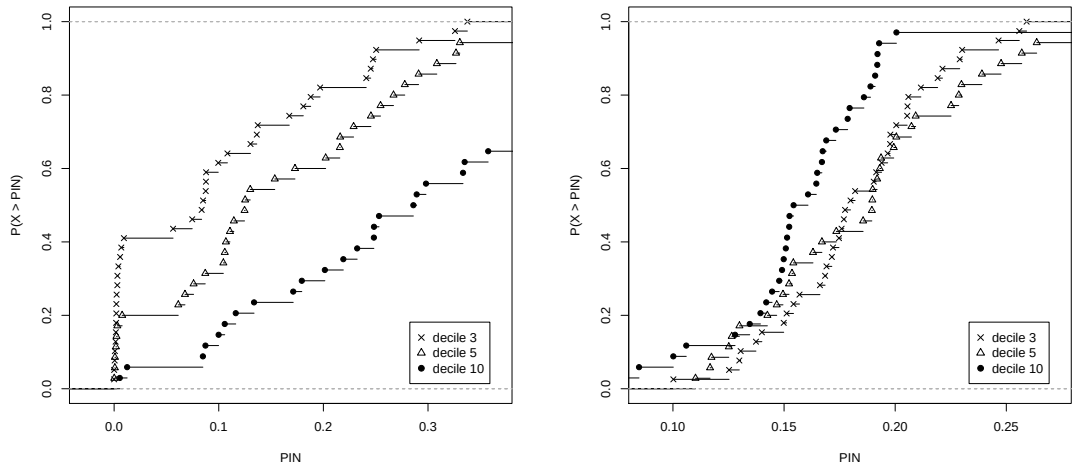
The plots of the empirical distribution functions of PINs in fig. 4.1 depict similar results for the equity markets (see fig. 4.1(b)) to those in Easley et al. (1996) with intersecting distributions

[33] Retail trading is generally assumed to be uninformative. See for instance Ferriani (2010) or Malinova et al. (2013).

[34] See for example the empirical studies of Madhavan et al. (1997) and Huang and Stoll (1997).

[35] All tests were performed in `R-3.0.2` using the `kruskal.test`-function for the Kruskal-Wallis tests and the `wilcox.test`-function with exact p-values for the Mann-Whitney tests.

for the lower deciles and clear dominance of the highest decile. In contrast, the probabilities of informed trading in bond markets (fig. 4.1(a)) indicate a reverse ordering with significant distances between the functions.



(a) (b)

Figure 4.1.: **Empirical distributions of PINs.** The figures depict empirical distribution functions for the PIN estimates in each decile in corporate bond markets (fig. 4.1(a)) and equity markets (fig. 4.1(b)).

The Kruskal-Wallis test statistics and their corresponding p-values in table B.7 all strongly reject the null-hypothesis of identical distributions, except the tests for the news probability α on equity markets.

Mann-Whitney tests were applied by comparing the 3rd and 5th, 3rd and 10th, and 5th and 10th deciles for all markets. The Mann-Whitney test results are presented in table B.8 and largely support the Kruskal-Wallis test results above. However, the difference between the parameters and PINs in the fixed-income markets is significant at the 1% level between the highest and lower deciles whereas both, stock and bond markets show no significant difference between the two lower deciles. The Mann-Whitney tests for equity markets reflect the results from Easley et al. (1996): The null hypothesis is strongly rejected in a comparison of the 3rd and 5th deciles with the 10th decile, however, for the news probability, α, no significant difference between deciles was observed. Overall, the tests reinforce the anomalous structure of information asymmetries in fixed-income markets relative to equity markets.

As both stocks and bonds are issued by the same company it could be considered contradictory if the probabilities of news, α, for the two securities differ.[36] For this reason I also conducted

[36] A more thorough discussion of this issue and a corresponding modeling approach can be found in Grammig et al. (2001), however, the authors investigate floor and screen trading of the same security.

Kruskal-Wallis and Mann-Whitney tests to test for the equality of the underlying population of estimated α-parameters between stock and corporate bond markets in each decile. Table B.9 presents the results. For decile 10, the null hypothesis of equal distributions is strongly rejected, but for decile 5, the hypothesis is rejected at the 5% level and for decile 3, this hypothesis cannot be rejected. The empirical distribution plots in fig. 4.2 indicate that for each decile the distribution function of stocks intersects the corresponding distribution function of bonds. This also explains the divergent results of the Kruskal-Wallis and the Mann-Whitney tests in deciles 3 and 5: Both tests are based on ranks and as the intersection in deciles 3 and 5 is nearly symmetrical, no significant difference can be detected between corresponding ranks.

Regarding the discrepancies between the news probabilities, α, of a bond and stock from a single company, I also investigated the news probabilities among different bonds issued by the same company and found similar discrepancies in my cross-sectional dataset used in section 4.1. The results are summarized in fig. 4.3 and confirm the findings in the last paragraph. The news probabilities of bonds issued by the same company often vary within a range of 60%, and this behavior appears inconsistent.

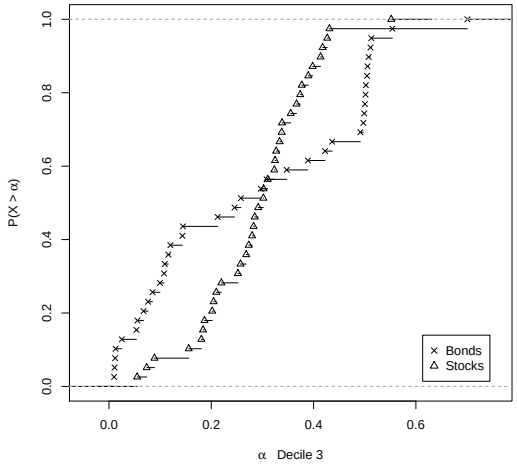
There are various explanations for why news probabilities in stock and corresponding corporate bond markets are not identical. First, it might be that the relevant news in the two markets is not identical, i.e., at least part of exploitable news in one market is independent of exploitable news in the other. However, the wide ranges of news probabilities from bonds issued by the same company would contradict this hypothesis. Second, the EKOP model and its compressed version could simply be inappropriate for applications involving highly fragmented markets. The last hypothesis is also supported by the large discrepancies between the arrival rates of informed and uninformed trading on the bond markets. The intensities of informed trading, μ, are between four and twenty times higher than the intensities of liquidity trading, ϵ. Intuitively, such a market should not function.

5. Conclusion

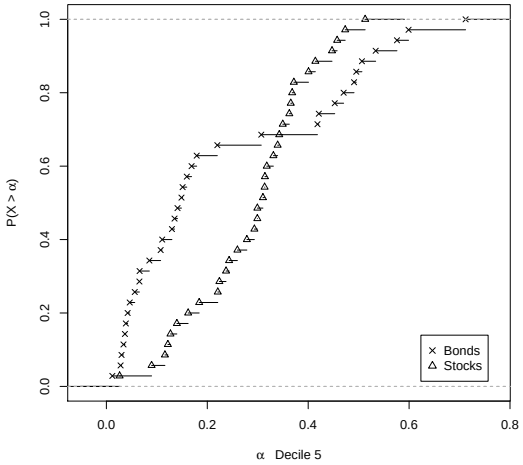
In this study I analyze the probability of informed trading on U.S. corporate bond markets by first investigating the causes of information asymmetries through cross-sectional regressions and second, by comparing information asymmetries in debt and corresponding equity markets.

In summary, the results from the cross-sectional analysis reveal that several variables appear to have a significant influence on information asymmetries in corporate bond markets. Several significant factors are related to bond characteristics and could become relevant to debt issuers as Han and Zhou (2013) demonstrated that information asymmetry is priced and investors demand higher debt yields in compensation for the adverse-selection risks they carry.

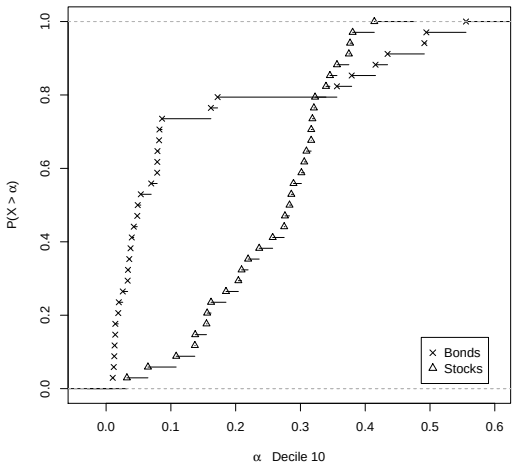
The findings from the comparison study suggest that the relationship between liquidity and information asymmetry in the two markets is different. In equity markets information asym-



(a)



(b)



(c)

Figure 4.2.: **Empirical distribution functions of news probabilities.** The figures compare empirical distribution functions for the α-estimates in corporate bond markets and equity markets for decile 3 (fig. 4.2(a)), decile 5 (fig. 4.2(b)), and decile 10 (fig. 4.2(c)).

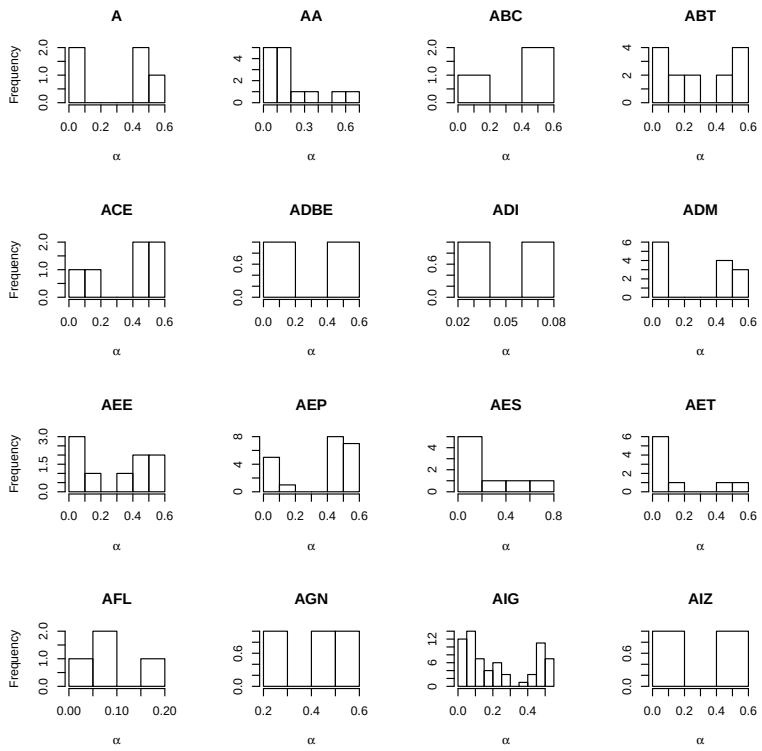


Figure 4.3.: **News probability distributions.** The figure plots histograms of the news probabilities of bonds issued by a single company. The companies are the first 16 companies in the cross sectional data set.

metry is greater for less active stocks than for stocks with high trading activity whereas infrequently traded bonds seem to suffer less under information asymmetry than more active bonds. Possible explanations for this behavior may be related to the specific market structure of corporate bond markets in the U.S., namely an OTC market complemented by interdealer brokerage.

My findings also suggest that the low relative rate of uninformed traders represents a crucial driver of the information asymmetries in bond markets. The growing segment of retail trading in the corporate bond market reported by the SEC and FINRA may adjust this imbalance over time demonstrating its importance to an efficiently working market.

However, I also observe significant differences in the news probabilities between stocks and their bonds which could be regarded as contradictory. Thus, one must question the appropriateness of traditional models for measuring information asymmetry. The EKOP (and comparably the compressed model) might not be suited for highly fragmented markets such as the U.S. corporate bond market investigated here.

There remain open questions that could not be covered in my study and are left for future research. First, if the results in this study can be considered reliable, then they demand an explanation of the reversed structure of information asymmetries in relation to liquidity. Traditional theoretical models concerning information asymmetry cannot explain this relationship. Higher transaction costs for frequently traded debt, as observed by Goldstein et al. (2007) are consistent with simultaneously higher information asymmetries, but there is currently no theory clarifying the causes. The literature holds that transaction costs are induced by information asymmetries and a model is necessary to describe this relationship for fixed-income securities.

Second, investors' pricing of information asymmetries was first investigated by Han and Zhou (2013) and cross-sectional regressions reveal additional variables that appear to affect the information asymmetry. It remains unclear whether there are certain debt-issuer characteristics that could become crucial to reducing the cost of debt.

Finally, regulators would benefit from deeper insights into the causes of information asymmetries in debt markets. Recently, the SEC announced the agenda for the roundtable of fixed-income markets, which includes numerous questions to be answered by research and investigation.[37] Some of these questions concretely address differences in information asymmetries caused by bond characteristics making this study a promising initial contribution to a fruitful literature on the topic of information asymmetries in the U.S. corporate bond markets.

[37] See `http://www.sec.gov/spotlight/fixed-income-markets/fixed-income-markets-agenda.htm` for the agenda of the roundtable for fixed income markets published by the SEC in April 2013.

A. Cleaning the enhanced TRACE data set

I cleaned my enhanced TRACE data set by eliminating same-day corrections and cancellations, trade reversals, duplicates from agency trades and interdealer trades. As Dick-Nielsen (2009) emphasizes, the TRACE system is essentially a one-day system, and reporting is only possible within the system's operating hours. This enables easy trade corrections or cancellations within the same day as the report was filed, but necessitates a specific process to reverse trade reports from past days. In the event that a member wishes to correct same-day reports she must send a report identified as a correction and referring to the original report via the original message sequence number. The same is true for trade cancellations. In the case of same-day corrections, the original report must be deleted, and in the case of a trade cancellation, the original report as well as the cancellation report must be removed from the data set. I use the algorithm developed by Dick-Nielsen (2013) to identify and delete redundant reports from same-day corrections and cancellations.

Corrections and cancellations filed after the day the original report was created are not captured by using the original message sequence number. Message sequence numbers are only unique for a given security on a single day; therefore members who wish to cancel trades on a later date must file a report identical to the original trade report indicated as a 'reversal'. If the member wishes to correct a trade from a previous day, she must send a reversal report and in addition a socalled 'as-of' report with identical information but corrected price and/or volume. Note that reversals could also cancel an as-of report. To identify reversals and as-of reports I combined the approaches developed by Dick-Nielsen (2013) and Asquith et al. (2013): to match reversals and as-of reports with their corresponding original filings, in addition to the bond symbol I used a POSIX timestamp (combining trade execution date and time), the reported price, the entire volume quoted, and the buy-sell flag. In contrast to the approach adopted by Dick-Nielsen (2013), I added the buy-sell flag, which was used in Asquith et al. (2013), to improve the matching. Furthermore, I followed the suggestion of Dick-Nielsen (2013) and only compared reports that have a trade execution date before the reversal or as-of filing's report date. First, I deleted all reversed trades from my data set and the matching reversals. I then repeated this step by matching reversals and as-of reports. Finally, all remaining unmatched reversals were also deleted.

Next I checked for reports filed with a delay; no such reports were contained in my data set.

Agency trades are a further source of error. Generally a dealer can act in a trade in a principal capacity or as an agent for another market participant. When she acts as an agent and another dealer is executing the trade for his customer, one trade results in three reports in TRACE: Two reports describing the execution of the interdealer trade for each of the two counterparties and one report describing the following customer trade between the socalled 'introducing dealer'

and her customer. Two of these three records must be deleted from the database. Similar to Dick-Nielsen (2013), I matched the associated trade reports using the bond symbol, trade execution date, reported price and the full volume quoted. Further matching was restricted to matches with at least one report indicated as an agency trade. When more than three reports could be matched I only deleted two of them. This procedure was conducted for sell- and buy-side agency trades.

In a next step, I followed Asquith et al. (2013) and deleted all duplicates from interdealer trades. If dealers trade among one another, both sides of the trade have an obligation to report and this results in two reports for one interdealer trade. I adopted a very conservative approach in cleaning my data set to eliminate interdealer trade duplicates: I first matched reports by bond symbol, POSIX timestamp (combining trade execution date and trade execution time), the reported price, the full volume quoted, and a contra-party flag, indicating that the trade was between dealers. In a next step, I repeated this procedure using the trade execution date only instead of a timestamp. As the dealers involved could report different trade execution times, this second step checks for missing matches in the first step. Finally, I reduced the number of reports by all unmatched buy-side interdealer trades as in Asquith et al. (2013).

The final cleaning step is described fully in Dick-Nielsen (2013) and deletes reports from trades when the security was issued, trades from primary markets, trades executed under special circumstances and in equity-linked securities and trades with non-standard settlement, without cash payment and with commission. In my data set, no such reports were found.

Statistics for deleted records in my enhanced TRACE data set are summarized in table B.1.

B. Tables

Table B.1.: **Clean-up statistics.** The table shows report statistics from cleaning the enhanced TRACE data set collected from WRDS. While the number of securities was reduced by the deletion of reports, the number companies remained stable at 389.

	Number of Securities	Number of reports
Raw data	**10,818**	**4,483,549**
Corrections		31,842
Cancellations		82,658
Reversed trades		43,900
Reversed as-of trades		3,085
Remaining reversals		10,205
Agency trades		626,631
Inter-dealer trades by date-time		656,026
Inter-dealer trades by date		901,328
Inter-dealer trades all buys		235,807
Deleted		**2,594,567**
Remaining	**10,735**	**1,888,982**

Table B.2.: **Daily trading volumes.** The table shows daily contract volumes on bond and corresponding stock markets used in the comparison analysis. Decile 3 contains 39 companies, decile 5 has 35 companies, and decile 10 is comprised of 34 companies. The average number of trades, 'N', is rounded up/down to the next integer value. For the bond data from TRACE single trade volumes were aggregated and for the stock data daily share volumes were collected from the Compustat Daily database. Means and medians were used to summarize daily volumes in a decile.

		Mean					**Median**			
	DECILE	N	SUM	MAX	MIN	STD	SUM	MAX	MIN	STD
Bonds										
	3	87	203,257.9	34,897.92	8.974359	5,683.577	204,492	31,550	5	5,332.102
	5	92	379,644.1	64,754.34	9.057143	10,732.319	382,346	43,244	5	7,694.139
	10	90	4,343,912.6	1,716,467.82	209.735294	205,891.987	3,018,882	1,656,765	37	198,204.050
Stocks										
	3	122	449,788,749	13,878,202	1,501,749	1,727,601	357,128,048	11,287,470	1,168,290	1,305,316
	5	121	582,675,282	20,347,779	1,811,060	2,686,582	486,864,990	14,036,380	1,359,658	1,701,813
	10	122	2,905,510,840	97,364,368	10,638,535	11,476,339	1,269,504,626	35,936,095	4,801,466	4,418,355

Table B.3.: **Number of trades.** The table shows trade summaries for 108 bonds and corresponding stocks used in the comparison analysis. Decile 3 contains 39 companies, decile 5 has 35 companies and decile 10 is comprised of 34 companies. The average number of days with trades, 'N' is rounded up/down to the next integer value. Data for the number of trades on bond markets were collected from reported trades in the enhanced TRACE data set and corresponding values for the stock markets were taken from the former online database of the Financial Markets Research Center of the Vanderbilt University.

		Mean					**Median**			
	DECILE	N	SUM	MEAN	MAX	MIN	SUM	MEAN	MAX	MIN
Bonds										
	3	87	320.6667	3.179432	14.69231	1.102564	189	2	9	1
	5	92	349.8857	3.641377	22.22857	1.028571	248	2	12	1
	10	90	1,314.0000	14.849650	163.02941	2.676471	1,018	10	176	1
Stocks										
	3	122	1,951,678	15,899.87	54,481.03	7,294.769	1,885,349	15,328	50,521	6,791
	5	121	2,466,261	20,253.49	66,569.63	8,758.600	2,353,870	19,137	53,360	7,421
	10	122	7,150,904	58,240.39	187,085.74	29,831.176	4,631,340	37,653	107,888	19,929

Table B.4.: **Variable summary.** The table shows a summary of the regressors used in cross-sectional regression. Counting variables show their sum and numeric values show median, mean and standard deviation.

	MEDIAN/SUM	MEAN	STD. DEVIATION
PIN	0.0128	0.15	0.188
Issuance			
Maturity (days)	2,526	3,694	3,596
Age (days)	1,359	1,724	1,459
Coupon	5.62	5.32	2.03
VariableCoupon	404		
ZeroCoupon	137		
InterestFrequency	2	2.9	3.7
OfferingAmount ($)	400,000,000	554,224,517	673,415,325
Denomination	1	4.268	18.36
Convertible	65		
Exchangeable	13		
InBankruptcy	5		
TenderOffer	443		
RatedInRange	203		
OffWatch	981		
AAA	84		
AA	560		
A	1,686		
BBB	1,102		
S&P	544		
Moody's	1,087		
Fitch	2,524		
Market			
Volume ($)	86,614,000	301,114,377	688,569,230
MM1	13	31.26	54.62
MM5	4	16.50	37.87
Amihud ($ mio.)	0.05432	0.1148	0.5636
Roll	1.486	1.607	0.8031
Dispersion	6.536e-03	8.741e-03	7.627e-03
PriceStd	0.5217	0.5994	0.4214
Issuer			
IssuerStockReturn (%)	0.05213	0.0642	0.3877
IssuerStockVolatility (%)	1.461	1.803	4.147
FinOrUtil	2,530		

Table B.5.: **Cross-sectional analysis.** The table shows results from two cross-sectional regressions using the probability of informed trading as dependent variable. Standard errors in the first regression were computed using a heteroskedasticity-consistent variance-covariance matrix defined in the `R`-package `sandwich` (namely `HC4m`), and the second regression uses cluster-robust standard errors where clusters are determined by the issuer. All estimations were conducted in `R-3.0.2`.

Dep. Var.: PIN	**Heteroskedastic**				**Clustered**			
	ESTIMATE	STD. ERROR	t-VALUE	P($>\|t\|$)	ESTIMATE	STD. ERROR	t-VALUE	P($>\|t\|$)
(Intercept)	-1.62e-01	2.51e-02	-6.44	0e+00	-1.62e-01	2.32e-02	-6.96	0e+00
Maturity	-2.09e-06	5.87e-07	-3.56	0.0004	-2.09e-06	5.63e-07	-3.72	0.0002
Age	-2.02e-06	1.18e-06	-1.72	0.0855	-2.02e-06	1.55e-06	-1.30	0.1934
Coupon	5.40e-03	1.10e-03	4.89	0e+00	5.40e-03	1.74e-03	3.10	0.0019
VariableCoupon	5.98e-02	6.69e-03	8.95	<2e-16	5.98e-02	1.03e-02	5.83	0e+00
log(OfferingAmount)	-1.47e-02	2.23e-03	-6.57	0e+00	-1.47e-02	2.79e-03	-5.26	0e+00
Convertible	4.66e-02	1.83e-02	2.54	0.0110	4.66e-02	1.95e-02	2.39	0.0171
InBankruptcy	-1.60e-01	6.65e-02	-2.40	0.0165	-1.60e-01	1.94e-02	-8.21	0e+00
TenderOffer	1.53e-02	5.50e-03	2.78	0.0055	1.53e-02	5.74e-03	2.66	0.0078
RatedInRange	2.78e-01	1.38e-02	20.18	<2e-16	2.78e-01	1.75e-02	15.89	<2e-16
AAAdum	-3.22e-02	1.32e-02	-2.45	0.0144	-3.22e-02	9.41e-03	-3.42	0.0006
AAdum	-1.33e-02	5.72e-03	-2.32	0.0203	-1.33e-02	7.15e-03	-1.86	0.0634
Adum	-1.62e-02	3.64e-03	-4.47	0e+00	-1.62e-02	5.03e-03	-3.23	0.0013
S&P	1.18e-02	5.20e-03	2.27	0.0235	1.18e-02	5.94e-03	1.98	0.0472
Moody's	1.69e-02	4.23e-03	4.00	1e-04	1.69e-02	5.02e-03	3.37	0.0008
log(Volume)	2.07e-02	2.16e-03	9.60	<2e-16	2.07e-02	2.20e-03	9.39	<2e-16
MM1	4.98e-04	1.40e-04	3.56	0.0004	4.98e-04	1.53e-04	3.27	0.0011
MM5	-9.45e-04	1.41e-04	-6.68	0e+00	-9.45e-04	1.38e-04	-6.86	0e+00
Dispersion	6.97e+00	4.30e-01	16.22	<2e-16	6.97e+00	5.15e-01	13.53	<2e-16
PriceStd	-2.82e-02	7.55e-03	-3.74	0.0002	-2.82e-02	7.13e-03	-3.96	1e-04
IssuerStockReturn	-3.69e-02	1.07e-02	-3.45	0.0006	-3.69e-02	1.09e-02	-3.39	0.0007
IssuerStockVolatility	3.34e-03	9.79e-04	3.41	0.0007	3.34e-03	1.00e-03	3.33	0.0009
Adj. R-squared	0.476				0.476			
F-statistic	112.7861			<2.22e-16	9,299.693			<2.22e-16
Residual Std. Err.	0.09837898				0.09837898			

Table B.6.: **Comparison analysis.** The table shows estimation results of the compressed model for fixed-income markets (Bonds) and equity markets (Stocks). The mean and the median are used to aggregate the results in each of the three deciles of a market. Results were obtained by an MCMC procedure using a self-coded `R`-package and `R-3.0.2`.

				Parameters				**Standard Errors**				
	DECILE	COMP	$\overline{\text{N}}$	α	ϵ	μ	PIN	α	ϵ	μ	PIN	LOGLIK
Bonds (avg.)												
	3	39	87	0.281	0.00346	0.0145	0.0998	0.207	0.00285	0.0221	0.105	-171
	5	35	92	0.234	0.00378	0.0306	0.1608	0.208	0.00205	0.0589	0.127	-194
	10	34	90	0.133	0.01409	0.2868	0.3123	0.169	0.01247	0.2347	0.193	-350
Stocks (avg.)												
	3	39	122	0.291	16.8	30.3	0.183	0.1051	10.4	25.9	0.0364	-53,619
	5	35	121	0.289	20.9	41.3	0.189	0.1164	10.6	31.0	0.0533	-73,602
	10	34	122	0.258	62.8	118.7	0.161	0.0957	47.5	141.2	0.0451	-176,127
Bonds (med.)												
	3	39	87	0.2579	0.00264	0.00432	0.085					-145
	5	35	92	0.1489	0.00305	0.00846	0.125					-171
	10	34	90	0.0512	0.01042	0.30043	0.288					-327
Stocks (med.)												
	3	39	122	0.302	16.1	25.3	0.180					-43,574
	5	35	121	0.310	20.1	33.8	0.190					-51,283
	10	34	122	0.284	40.5	73.4	0.158					-105,900

Table B.7.: **Kruskal-Wallis tests.** The table shows the results of Kruskal-Wallis tests applied to the parameters of the compressed model and the PIN of all deciles for the bond market (Bonds) and the equity markets (Stocks). Tests were conducted in `R-3.0.2` using the function `kruskal.test`.

	STATISTIC	PROBABILITY
Bonds		
α	13.06	1.46e-03
ϵ	48.78	2.55e-11
μ	44.10	2.65e-10
PIN	27.62	1.01e-06
Stocks		
α	2.06	3.57e-01
ϵ	42.14	7.09e-10
μ	30.64	2.22e-07
PIN	9.31	9.50e-03

Table B.8.: **Whitney-Mann tests.** The table shows the Whitney-Mann statistics for the parameters of the compressed model for all combinations of deciles of bond markets (Bonds) and stock markets (Stocks). All tests were conducted in `R-3.0.2` using the `wilcox.test` with exact p-values.

	3rd to 5th		**3rd to 10th**		**5th to 10th**	
	STAT.	PROB.	STAT.	PROB.	STAT.	PROB.
Bonds						
α	766	0.3710	967	6.27e-04	824	5.58e-03
ϵ	516	0.0722	102	5.13e-12	120	4.94e-10
μ	544	0.1358	118	3.15e-11	153	1.28e-08
PIN	493	0.0402	207	9.53e-08	310	4.89e-04
Stocks						
α	676	0.9485	781	1.95e-01	694	2.39e-01
ϵ	495	0.0424	134	1.68e-10	173	7.31e-08
μ	512	0.0655	191	2.77e-08	268	5.33e-05
PIN	677	0.9571	927	3.18e-03	792	1.77e-02

Table B.9.: **News probability tests.** The table shows the results of Kruskal-Wallis and Whitney-Mann tests applied to the parameter α between bond and stock markets. Tests were conducted in `R-3.0.2` using the functions `kruskal.test` and `wilcoxon.test`.

DECILE	STATISTIC	PROBABILITY
Kruskal-Wallis		
3	0.09	7.68e-01
5	3.34	6.78e-02
10	14.36	1.51e-04
Whitney-Mann		
3	790	7.73e-01
5	768	6.85e-02
10	887	9.92e-05

Bibliography

Akaike, H., 1974. A new look at the statistical model identification. Automatic Control, IEEE Transactions on 19 (6), 716–723.

Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. Journal of financial markets 5 (1), 31–56.

Asquith, P., Covert, T., Pathak, P., 2013. The effects of mandatory transparency in financial market design: Evidence from the corporate bond market. Tech. rep., National Bureau of Economic Research.

Bessembinder, H., Maxwell, W., 2008. Markets: Transparency and the corporate bond market. The Journal of Economic Perspectives 22 (2), 217–234.

Bessembinder, H., Maxwell, W., Venkataraman, K., 2006. Market transparency, liquidity externalities, and institutional trading costs in corporate bonds. Journal of Financial Economics 82 (2), 251–288.

Biais, B., Green, R. C., 2007. The microstructure of the bond market in the 20th century. Tepper School of Business, 134.

Bjønnes, G. H., Rime, D., 2005. Dealer behavior and trading systems in foreign exchange markets. Journal of Financial Economics 75 (3), 571–605.

Bond Market Association, December 2005. eCommerce in the fixed-income markets. online.

Cai, F., Han, S., Li, D., 2012. Institutional herding in the corporate bond market. Federal Reserve Board.

Calcagno, R., Lovo, S., 2006. Bid-ask price competition with asymmetric information between market-makers. The Review of Economic Studies 73 (2), 329–355.

Cameron, A. C., Gelbach, J. B., Miller, D. L., 2011. Robust inference with multiway clustering. Journal of Business & Economic Statistics 29 (2).

Copeland, T. E., Galai, D., 1983. Information effects on the bid-ask spread. the Journal of Finance 38 (5), 1457–1469.

Cribari-Neto, F., da Silva, W. B., 2011. A new heteroskedasticity-consistent covariance matrix estimator for the linear regression model. AStA Advances in Statistical Analysis 95 (2), 129–146.

Dick-Nielsen, J., 2009. Liquidity biases in trace. Journal of Fixed Income 19 (2), 43.

Dick-Nielsen, J., 2013. How to clean enhanced trace data. Available at SSRN 2337908.

Dick-Nielsen, J., Feldhütter, P., Lando, D., 2012. Corporate bond liquidity before and after the onset of the subprime crisis. Journal of Financial Economics 103 (3), 471–492.

Easley, D., Engle, R. F., O'Hara, M., Wu, L., 2008. Time-varying arrival rates of informed and uninformed trades. Journal of Financial Econometrics 6 (2), 171–207.

Easley, D., Kiefer, N., O'Hara, M., Paperman, J., 1996. Liquidity, information, and infrequently traded stocks. Journal of Finance 51, 1405–1436.

Easley, D., O'Hara, M., 1987. Price, trade size, and information in securities markets. Journal of Financial economics 19 (1), 69–90.

Eddelbuettel, D., Sanderson, C., 2013. RcppArmadillo: accelerating R with high-performance C++ linear algebra. ComputationalStatistics and Data Analysis in press.
URL http://dx.doi.org/10.1016/j.csda.2013.02.005

Edwards, A. K., February 2006. Corporate bond market microstruture and transparency - the US experience. BIS papers (26).

Edwards, A. K., Nimalendran, M., Piwowar, M. S., 2006. Corporate bond market transparency: Liquidity concentration, informational efficiency, and competition. document de travail, Securities Exchange Commission.

Evans, M. D., Lyons, R. K., inter-dealer, information 1999. Order flow and exchange rate dynamics. Tech. rep., National Bureau of Economic Research.

Feldhütter, P., 2012. The same bond at different prices: identifying search frictions and selling pressures. Review of Financial Studies 25 (4), 1155–1206.

Ferriani, F., 2010. Informed and uninformed traders at work: evidence from the french market.

FINRA, February 2005. NASD's fully implemented "trace" brings unprecedented transparency to corporate bond market. Tech. rep., FINRA.
URL http://www.finra.org/Newsroom/NewsReleases/2005/p013274

Glosten, L., Harris, L., 1988. Estimating the Components of the Bid-Ask Spread. Journal of Financial Economics 21 (1), 123–142.

Glosten, L. R., Milgrom, P. R., 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. Journal of financial economics 14 (1), 71–100.

Goldstein, M. A., Hotchkiss, E. S., Sirri, E. R., 2007. Transparency and liquidity: A controlled experiment on corporate bonds. Review of Financial Studies 20 (2), 235–273.

Grammig, J., Schiereck, D., Theissen, E., 2001. Knowing me, knowing you:: Trader anonymity and informed trading in parallel markets. Journal of Financial Markets 4 (4), 385–412.

Grammig, J., Theissen, E., Zehnder, L. S., October 2014. Bayesian estimation of the probability of informed trading.

Gu, Z., Zhao, J. Y., 2006. Information precision and the cost of debt.

Han, S., Zhou, X., 2011. Informed bond trading, corporate yield spreads, and corporate default prediction. Corporate Yield Spreads, and Corporate Default Prediction (April 15, 2013).

Han, S., Zhou, X., 2013. Informed bond trading, corporate yield spreads, and corporate default prediction. Management Science 0 (0), 1–20.

Harris, L., 2002. Trading and exchanges: Market microstructure for practitioners. Oxford university press.

Hasbrouck, J., 1988. Trades, quotes, inventories, and information. Journal of Financial Economics 22 (2), 229–252.

Ho, T., Stoll, H. R., 1981. Optimal dealer pricing under transactions and return uncertainty. Journal of Financial economics 9 (1), 47–73.

Huang, R., Stoll, H., 1997. The components of the bid-ask spread: A general approach. Review of Financial Studies 10 (4), 995–1034.

Huang, R. D., Cai, J., Wang, X., 2002. Information-based trading in the treasury note interdealer broker market. Journal of Financial Intermediation 11 (3), 269–296.

Kishore, A., 2013. Machine learning and algo trading in fixed income markets. Available at SSRN 2305886.

Kokot, S., 2004. The Econometrics of Sequential Trade Models: Theory and Applications Using High Frequency Data. Springer.

Levitt, A., September 1998. The importance of transparency in America's debt market. `http://www.sec.gov/news/speech/speecharchive/1998/spch218.htm`.

Li, D., Schürhoff, N., 2012. Dealer networks, available at ssrn.com.

Lyons, R. K., 1997. A simultaneous trade model of the foreign exchange hot potato. Journal of International Economics 42 (3-4), 275–298.

Madhavan, A., Richardson, M., Roomans, M., 1997. Why do security prices change? A transaction-level analysis of NYSE stocks. Review of Financial Studies 10 (4), 1035–1064.

Malinova, K., Park, A., Riordan, R., 2013. Do retail traders benefit from improvements in liquidity? Tech. rep., Working paper, Nov.

Marsh, I., O'Rourke, C., 2005. Customer order flow and exchange rate movements: is there really information content? Cass Business School Research Paper.

Menkhoff, L., Osler, C. L., Schmeling, M., 2010. Limit-order submission strategies under asymmetric information. Journal of Banking & Finance 34 (11), 2665–2677.

Menkhoff, L., Schmeling, M., 2008. Local information in foreign exchange markets. Journal of International Money and Finance 27 (8), 1383–1406.

Menkhoff, L., Schmeling, M., 2010a. Trader see, trader do: How do (small) FX traders react to large counterparties' trades? Journal of International Money and Finance 29 (7), 1283–1302.

Menkhoff, L., Schmeling, M., 2010b. Whose trades convey information? evidence from a cross-section of traders. Journal of Financial Markets 13 (1), 101–128.

Payne, R., 2003. Informed trade in spot foreign exchange markets: an empirical investigation. Journal of International Economics 61 (2), 307–329.

Randall, O., January 2013. Pricing and liquidity in the US corporate bond market.

Roll, R., 1984. A simple implicit measure of the effective bid-ask spread in an efficient market. The Journal of Finance 39 (4), 1127–1139.

Saunders, A., Srinivasan, A., Walter, I., 2002. Price formation in the otc corporate bond markets: a field study of the inter-dealer market. Journal of Economics and Business 54 (1), 95–113.

Securities Industry and Financial Markets Association, 2006. The role of interdealer brokers in the fixed income markets.

Stoll, H. R., 1978. The supply of dealer services in securities markets. The Journal of Finance 33 (4), 1133–1151.

Wittenberg-Moerman, R., 2008. The role of information asymmetry and financial reporting quality in debt trading: Evidence from the secondary loan market. Journal of Accounting and Economics 46 (2), 240–260.

Wuensche, O., 2007. Using mixed Poisson distributions in sequential trade models. Available at SSRN 1010989.