Estimating the Probability of Informed Trading - Does Trade Misclassification Matter?

Joachim Grammig und Erik Theissen

Januar 2003 Discussion Paper no. 2003-01

Department of Economics University of St. Gallen

Editor: Prof. Jörg Baumberger
University of St. Gallen
Department of Economics
Bodanstr. 1
CH-9000 St. Gallen
Phone ++41 71 224 22 41
Fax ++41 71 224 28 85
Email joerg.baumberger@unisg.ch

Publisher: Forschungsgemeinschaft für Nationalökonomie
an der Universität St. Gallen
Dufourstrasse 48
CH-9000 St. Gallen
Phone ++41 71 224 23 00
Fax ++41 71 224 26 46

Electronic Publication: www.fgn.unisg.ch/public/public.htm

Estimating the Probability of Informed Trading - Does Trade Misclassification Matter?

Joachim Grammig und Erik Theissen

Author's address:

Prof. Dr. Joachim Grammig
Schweizerisches Institut für Banken und Finanzen
Universität St. Gallen
Rosenbergstr. 52
CH-9000 St. Gallen
Tel. ++41 71 2247090
Fax ++41 71 2247088
Email joachim.grammig@unisg.ch
Website www.sbf.unisg.ch

We thank seminar participants at the University of St. Gallen for helpful comments. Joachim Grammig, University of St. Gallen, Swiss Institute of Banking and Finance, Rosenbergstr. 52, 9000 St. Gallen, Switzerland, Email: Joachim.Grammig@unisg.ch; Erik Theissen, University of Bonn, BWL I, Adenauerallee 24-42, 53113 Bonn, Germany, Email: theissen@uni-bonn.de.

Abstract: Easley / Kiefer / O'Hara / Paperman (1996) (EKOP) have proposed an empirical methodology that allows to estimate the probability of informed trading and that has subsequently been used to address a wide range of issues in market microstructure. The data needed for estimation is the number of buyer- and seller-initiated trades. This information often has to be inferred by applying trade classification algorithms like the one proposed by Lee / Ready (1991). These algorithms are known to be inaccurate. In this paper we perform extensive simulations to show that inaccurate trade classification leads to biased estimation of the probability of informed trading when applying the EKOP methodology. The estimate is biased downward and the magnitude of the bias is related to the trading intensity of the stock in question. Scrutinizing prior empirical studies using the EKOP methodology, we conclude that the bias may severely affect the results of empirical microstructure studies.

Keywords:

Informed trading, market microstructure, trade classification

JEL classification:

C52, G10, G14

1 Introduction

The interest in market microstructure research, particularly in empirical research, has increased substantially in recent years. New methods enable researchers to address questions that have not been amenable to empirical investigation before. An important contribution in that respect is the structural model first proposed by Easley et al. (1996), henceforth referred to as EKOP. It builds on the theoretical work of Easley / O'Hara (1987, 1992) and allows to directly estimate the (unconditional) probability of informed trading.

An attractive feature of the EKOP methodology is its apparently modest data requirement. All that is needed to estimate the model is to count the number of buyer- and seller-initiated trades for each stock and each trading day. This requirement is, however, less innocuous than it appears. For many markets (e.g., the New York Stock Exchange, NASDAQ, and the Frankfurt Stock Exchange, to name but a few), this kind of data is not readily available. The trade classification is usually inferred from trade and quote data using the method proposed by Lee / Ready (1991) or modifications thereof. These algorithms are known to be less than perfectly accurate. Estimates of the misclassification frequency range from 7% (Lee / Radhakrishna 2000 for the NYSE) to 25% (Theissen 2001 for the Frankfurt Stock Exchange).[1]

In the present paper we argue that such misclassification may lead to biased results, in particular to an underestimation of the probability of informed trading, the key variable in most applications of the methodology. We provide the intuition for our claim and perform extensive simulations to substantiate it. We calibrate the parameters of our simulation model in a way that allows us to analyze the extent to which the results of previous studies are affected by this bias.

The results indicate that the bias introduced by trade misclassification is substantial. As an example, consider parameter values that are representative of the 5th volume decile of NYSE stocks. The probability of informed trading is 21.44%. However, with 15% misclassification (the misclassification rate reported by Odders-White 2000 for the NYSE), the estimate of the probability of informed trading is only 14 %. Our results further show that the bias is more pronounced for less liquid stocks. Consequently, differences in the probability of informed trading between liquid and less liquid stocks are likely to be even higher than previously estimated.

We reconsider recent empirical results obtained by applying the EKOP methodology to a variety of issues in empirical finance (e.g. market design, liquidity and asset pricing) and argue that some of the conclusions drawn in those papers may, at least partially, be an artefact of the downward bias that affects the estimation of the probability of informed trading.

The remainder of the paper is organized as follows. In section 2 we lay out our argument that trade misclassification leads to biased results in more detail. In section 3 we describe our simulations and document their results. Section 4 discusses implications for the interpretation of previous results obtained using the EKOP methodology, section 5 concludes.

2 Theory

The EKOP model assumes the existence of competitive market makers, informed traders and uninformed liquidity traders. Liquidity traders are assumed to be equally likely to buy or sell shares. The arrival of their buy and sell orders at the market maker's desk is modeled as two independent Poisson processes with identical intensity parameter ε. Before trading starts, an information event occurs with probability α. Informed agents only trade when an information

---

[1] Other papers analyzing the accuracy of the Lee / Ready trade classification algorithm include Ellis / Michaely /

event has occurred. If the information is good news (this happens with probability 1-δ) informed traders buy, if it is bad news (this event occurs with probability δ) they sell. The arrival of informed 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 an information event is negative (δ) and the order arrival rates of informed and uninformed traders (μ and ε) – can straightforwardly be estimated by Maximum Likelihood.

In their original paper, Easley et al. (1996) estimate the unconditional probability of informed trading

$$PI = \frac{\alpha\mu}{\alpha\mu + 2\varepsilon} \tag{1}$$

for stocks with differing trading volume. This methodology has subsequently been adopted to address a variety of issues. These comprise (but are not limited to) informational aspects of analyst coverage (Easley / O'Hara / Paperman 1998) and stock splits (Easley / O'Hara / Saar 2001), the relation between information risk and expected returns (Easley / Hvidkjaer / O'Hara 2002) and the role of anonymity in the trading process (Grammig / Schiereck / Theissen 2001).

Most of these papers center around the question of whether certain characteristics of a stock (e.g., its trading volume) or of the trading process (e.g., the degree of anonymity) affect PI. Addressing this question using the EKOP methodology appears to be a straightforward exercise: All that is needed to estimate the crucial parameters (namely, the arrival rates ε and μ and the news probability α) is to count the number of buyer and seller initiated trades per day for a sample of trading days. The key question that motivates the present paper is: what happens if

---

O'Hara (2000), Finucane (2000), Odders-White (2000) and Savickas / Wilson (2001).

trades cannot be accurately classified as buyer- or seller-initiated, i.e., if trade classification errors occur. In the following we show that the estimate of PI will be biased downwards in the presence of trade misclassification.

Let $\tilde{x}_{i,t}$ be a random variable characterizing the true trade classification of a transaction at time t of trading day i.[2] Let $\tilde{x}_t$ be 1 for a buyer-initiated transaction with probability p and -1 for a seller-initiated transaction with probability $1-p$. The structural model underlying the EKOP methodology implies that, on a given trading day i, the $\tilde{x}_t$ are i.i.d. with constant probability p equal to

- $\frac{\varepsilon+\mu}{2\varepsilon+\mu}$ on a day with a positive information event,
- $\frac{\varepsilon}{2\varepsilon+\mu}$ on a day with a negative information event,
- 0.5 on a day without information event.

Now consider misclassification of trades. Let $\tilde{\eta}_t$ be a binary random variable indicating whether the trade at time t is correctly classified, $\Pr(\tilde{\eta}_t = 1)=q$, or misclassified, $\Pr(\tilde{\eta}_t = -1)=1-q$. We will refer to $1-q$ as the misclassification rate. For simplicity, assume that the $\tilde{\eta}_t$ are i.i.d. and independent of the $\tilde{x}_t$. The **observed** trade classification is $\tilde{y}_t = \tilde{x}_t\tilde{\eta}_t$. The trade indicator variable $\tilde{y}_t$ takes on the value 1 if the trade is classified as a buyer initiated trade. This comprises correctly classified buyer initiated trades (i.e. $\tilde{x}_t = 1$ and $\tilde{\eta}_t = 1$) and mis-classified seller-initiated trades (i.e. $\tilde{x}_t = -1$ and $\tilde{\eta}_t = -1$). $\tilde{y}_t$ takes on the value –1, thereby identifying a seller initiated trade, either if the trade is correctly classified as seller initated (i.e.

[2] For ease of notation we suppress the index i in the remainder of this section.

$\tilde{x}_t = -1$ and $\tilde{\eta}_t = 1$) or if a buyer-initated trade is misclassified (i.e. $\tilde{x}_t = 1$ and $\tilde{\eta}_t = -1$). This implies that the probability of classifying a trade as buyer initiated is

$$\Pr(\tilde{y}_t = 1) = (1 - p - q + 2pq) \tag{2}$$

and the probability for classifying a trade as seller initiated is

$$\Pr(\tilde{y}_t = -1) = (p + q - 2pq) \tag{3}$$

Define the classification bias as the difference between the observed and the true probability of a buyer-initiated trade. The bias is thus

$$(1 - p - q + 2pq) - p = (2p - 1)(1 - q) \tag{4}$$

It is immediately apparent that the bias is nonzero whenever misclassification occurs (i.e., q is less than one) and the true probabilities of observing a buyer- or seller initiated trade are not identical (i.e., $p \neq 0.5$). Furthermore, in the presence of misclassification, the probability of a buyer-initiated trade is **overestimated** whenever $p < 0.5$ and is **underestimated** in case of $p > 0.5$.

Figure 1 graphs the observed probability of a buyer-initiated trade as a function of the true probability for different misclassification rates. It is evident that the bias can be substantial. To gain an impression of the size of the bias, consider the parameter values $\varepsilon = 0.175, \mu = 0.14, q = 0.85$. The values for the trading intensities correspond to those estimated for the top panel of NYSE stocks by EKOP and the assumed misclassification rate of 0.15 conforms to empirical results for the NYSE (e.g. Odders-White 2000). The true probability of observing a buyer-initiated trade on a day with a positive information event amounts to $\frac{\varepsilon + \mu}{2\varepsilon + \mu} = 0.643$ and the true probability of observing a seller-initiated trade is 1-0.643=0.357.

However, the observed probabilities, calculated using (2) and (3), are 0.6 and 0.4, respectively. Thus, the probability of a buyer-initiated trade is underestimated by 6.7% on days with positive information events and is overestimated by 12% on days with a negative information event.

Insert Figure 1 about here

What are the consequences of the classification bias for the estimation of EKOP type models? Before we address this question by means of a simulation study in the next section, let us first provide some intuition: In the absence of informed trading, buyer- and seller-initiated trades would be equally likely. If, on a given trading day, a large discrepancy between the number of buyer- and seller-initiated trades is observed, this indicates that informed traders are present. The estimated probability of informed trading tends to be larger the more frequent, and the more pronounced, the discrepancies between the number of buyer- and seller-initiated trades in a sample of i=1,..., M trading days are. Maximization of the log-likelihood-function of the EKOP model will then produce, ceteris paribus a relatively high estimate of the ratio of the arrival rates $\frac{\mu}{\varepsilon}$, indicating activity of informed traders. Now, as argued above, in the presence of misclassification the number of buyer-initiated trades will be understated whenever $p > 0.5$ and overstated whenever $p < 0.5$. The reverse is true for seller-initiated trades. Since p is larger [smaller] than 0.5 when a positive [negative] information event has occurred, it follows that misclassification leads to a reduction of the discrepancy between the number of buyer- and seller-initiated trades. Consequently, the EKOP methodology will underestimate both the ratio of informed to uninformed arrival intensities $\frac{\mu}{\varepsilon}$ and, hence, **underestimate** the probability of informed trading in the presence of trade misclassification. This can easily be seen by rewriting equation (1) as

$$PI = \frac{1}{1 + \frac{2\varepsilon}{\alpha\mu}}$$

The magnitude of the bias depends on the probability of observing a buyer-initiated trade, p, and the misclassification rate, (1-q). Since p depends on the intensity parameters ε, μ, we hypothesize that the downward bias of the estimated probability of informed trading will, ceteris paribus, be larger

- the smaller the intensity of liquidity trading, ε (because a decrease in ε moves the p on days with informed trading away from 0.5 which, in turn, increases the bias),
- the larger the intensity of informed trading, μ (because an increase in μ moves p away from 0.5 which, in turn, increases the bias) and
- the larger the misclassification rate (1 - q).

So far, we have substantiated our claim that trade misclassification is likely to bias estimates of the probability of informed trading obtained using the EKOP methodology. It remains to be seen, however, whether the magnitude of the bias is economically significant. This is the issue to which we will now turn.

3 Simulations

To assess the sensitivity of the estimate of the probability of informed trading to trade classification errors and to address the hypotheses formulated above we design the following Monte Carlo study. The basic idea is to simulate the trading process described in the sequential trading model proposed by Easley et al. (1996) (EKOP) with and without trade classification errors and to study the effect of trade misclassification on the parameter estimates. The model parameters α, δ, ε, μ vary across simulation designs, but other model characteristics are held

fixed, namely the length of the trading day, *T*, and the number of trading days, *M*. *T* is set equal to 6.5 hours (the length of a NYSE trading day) and *M*=60. This corresponds to the number of trading days considered in Easley et al. (1996). By simulating the trading process implied by the EKOP model we obtain, for each trading day *i,* the (true) number of buyer and seller initiated trades B_i and S_i. Based on the simulated data, we can estimate the model parameters by maximizing the log likelihood function implied by the EKOP model which is given by:

$$L = \sum_{i=1}^{M} ln\left[\alpha\delta e^{-\varepsilon T} \frac{(\varepsilon T)^{B_i}}{B_i!} e^{-(\varepsilon+\mu)T} \frac{[(\varepsilon+\mu)T]^{S_i}}{S_i!} + (1-\alpha) e^{-\varepsilon T} \frac{(\varepsilon T)^{B_i}}{B_i!} e^{-\varepsilon T} \frac{(\varepsilon T)^{S_i}}{S_i!} + \alpha(1-\delta) e^{-(\varepsilon+\mu)T} \frac{[(\varepsilon+\mu)T]^{B_i}}{B_i!} e^{-\varepsilon T} \frac{(\varepsilon T)^{S_i}}{S_i!} \right]. \quad (5)$$

The three terms that constitute the sum in brackets denote the probability of the "information characteristic" of the trading day (bad news, no news, good news) multiplied by the probability of observing the given number of buyer and seller initiated trades on this day (one can see that this is simply the product of two univariate Poisson probability density functions). Having obtained the ML estimates based on the true trade classification we introduce classification errors. We assume that each buyer-initiated trade may be falsely classified as seller-initiated, and each seller-initiated trade may be falsely classified as buyer-initiated. Misclassification rates are set to 0.1 and 0.15. This is the order of magnitude of misclassification rates found for the NYSE and Nasdaq. This “contamination” of the data generates new sequences of "observed" buyer and seller initiated trades which differ from the original series $\{B_i\}_{i=1}^{M}$ and $\{S_i\}_{i=1}^{M}$. Based on this misclassified data, ML estimation of the model parameters is repeated. The whole procedure is replicated *K*=100 times for each set of parameters, and both the ML estimates based on the correctly classified data and the “bogus” parameter estimates obtained in each replica-

tion *k* are stored. The bias associated with the estimation of the probability of informed trading is computed by substracting the estimated *PI* obtained in replication *k* from the true *PI* and averaging the difference over replications. The root mean squared error (RMSE) of the *PI* estimates is computed as the square root of the average squared difference between estimated and true *PI.* The exact formulas for bias and RMSE are given in the caption of Table I.

Insert Table I about here

Each row in Table I corresponds to a unique parameter constellation for which this simulation exercise is conducted. The first columns contain the set of true parameter values and the *PI* that is implied by them. The other columns report bias and RMSE associated with the estimation of *PI* that is induced by misclassification rates of 0.1 and 0.15, respectively.

To address the questions raised in the previous section, the news probabilities and the trade intensities, ε and μ, are varied over a range of meaningful values. In the first three rows of Table I, the model parameters α, ε, δ and μ are chosen to match the estimates in Easley et al. (1996) who have sorted NYSE stocks into trade volume deciles and estimated the model parameters for each stock by ML. The first row parameters correspond to the mean of the parameter estimates for the first trade volume decile, the second row takes the mean estimates of the fifth decile and the third row contains the mean parameters estimates of the eighth decile.

The simulation results confirm the hypotheses of the previous section. Trade misclassification induces a negative bias in the estimation of the probability of informed trade. The bias can be considerable both in absolute and in relative terms. Take as an example the small decile parmeters (simulation 3): The true probability of informed trade is 0.2253, whilst the bias when estimating *PI* on data that are misclassified with probability 0.15 amounts to -0.1085. Comparing RMSE and bias it is evident that the bias clearly dominates parameter estimation variance. Take as an example the medium volume decile where the RMSE of the *PI* estimate with

misclassification rate 0.15 is 0.079 and the bias amounts to -0.0742. This implies that trade misclassification does not increase estimation *variance*, as one could expect, but mainly *biases* the PI estimate. Table I shows that, as hypothesized above, the magnitude of the bias increases, ceteris paribus, with the misclassification rate and with increasing [decreasing] trade intensity of informed [uninformed] traders μ (ε).

An interesting result with considerable implications for empirical research is that the bias of the *PI* estimate is more pronounced for less frequently traded stocks. To see this, fix the news probability α and increase the arrival rates of informed and uninformed traders, but keep the composition of the trader population $\frac{\varepsilon}{\mu}$ constant. By equation (1) this implies that the true probability of informed trade is also constant. For example, in simulations 7, 39, 43 and 47 the trader population , $\frac{\varepsilon}{\mu}=1$, and the news probability, α=0.5, are identical.[3] This implies that $PI=\frac{1}{1+\frac{2}{0.5}}=0.2$ is also constant. However, the four simulation designs differ in the trading intensity which is highest in simulation 39 ($\varepsilon=\mu=0.2$) and smallest in simulation 7 $(\varepsilon=\mu=0.02)$.

Insert Figure 2 about here

Figure 2 clearly shows that both the bias in absolute terms and the relative bias is larger for less frequently traded stocks. As one of the classical application of the EKOP methodology is the analysis of the probability of informed trading for a cross section of stocks, this result exerts considerable consequences for empirical research. We will turn to this issue in the next section.

---

[3] The simulation results in table 1 also allow to fix the news probability α at different values (0.4. 0.3 and 0.2) and conduct the same analysis. The conclusions are not qualitatively different.

Insert Figure 3 about here

Figure 3 identifies the sources of the bias that haunts the estimation of *PI* from misclassified data. For this purpose we focus on simulation design 1 which takes Easley et al's (1996) mean estimates of the first volume decile of NYSE stocks. The simulation exercise is performed as described above, with the only difference that the number of replications is increased to K=1000 to produce smoother density plots. The bias affecting the model parameters under misclassification is displayed using kernel densities based on the empirical distribution of parameter estimates obtained in the 1000 replications. The graphs in the first row of Figure 3 show that the estimates of the probabilities α and δ are unbiased under misclassification: The kernel densities based on the ML estimates obtained when using correctly classified data on the one hand and the kernel densities using parameters estimated on misclassified data on the other barely differ. However, it can be seen in the second row of Figure 3 that trade classification error implies biased arrival intensity estimates. The arrival rate of uninformed traders is overestimated, whilst the arrival rate of informed traders is overestimated under trade misclassification. As outlined in the previous section, the fact that the ratio $\frac{\varepsilon}{\mu}$ is thus underestimated induces the downward bias when estimating *PI*. This bias is graphically highlighted by the kernel densities of the *PI* estimates in the third row of Figure 3.

4 Discussion

Our simulation results support our assertion that trade misclassification leads to downward-biased estimates of the probability of informed trading. The results of several papers making use of the EKOP methodology are thus likely to be affected by this bias. In what follows we scrutinize recent papers and analyze whether their results suffer from such a bias and, if so, whether the bias calls into question the qualitative implications of the papers.

In their original paper EKOP analyze whether the risk of information-based trading differs between low and high volume stocks and document that the probability of informed trading increases with volume. Therefore, the larger spreads for low-volume stocks can (at least partially) be explained by higher information risk. The results are affected by the bias because trade classification is inferred by applying the Lee / Ready algorithm to NYSE data.

Our simulations have shown that the estimated probability of informed trading, PI, is biased downward and that the bias is more pronounced for low-volume stocks. Consequently, with correct trade classification (if it were possible), the inverse relation between volume and PI would even be *stronger*. Hence, the qualitative result of the paper - low volume stocks have higher risk of information based trading - is clearly not due to misclassification bias.

The same applies to the results of Easley / O'Hara / Paperman (1998). They form pairs of stocks that differ in analyst coverage and find that trading volume is higher and PIs are lower for stocks with more extensive analyst coverage. As the Lee / Ready algorithm is applied, the downward bias in estimating PI is likely to be present. However, as stocks with higher analyst coverage have higher volume, our simulation results suggest that the downward bias is less pronounced compared to the group of stocks with low analyst coverage. The result that the PIs in the latter group are higher is thus not due to the bias. Hence, the paper's main conclusion that analysts serve to increase trading volume by showcasing stocks to uninformed traders, but not contribute significantly to price discovery, is strengthened.

Easley / O'Hara / Saar (2001) analyze whether stock splits affect the risk of information based trading. To that end, they analyze a sample of NYSE-listed stocks that experienced one or more stock splits in 1995 and estimate pre- and post-split PIs. They find that the order arrival rates (μ for informed traders and ε for uninformed traders) increase and that the PIs decrease, though not significantly. The authors take this as evidence that, contrary to a popular hypothe-

sis, there is no evidence that stock splits reduce information asymmetries. According to our simulation results, however, increased trading intensity reduces the downward bias in PI. Thus, if the true PIs remained constant, the estimated PIs should increase because of the reduced bias. The fact that Easley / O'Hara / Saar (2001) find *lower* PIs after the split indicates that there seems to be a reduction in information risk that is large enough to overcompensate the increase in the estimated PIs that is due to the reduced bias. Thus, if the PIs could be estimated without bias, the reduction in PI associated with a stock split would be more pronounced than reported by Easley / O'Hara / Saar (2001) and could well be significant, then supporting the hypothesis that stock splits reduce information asymmetries.

Another result that is challenged by our evidence is that information risk is a priced factor, an argument that was recently put forward by Easley / Hvidkjaer / O'Hara (2002). As argued above, low-volume stocks have higher PIs and exhibit a more pronounced downward bias in the estimation of PI. This implies that the true dispersion of the PIs is larger than the estimated dispersion. This, in turn, has implications for the results reported by Easley / Hvidkjaer / O'Hara (2002). They analyze the cross-section of asset returns and ask whether stocks with higher information risk (measured by a higher probability of informed trading) have higher expected returns. They use NYSE data, classify trades using the Lee / Ready method and then employ a two-step estimation procedure. In the first step, PI is estimated for each stock separately. In a second-pass regression, the PIs are related to realized returns. The significant and positive relation between PI and returns suggests that investors command (and, in equilibrium, receive) a higher return if they are to hold assets with higher information risk. Since, as noted above, the true dispersion of the PIs is larger than the estimated dispersion. This has two consequences. First, there is an errors-in-variables problem that is not cured by the portfolio approach described in Easley / Hvidkjaer / O'Hara (2002, p. 2213-2214) because the bias in the

PIs is systematic. Second, when the dispersion of the PIs in the sample is lower than the true dispersion, the coefficient on PI in the second-pass regression (which has a natural interpretation as the premium for information risk) is likely to be overstated.

As a final example we consider the study by Grammig / Schiereck / Theissen (2001). The authors use data from the German stock market and estimate the probability of informed trading for stocks that are simultaneously traded in a non-anonymous floor trading system and in an anonymous electronic auction market. They find that the PIs are higher in the latter, confirming their hypothesis that anonymity aggravates adverse selection problems. However, trades on the floor have to be classified using the Lee / Ready method whereas trade classification in the electronic trading system is accurate.[4] Consequently, the estimated PIs for the floor are downward-biased and the result that they are lower than those in the electronic trading system may (at least in part) be ascribed to trade misclassification on the floor.

5 Summary and conclusion

The empirical approach pioneered by Easley et al. (1996) has the attractive feature that it allows to estimate the probability of informed trading from, as it appears, easily available data - all that is needed is the number of buyer- and seller-initiated trades. However, data on trade classification is often unavailable. Empirical researchers try to get around this problem by applying trade classification algorithms like the one proposed by Lee / Ready (1991). However, these algorithms are known to be inaccurate.

In this paper we show that trade misclassification has severe consequences when applying the EKOP methodology. We argue that the probability of informed trading will be estimated with

---

[4] This is a consequence of the design of the trading system. The market under scrutiny, IBIS, was a hit-and-take system and transactions at prices inside the spread were impossible.

a downward bias and that the magnitude of the bias is related to the trading intensity of the stock in question. This claim is substantiated in extensive simulations.

We reconsider some prior empirical studies addressing issues in market microstructure and asset pricing using the EKOP methodology. Some of the papers are "on the conservative side" - their results would be strengthened if the probability of informed trading could be estimated without a bias. In other cases, however, the empirical results may be (at least in part) be due to biased estimation.

What do our results imply for the use of the EKOP methodology? When trades can be classified without error (as it is often the case in electronic auction markets) the method can be applied without any problem. When trade classification is prone to error, however, care should be taken. Obvious pitfalls lurk when one aims at comparing trading venues or stocks with different trading intensity and / or misclassification rates. This is a quite frequent design, and imprudent application of the EKOP model might call into question many an empirical result.

References

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

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

Easley, D. and M. O'Hara. "Time and the Process of Security Price Adjustment." *Journal of Finance*, 47 (1992), 577-606.

Easley, D., M. O'Hara and J. Paperman. "Financial Analysts and Information-Based Trade." *Journal of Financial Markets*, 1 (1998), 175-201.

Easley, D., N. Kiefer and M. O'Hara. "Cream-Skimming or Profit-Sharing?, The Curious Role of Purchased Order Flow." *Journal of Finance*, 51 (1996), 811-833.

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

Easley, D., M. O'Hara and G. Saar. "How Stock Splits Affect Trading: A Microstructure Approach." *Journal of Financial and Quantitative Analysis*, 36 (2001), 25-51.

Ellis, K., R. Michaely and M. O'Hara. "The Accuracy of Trade Classification Rules: Evidence from Nasdaq." *Journal of Financial and Quantitative Analysis*, 35 (2000), 529-551.

Finucane, Th. "A Direct Test of Methods for Inferring Trade Direction from Intra-Day data." *Journal of Financial and Quantitative Analysis*, 35 (2000), 553-576.

Grammig, J., D. Schiereck and E. Theissen. "Knowing Me, Knowing You: Trader Anonymity and Informed Trading in Parallel Markets." *Journal of Financial Markets*, 4 (2001), 385-412.

Lee, Ch. and B. Radhakrishna. "Inferring Investor Behavior: Evidence from TORQ." *Journal of Financial Market*, 3 (2000), 83-111.

Lee, C. and M. Ready. "Inferring Trade Direction from Intraday Data." *Journal of Finance*, 46 (1991), 733-746.

Odders-White, E. "On the Occurrence and Consequences of Inaccurate Trade Classification." *Journal of Financial Markets*, 3 (2000), 259-286.

Savickas, R. and A. Wilson. "On Inferring the Direction of Option Trades." Working Paper, George Washington University, 2001.

Silverman, B. "Kernel density estmation for statistics and data analysis." Chapman & Hall, London, 1986.

Theissen, E. "A Test of the Accuracy of the Lee/Ready Trade Classification Algorithm." *Journal of International Financial Markets, Institutions and Money*, 11 (2001), 147-165.

Table I: Simulation results

s	α [%]	ε	δ [%]	μ	PI [%]	bias (PI) q=0.9	bias (PI) q=0.85	RMSE (PI) q=0.9	RMSE (PI) q=0.85
1	**50.0294**	**0.17574**	**34.908**	**0.131970**	**15.81**	**-2.26**	**-3.36**	**2.78**	**3.67**
2	**43.3952**	**0.02397**	**44.439**	**0.030148**	**21.44**	**-5.65**	**-7.42**	**6.37**	**7.90**
3	**35.6320**	**0.00961**	**50.179**	**0.015696**	**22.53**	**-9.94**	**-10.85**	**11.19**	**11.97**
4	20	0.02	50	0.02	9.09	-3.16	-3.16	4.55	4.66
5	30	0.02	50	0.02	13.04	-4.08	-5.31	5.79	6.58
6	40	0.02	50	0.02	16.67	-5.47	-6.85	6.59	7.63
7	50	0.02	50	0.02	20.00	-6.59	-8.35	7.63	9.11
8	20	0.05	50	0.02	3.85	0.31	0.50	2.57	2.51
9	30	0.05	50	0.02	5.66	-0.50	-1.19	2.68	2.69
10	40	0.05	50	0.02	7.41	-1.43	-1.81	3.51	3.48
11	50	0.05	50	0.02	9.09	-2.16	-3.20	3.93	4.31
12	20	0.05	50	0.20	28.57	-4.00	-5.67	5.86	6.98
13	30	0.05	50	0.20	37.50	-6.10	-8.21	7.38	9.07
14	40	0.05	50	0.20	44.44	-6.72	-9.54	7.54	10.00
15	50	0.05	50	0.20	50.00	-8.33	-11.54	8.78	11.78
16	20	0.02	50	0.03	11.11	-3.06	-4.13	4.26	4.88
17	30	0.02	50	0.03	15.79	-4.35	-5.69	5.88	6.58
18	40	0.02	50	0.03	20.00	-5.44	-7.22	6.36	7.98
19	50	0.02	50	0.03	23.81	-7.01	-8.97	7.74	9.54
20	20	0.05	50	0.03	4.76	-0.66	-0.49	2.36	2.43
21	30	0.05	50	0.03	6.98	-1.56	-2.36	3.01	3.28
22	40	0.05	50	0.03	9.09	-2.41	-3.11	3.51	3.95
23	50	0.05	50	0.03	11.11	-3.15	-3.95	4.15	4.78
24	20	0.02	50	0.05	20.00	-4.64	-5.75	5.81	6.64
25	30	0.02	50	0.05	27.27	-6.00	-7.58	6.97	8.29
26	40	0.02	50	0.05	33.33	-7.26	-9.51	7.89	9.96
27	50	0.02	50	0.05	38.46	-8.39	-11.28	8.88	11.59
28	20	0.20	50	0.05	2.44	0.17	-0.19	1.46	1.08
29	30	0.20	50	0.05	3.61	-0.69	-0.94	1.58	1.65
30	40	0.20	50	0.05	4.76	-0.81	-1.30	1.71	1.98
31	50	0.20	50	0.05	5.88	-1.03	-1.66	1.92	2.22
32	20	0.03	50	0.05	16.67	-4.53	-5.36	5.59	6.18
33	30	0.03	50	0.05	23.08	-4.31	-5.74	5.37	6.46
34	40	0.03	50	0.05	28.57	-6.32	-8.40	6.95	8.80
35	50	0.03	50	0.05	33.33	-7.20	-9.76	7.73	10.13
36	20	0.20	50	0.20	9.09	-1.19	-1.71	2.08	2.34
37	30	0.20	50	0.20	13.04	-1.40	-2.19	2.41	2.84
38	40	0.20	50	0.20	16.67	-2.49	-3.53	3.15	3.94
39	50	0.20	50	0.20	20.00	-2.67	-4.04	3.12	4.29
40	20	0.03	50	0.03	9.09	-2.23	-2.92	3.55	4.35
41	30	0.03	50	0.03	13.04	-3.71	-4.37	4.83	5.36
42	40	0.03	50	0.03	16.67	-5.02	-6.24	5.98	7.05
43	50	0.03	50	0.03	20.00	-5.43	-7.04	6.44	7.80
44	20	0.05	50	0.05	9.09	-1.33	-2.51	2.59	3.24
45	30	0.05	50	0.05	13.04	-2.49	-3.39	3.33	4.00
46	40	0.05	50	0.05	16.67	-3.29	-4.61	3.93	5.04
47	50	0.05	50	0.05	20.00	-3.91	-5.52	4.63	5.99

The table reports bias and root mean squared error (RMSE) of the estimated probability of informed trade *PI* when the trade misclassification rate 1-q is equal to 10% and 15 %. In the first three rows, the model parameters α, ε, δ and μ are chosen to match the estimates in Easley et al. (1996) who have sorted NYSE stocks into trade volume deciles and estimated the model parameters by ML. The first row parameters are the mean of the parameter estimates for the first trade volume decile, the second row takes the mean estimates of the fifth decile and the third row contains the mean parameter estimates of the eighth decile. In the remaining rows the model parameters α, ε and μ are varied over a range of values. The implied probability of informed trade is calculated as $PI = \alpha\mu(\alpha\mu + 2\varepsilon)^{-1}$. Bias and RMSE are computed as: $Bias(PI) = K^{-1}\sum_{k=1}^{K}\hat{PI}_k - PI$ and $RMSE(PI) = \left(K^{-1}\sum_{k=1}^{K}\left(\hat{PI}_k - PI\right)^2\right)^{0.5}$ where $\hat{PI}_k$ denotes the estimated probability of informed trade calculated using the ML estimates based on misclassified data obtained in replication *k*, $\hat{PI} = \hat{\alpha}_k\hat{\mu}_k\left(\hat{\alpha}_k\hat{\mu}_k + 2\hat{\varepsilon}_k\right)^{-1}$. The number of replications *K* within each simulation design is equal to 100. To simulate the number of buyer and seller initiated trades, the length of the trading day is set to *T*=6.5 and the number of trading days is set to *M*=60.

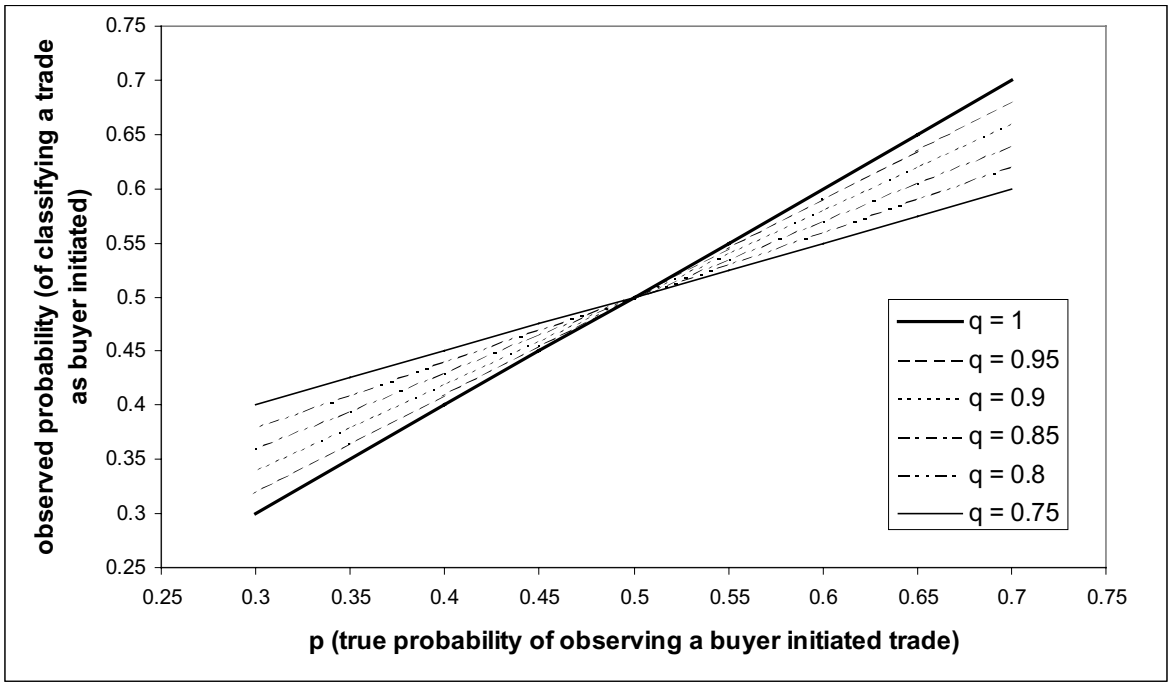

Figure 1: Classification bias and misclassification rate

The graph shows the relation between the true probability of a buyer-initiated trade and the observed probability of a buyer-initiated trade for different values of the parameter q that measures classification accuracy. The relation between the true probability of a buyer-initiated trade, p, and the observed probability of a buyer-initiated trade, $\Pr(\tilde{y}_t = 1)$, is

$$\Pr(\tilde{y}_t = 1) = (1 - p - q + 2pq).$$

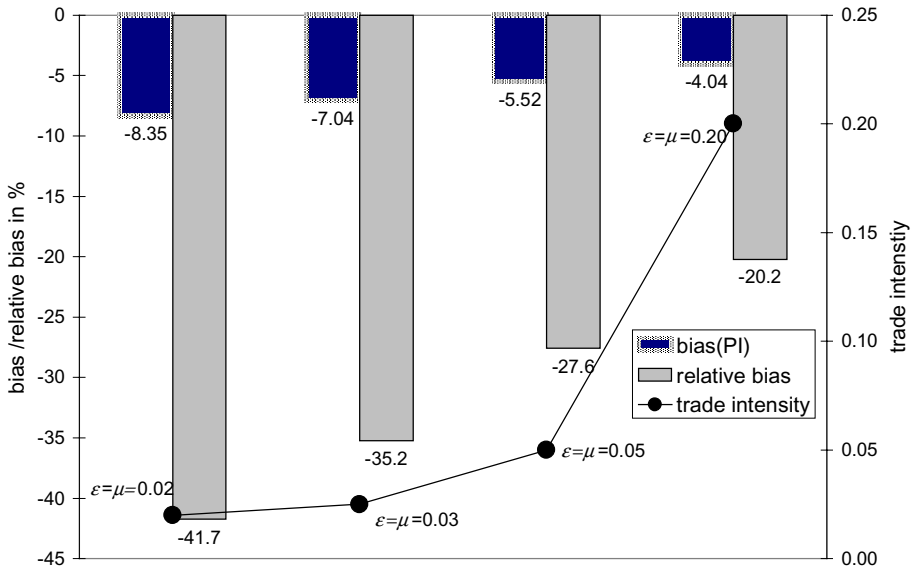

Figure 2: Bias and trade intensity

The graph depicts how the bias of the estimate of the probability of informed trading that is induced by a trade misclassification rate of 0.15 is changing with varying trade intensities. The relative bias is computed by dividing the bias by the true probability of informed trade. Four simulation designs from table 1 are selected which have the identical composition of the trader population $\varepsilon/\mu=1$ and the same news probabilities $\alpha=0.5$. The implied probability of informed trade is thus equal to 0.2 in the four designs. The simulation designs differ in their trading intensities. The figure displays from left to right and with the numbering of table 1: design 7 with smallest transaction intensity ($\varepsilon=\mu=0.02$), design 43 ($\varepsilon=\mu=0.03$), design 47 ($\varepsilon=\mu=0.05$), and design 39 which assumes the highest trade intensity ($\varepsilon=\mu=0.20$). The simulation procedure is as described in the caption of table 1.

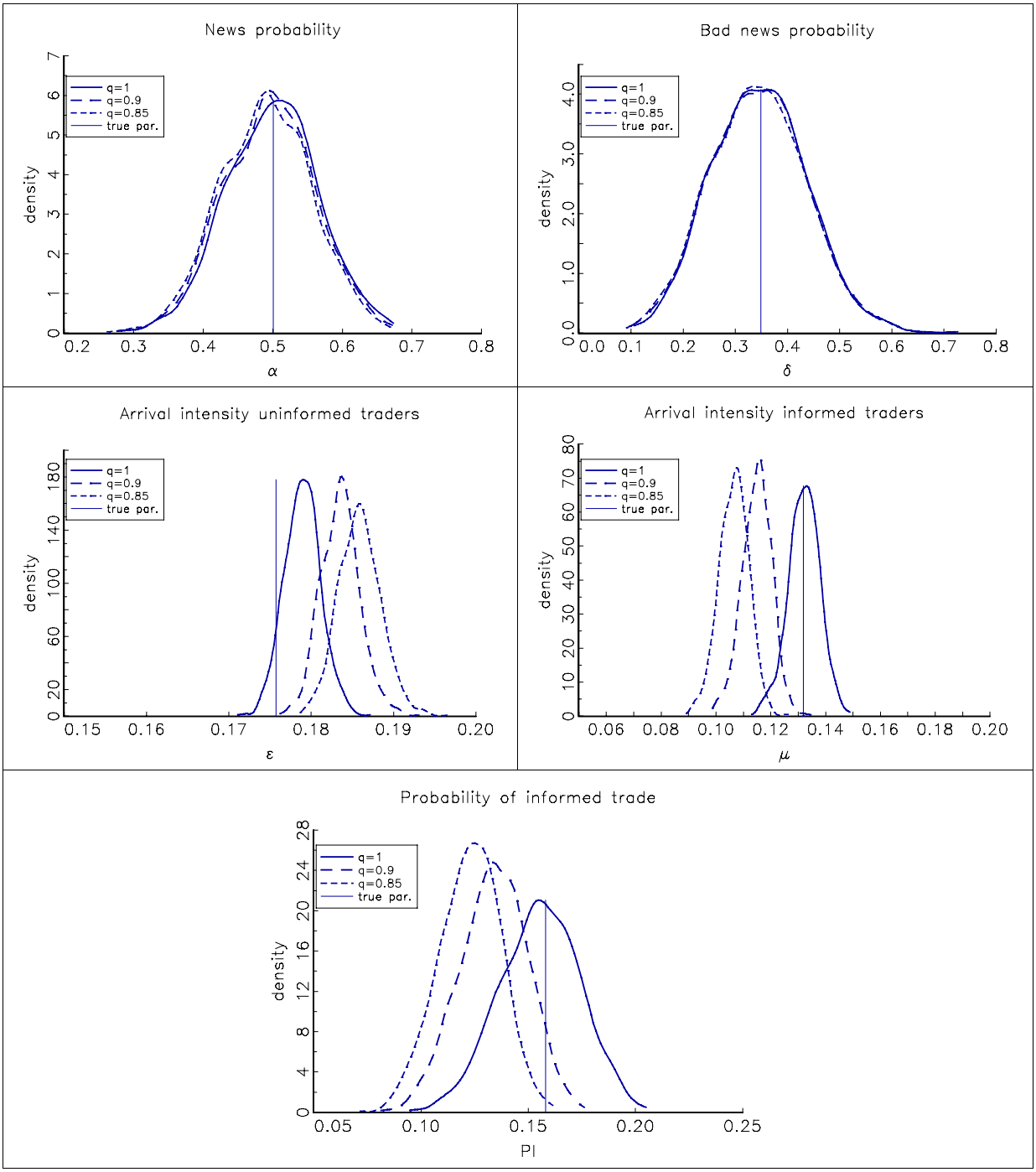

Figure 3: Illustration and decomposition of the bias

The graph depicts kernel density estimates of the distribution of the ML estimates based on correctly classified and misclassified simulated data. The true parameters are taken from the first row of table 1, i.e. they represent the mean estimates of the first volume decile of NYSE stocks in Easley et al (1996). The graph shows the distribution of the parameter estimates based on simulated data with correct trade classification q=1 and with misclassification rate 1-q equal to 0.1 and 0.15, respectively. In each replication a simulated sample is generated. Using this sample the model parameters are estimated by Maximum Likelihood. Then the sample is reshuffled allowing for misclassification of each trade with probability 1-q. ML estimation is repeated based on the modified data. To generate smooth densities, the procedure is repeated K=1000 times. As above, M=60 and T=6.5. Kernel density estimation is based on the sample of 1000 estimated parameter sets. The Gaussian kernel is employed with Silverman's (1986, p. 48) rule to select the optimal bandwidth.