

# Modeling Transaction Data of Trade Direction and Estimation of Probability of Informed Trading

**Anthony Tay**  
January 2007

Paper No. 13-2007

# Modeling Transaction Data of Trade Direction and Estimation of Probability of Informed Trading

Anthony Tay

School of Economics and Social Sciences,  
Singapore Management University

Christopher Ting

Lee Kong Chian School of Business,  
Singapore Management University

Yiu Kuen Tse\*

School of Economics and Social Sciences,  
Singapore Management University

Mitch Warachka

Lee Kong Chian School of Business,  
Singapore Management University

January 2007

## Abstract

This paper implements the Asymmetric Autoregressive Conditional Duration (AACD) model of Bauwens and Giot (2003) to analyze irregularly spaced transaction data of trade direction, namely buy versus sell orders. We examine the influence of lagged transaction duration, lagged volume and lagged trade direction on transaction duration and direction. Our results are applied to estimate the probability of informed trading (PIN) based on the Easley, Hvidkjaer and O'Hara (2002) framework. Unlike the Easley- Hvidkjaer-O'Hara model, which uses the daily aggregate number of buy and sell orders, the AACD model makes full use of transaction data and allows for interactions between buy and sell orders.

JEL Codes: C410, G120

Keywords: Autoregressive Conditional Duration, Market Microstructure, Probability of Informed Trading, Transaction Data, Weibull Distribution

---

\*Corresponding author, 90 Stamford Road, Singapore Management University, Singapore 178903, Email: yktse@smu.edu.sg, Phone: (65) 6828 0257, Fax: (65) 6828 0833. This paper, and its companion paper "The Impact of Transaction Duration, Volume and Direction on Price Dynamics and Volatility", supersedes a previously circulated paper titled "An Autoregressive Conditional Marked Duration Model for Transaction data", which was presented at the First Symposium on Econometric Theory and Application (SETA), Taipei, 2005 and the Econometric Society World Congress, London, 2005. We thank all who have commented on these papers.

# Modeling Transaction Data of Trade Direction and Estimation of Probability of Informed Trading

January 2007

## **Abstract**

This paper implements the Asymmetric Autoregressive Conditional Duration (AACD) model of Bauwens and Giot (2003) to analyze irregularly spaced transaction data of trade direction, namely buy versus sell orders. We examine the influence of lagged transaction duration, lagged volume and lagged trade direction on transaction duration and direction. Our results are applied to estimate the probability of informed trading (PIN) based on the Easley, Hvidkjaer and O'Hara (2002) framework. Unlike the Easley-Hvidkjaer-O'Hara model, which uses the daily aggregate number of buy and sell orders, the AACD model makes full use of transaction data and allows for interactions between buy and sell orders.

# 1 Introduction

In this paper we implement the Asymmetric Autoregressive Conditional Duration (AACD) model of Bauwens and Giot (2003) to analyze the trade direction of stock transactions (i.e., buy- versus sell-initiated stock trading) and the duration of these transaction *jointly*. We apply the model to examine the intensity of informed versus uninformed trading and propose a method to estimate the probability of informed trading (PIN) under the Easley, Hvidkjaer and O'Hara (2002) (EHO hereafter) framework. Unlike the EHO method, however, which uses the daily aggregate number of buy and sell orders to estimate PIN, our model enables us to estimate PIN using irregularly spaced transaction data.

Easley, Kiefer and O'Hara (1996, 1997) develop the methodology of PIN and utilize it to investigate the role of purchase order flows. Easley, Kiefer, O'Hara and Paperman (1996) further extend the method to investigate stocks that trade infrequently. Since then the PIN measure has been widely used in the empirical finance literature. For example, Easley, O'Hara and Saar (2001) apply PIN to study the impact of stock splits on uninformed traders. More recently, Aslan, Easley, Hvidkjaer and O'Hara (2006) use PIN to examine the link between microstructure and asset pricing. Henry (2006) employs the PIN measure to study the interaction between short selling and informed trading, while Benos and Jochev (2006) confirm that PIN captures informed trading around earnings announcements. Chung and Li (2003) utilize PIN to verify the appropriateness of decomposing bid-ask spreads into adverse-selection and transitory components.

The AACD model is an extension of the Autoregressive Conditional Duration (ACD) model introduced by Engle and Russell (1998) and Engle (2000). Recently, the literature on the ACD model has expanded quickly, with contributions by Bauwens and Veredas (2004), Fernandes and Grammig (2005), Ghysels, Gouriéroux and Jasiak (2004), Grammig and Maurer (2000), and Zhang, Russell and Tsay (2001), among others. For a comprehensive survey of ACD models, see Pacurar (2006).

The ACD model analyzes the duration between two transactions, irrespective of the nature of the transaction (such as an increase or decrease in the traded price, or a trade initiated by a buy versus and sell order). Bauwens and Giot (2003) extend the ACD model

to study the mid price of bid-ask quotes. They propose a two-state AACD model to analyze mid-price decrease and mid-price increase jointly with the trade duration. In their model the conditional expected duration of each state varies differently with the conditional information, which includes the lagged duration, the lagged volume and the lagged spread.

In this paper we apply the AACD approach to a two-state model of transaction data, where the two states represent a transaction initiated by a buy versus a sell order, called the transaction direction. Following Bauwens and Giot (2003), we allow the expected duration to vary with some covariates, including the lagged duration, the lagged expected duration, the lagged direction of trade and the lagged trade volume. We construct AACD equations that reflect the changes in the trading activities due to informed traders. The AACD model is then used to estimate PIN. Our new method makes full use of transaction data and relaxes several assumptions in the EHO model.

The balance of the paper is as follows. In the next section, we summarize the AACD model as applied to trade direction and trade duration data. We outline the model as one based on competing risks, in which the underlying competing latent processes have inter-arrival times that are distributed as two-parameter Weibull random variables. In Section 3 we discuss the application of the AACD model under the EHO framework, which assumes trading intensities vary in trading days with no news, good news and bad news. Section 4 describes the data, and Section 5 reports our empirical results. Our conclusions are summarized in Section 6.

## 2 The AACD Model of Trade Direction

We consider a two-state AACD model of trade direction. Let  $w_i$  denote the trade direction of the  $i$ th trade at time  $t_i$ , which may take values of  $j = -1$  or  $1$  representing a sell-initiated and buy-initiated trade, respectively. We denote  $\Phi_{i-1}$  as the information set after the  $(i-1)$ th trade.  $\Phi_{i-1}$  may consist of past trade directions, volumes of transaction and lagged durations. Given  $\Phi_{i-1}$  we assume each of the two *potential* trade directions of the trade at time  $t_i$  follows a latent stochastic point process whose inter-arrival times have independent Weibull distributions. The realized (observed) trade direction is the outcome

of a competition between the two underlying point processes to be the first arrival.

Specifically, we denote the random duration for the next potential trade direction by  $T_{ji}$  for  $j = -1$  and  $1$  (the subscript  $i$  refers to the  $i$ th trade). Conditional on  $\Phi_{i-1}$ ,  $T_{ji}$  are assumed to be independently distributed as two-parameter Weibull random variables with scale parameter  $\psi_{ji}$  and shape parameter  $\phi_j$ . Both the shape and scale parameters are assumed to be positive. In particular, we allow the shape parameter  $\phi_j$ , which are time invariant, to vary with the latent process of trade direction  $j$ . On the other hand, the scale parameter  $\psi_{ji}$  of each latent process changes after each transaction. Thus, conditional on the information set  $\Phi_{i-1}$  after the  $(i-1)$ th trade, the inter-arrival time random variables of the latent processes have the following density function

$$f_{T_{ji}}(t) = \frac{\phi_j}{\psi_{ji}} \left( \frac{t}{\psi_{ji}} \right)^{\phi_j-1} \exp \left[ - \left( \frac{t}{\psi_{ji}} \right)^{\phi_j} \right], \quad j = -1, 1, \quad (1)$$

and survival function (complement of the distribution function, i.e.,  $1 - F_{T_{ji}}(t)$ , where  $F_{T_{ji}}(t)$  is the distribution function)

$$S_{T_{ji}}(t) = \exp \left[ - \left( \frac{t}{\psi_{ji}} \right)^{\phi_j} \right], \quad j = -1, 1. \quad (2)$$

Furthermore, the expected values of the inter-arrival time variables are

$$E(T_{ji}) = \psi_{ji} \Gamma \left( \frac{1}{\phi_j} + 1 \right), \quad (3)$$

where  $\Gamma(\cdot)$  is the gamma function. Thus, the expected duration of the inter-arrival time is proportional to the scale parameter.

Denoting the duration of the  $i$ th trade (i.e., the waiting time from time  $t_{i-1}$  of the  $(i-1)$ th trade to time  $t_i$  of the  $i$ th trade) by  $x_i = t_i - t_{i-1}$  and the direction of the  $i$ th trade by  $w_i$  (note that  $w_i$  is either  $-1$  or  $1$ ), the conditional joint density of  $(w_i, t_i)$  (or equivalently  $(w_i, x_i)$ ), denoted by  $p_i(w_i, t_i | \Phi_{i-1})$ , is

$$\begin{aligned} p_i(k, t_i | \Phi_{i-1}) &= \Pr \left( \bigcap_{j=-1,1} \{T_{ji} > x_i\} \right) f_{T_{ki}}(x_i | T_{hi} > x_i; h = -1, 1) \\ &= \left( \prod_{j=-1,1} S_{T_{ji}}(x_i) \right) \frac{f_{T_{ki}}(x_i)}{S_{T_{ki}}(x_i)} \\ &= \left( \prod_{j=-1,1} S_{T_{ji}}(x_i) \right) \frac{\phi_k}{\psi_{ki}} \left( \frac{x_i}{\psi_{ki}} \right)^{\phi_k-1}, \quad k = -1, 1. \end{aligned} \quad (4)$$

Hence, if we define  $D_h(j) = 1$  for  $h = j$ , and 0 otherwise, the conditional joint density of  $(w_i, t_i)$  is

$$p_i(k, t_i | \Phi_{i-1}) = \prod_{j=-1,1} \left[ \frac{\phi_j}{\psi_{ji}} \left( \frac{x_i}{\psi_{ji}} \right)^{\phi_j - 1} \right]^{D_k(j)} \exp \left[ - \left( \frac{x_i}{\psi_{ji}} \right)^{\phi_j} \right], \quad k = -1, 1. \quad (5)$$

Given a sample of observations  $\{w_i, t_i\}$  for  $i = 1, \dots, N$ , the log-likelihood function is

$$\sum_{i=1}^N \ln p_i(w_i, t_i | \Phi_{i-1}) = - \sum_{i=1}^N \left( \sum_{j=-1,1} \left( \frac{x_i}{\psi_{ji}} \right)^{\phi_j} - \sum_{j=-1,1} D_{w_i}(j) \ln \left[ \frac{\phi_j}{\psi_{ji}} \left( \frac{x_i}{\psi_{ji}} \right)^{\phi_j - 1} \right] \right). \quad (6)$$

Thus, the parameters of the model can be estimated using maximum likelihood estimation (MLE) method once the functional forms of the scale parameters  $\psi_{ji}$  are specified. As the conditional expected duration is proportional to the scale parameter, we apply the ACD model to the scale parameter. For example, the conditional scale parameter may be updated according to the following equation

$$\ln \psi_{ji} = \nu_{j,-1} D_{-1}(w_{i-1}) + \nu_{j1} D_1(w_{i-1}) + \alpha_j \ln \psi_{j,i-1} + \beta_j \ln x_{i-1}, \quad j = -1, 1. \quad (7)$$

In equation (7), we have an extended ACD(1,1) structure, where the constant term in the usual ACD equation is replaced by the intercepts  $\nu_{j,-1}$  and  $\nu_{j1}$  that vary according to the previous trade direction  $w_{i-1}$ . As seen in equation (3), an increase (decrease) in  $\psi_{ji}$  implies a larger (smaller) expected duration, which in turn implies a reduced (increased) probability that the transaction at time  $t_i$  is of type  $j$ . The intercepts  $\nu_{jk}$  represent the sensitivity of the next trade direction  $j$  to the prior trade direction  $k$ . Thus, if the previous trade direction is of type  $k$ , the intercept for  $\ln \psi_{ji}$  is  $\nu_{jk}$ . A larger (smaller)  $\nu_{jk}$  implies that trade direction  $k$  induces a larger (smaller) expected duration of the next trade direction being of type  $j$ . The remaining sets of  $\alpha$  and  $\beta$  coefficients are those of an ACD(1,1) model. If  $\beta_{-1}$  is smaller than  $\beta_1$ , a longer lagged duration implies a lower conditional expected duration of a sell-initiated trade than a buy-initiated trade.

If the shape parameters of the two latent competing processes of trade directions are equal, i.e.,  $\phi_{-1} = \phi_1 = \phi$ , say, equation (5) can be simplified as follows

$$p_i(k, t_i | \Phi_{i-1}) = \frac{\phi}{\psi_{ki}} \left( \frac{x_i}{\psi_{ki}} \right)^{\phi - 1} \exp \left[ - \sum_{j=-1,1} \left( \frac{x_i}{\psi_{ji}} \right)^{\phi} \right], \quad k = -1, 1. \quad (8)$$

In this special case it is straightforward to obtain the marginal densities of  $w_i$  and  $x_i$ . Thus, if we define  $\psi_i$  by

$$\psi_i = \left( \frac{1}{\psi_{-1,i}^\phi} + \frac{1}{\psi_{1,i}^\phi} \right)^{-\frac{1}{\phi}}, \quad (9)$$

the conditional marginal density of  $x_i$  is

$$f_{x_i}(x|\Phi_{i-1}) = \frac{\phi x^{\phi-1}}{\psi_i^\phi} \exp \left[ - \left( \frac{x}{\psi_i} \right)^\phi \right]. \quad (10)$$

Also, the conditional marginal density of  $w_i$  is

$$f_{w_i}(k|\Phi_{i-1}) = \left( \frac{\psi_i}{\psi_{ki}} \right)^\phi, \quad k = -1, 1. \quad (11)$$

Hence, conditional on  $\Phi_{i-1}$ ,  $x_i$  has a two-parameter Weibull distribution with shape parameter  $\phi$  and scale parameter  $\psi_i$ . Likewise, conditional on  $\Phi_{i-1}$ ,  $x_i$  has a multinomial distribution with probabilities proportional to  $1/\psi_{ki}^\phi$  for  $k = -1, 1$ .

Note that the product of the expressions in (10) and (11) is equal to that in (8). Thus, under the special case when the shape parameters of the latent processes are equal, the trade direction  $w_i$  and the trade duration  $x_i$  are independent conditional upon the information  $\Phi_{i-1}$ . As shown in Section 4, the estimated shape parameters of the latent processes are very close for all the data sets considered. Furthermore, the estimates of the other parameters in the model are quite similar whether the equal-shape assumption is imposed or not. Thus, imposing equal shape is theoretically convenient and appears to be empirically robust. In an extensive study to investigate various ACD models, Bauwens, Giot, Grammig and Veredas (2004) compare different distribution assumptions, including the Poisson, Weibull, Burr, generalized Gamma, threshold and stochastic conditional duration models. They conclude that “simpler approaches perform at least as well as more complex methods”. In this paper we adopt the assumption of Weibull distributions with equal shape parameters for its theoretical convenience without compromising its empirical performance.

### 3 Estimation of PIN

EHO analyze trade-direction data to estimate the proportion of trades initiated by informed traders. Their model is based on the numbers of buy and sell orders in each day, the



intensities of which depend on the existence of “news” or information. Conditional on the arrival of news, information is further classified as being either “good” or “bad”. EHO model the aggregate numbers of buy- and sell-initiated trades in each day as independent Poisson variables, with different intensities for days with no news, good news and bad news. The characterization of each trading day is unknown, and its likelihood is based on the mixture-of-Poisson distribution. The probability of informed trading (PIN) is then calculated as the ratio of the combined intensity for buy- and sell-initiated trades for days with news divided by the total intensity over all days.

In the EHO framework each trading day is characterized by good news ( $G$ ), no news ( $N$ ) and bad news ( $B$ ) to form the set  $S = \{G, N, B\}$ . We denote  $\pi_s$  as the probability of state  $s$  in  $S$ . Let the probability of a day containing news be  $\theta_E$ . Furthermore, conditional on the arrival of news, denote the probability of bad news as  $\theta_B$ . Thus, the probability of a no-news day is  $\pi_N = 1 - \theta_E$ , and the probabilities of good- and bad-news days are  $\pi_G = \theta_E (1 - \theta_B)$  and  $\pi_B = \theta_E \theta_B$ , respectively. EHO assume the aggregate numbers of buy and sell orders in a trading day follow independent Poisson distributions, where the intensities of sell and buy orders on a no-news day, denoted by  $\lambda_{-1}$  and  $\lambda_1$ , respectively, are constant throughout the sample period. On a good-news day, the buy intensity increases by a positive amount  $\delta$ , with no change in the sell intensity. Likewise, on a bad-news day, the sell intensity increases by  $\delta$  while the buy intensity remains unchanged. Suppose there are  $D$  days of data. Under the mixture-of-distributions assumption, the likelihood function of the model is given by

$$\prod_{d=1}^D \left\{ (1 - \theta_E) \frac{\lambda_1^{B_d} e^{-\lambda_1}}{B_d!} \frac{\lambda_{-1}^{S_d} e^{-\lambda_{-1}}}{S_d!} + \theta_E \theta_B \frac{\lambda_1^{B_d} e^{-\lambda_1}}{B_d!} \frac{(\lambda_{-1} + \delta)^{S_d} e^{-(\lambda_{-1} + \delta)}}{S_d!} \right. \\ \left. + \theta_E (1 - \theta_B) \frac{(\lambda_1 + \delta)^{B_d} e^{-(\lambda_1 + \delta)}}{B_d!} \frac{\lambda_{-1}^{S_d} e^{-\lambda_{-1}}}{S_d!} \right\}, \quad (12)$$

where  $B_d$  and  $S_d$  are the respective aggregate number of buy and sell orders on day  $d$ . From this model, EHO’s suggested estimate of PIN is calculated as

$$\text{PIN} = \frac{\theta_E \delta}{\theta_E \delta + \lambda_{-1} + \lambda_1}. \quad (13)$$

We propose to model the buy- and sell-initiated orders by applying the AACD approach to a two-mark process describing the buy- and sell-initiated transactions. Our model allows the parameters to vary according to whether there is news for the trading day. We refer to

this model as the PIN-AACD model, as opposed to the PIN-EHO model of Easley, Hvidkjaer and O’Hara (2002). Our PIN-AACD implementation relaxes several PIN-EHO assumptions. Specifically, the following extensions are made: (i) volume is incorporated, (ii) independence between the number of buy and sell orders each day is relaxed, (iii) trade orders are allowed to be serially correlated, and (iv) transaction duration and trade direction are modeled jointly.

The PIN-AACD model uses transaction data of trade directions and does not just model the daily aggregates of buy and sell orders. Summarizing the cumulative trade flow in daily aggregates may result in a loss of information. Modeling transaction data, the PIN-AACD model accounts for autocorrelation in trade directions and can incorporate trade volume into the estimation of PIN.<sup>1</sup> Another potential limitation of the PIN-EHO approach is the assumption that the number of daily buy and sell orders are independent. In contrast, the PIN-AACD model allows the interactions between consecutive buy and sell orders to be updated after each transaction. This implies the conditional durations, and thus the associated intensities, of buy and sell orders influence one another.

We modify equation (7) to describe the evolution of the conditional scale parameters in trading days with or without news. First we denote  $\psi_{ji}^s$  as the conditional scale parameter of trade direction  $j$  in state  $s \in S$  given information  $\Phi_{i-1}$  after the trade at time  $t_{i-1}$ , where the specification of  $\psi_{ji}^s$  must reflect the activities of informed and uninformed traders. We then define the following function  $f_{ji}^s$ , which forms the basis of the equations for the conditional scale parameter in each of the three states in  $S$ ,

$$f_{ji}^s \equiv \nu_{j,-1} D_{-1}(w_{i-1}) + \nu_{j,1} D_1(w_{i-1}) + \alpha_j \ln \psi_{j,i-1}^s + \beta_j \ln x_{i-1}, \quad j = -1, 1. \quad (14)$$

Thus, the basis  $f_{ji}^s$  depends on whether the previous transaction is a buy- or sell-initiated order, as well as the lagged duration and previous conditional scale parameter of the order mark.

According to the assumptions of EHO, only uninformed traders are active in the absence of any news. When there is good news, informed traders purchase shares, increasing the trading intensity of buy orders. Conversely, when there is bad news, informed traders sell

---

<sup>1</sup>Easley, Hvidkjaer and O’Hara (2002) report that informed trading estimates are negatively correlated with volume, which remains a determinant of asset prices, but they omit this variable for computational tractability.

shares and increase the trading intensity of sell orders. However, the trading intensity of sell orders on a good-news day and the trading intensity of buy orders on a bad-news day are identical to their counterparts on a no-news day.

Thus, for a no-news day ( $s = N$ ), we assume a simple functional form for the logarithmic conditional scale parameter, with

$$\ln \psi_{ji}^N = f_{ji}^N, \quad j = -1, 1. \quad (15)$$

For the buy-orders ( $j = 1$ ) on a good-news day ( $s = G$ ) we reduce  $f_{1i}^N$  by a positive constant  $\mu$  to yield the following logarithmic conditional scale parameter

$$\ln \psi_{1i}^G = f_{1i}^G - \mu, \quad (16)$$

while the logarithmic conditional scale parameter for sell trade is the same as the basis function  $f_{-1,i}^G$ , i.e.,

$$\ln \psi_{-1,i}^G = f_{-1,i}^G. \quad (17)$$

On the other hand, for  $s = B$ , we have

$$\ln \psi_{1i}^B = f_{1i}^B, \quad (18)$$

and

$$\ln \psi_{-1,i}^B = f_{-1,i}^B - \mu. \quad (19)$$

According to equation (18),  $\psi_{1i}^B$  represents the conditional scale parameter of a buy order on a bad-news day, which is based on the benchmark function  $f_{1i}^B$  without adjustment. However, the logarithmic conditional scale parameter of a sell order  $\ln \psi_{-1,i}^B$  on a bad-news day decreases by  $\mu$  due to selling by informed traders. As seen in equation (16), on a good-news day, the logarithmic conditional scale parameter of a buy order decreases by  $\mu$  due to the buying of informed traders. However, the logarithmic conditional scale parameter of sell orders on good-news days,  $\ln \psi_{-1,i}^G$ , remain unchanged versus that of a no-news day, as seen in equation (17).

Given that a certain day is of type  $s$ , the joint density of  $\{w_i, t_i\}$  conditional on the information set  $\Phi_{i-1}$  is given in equation (5), which is rewritten below to incorporate the variations with respect to the state of the news:

$$p_{si}(k, t_i | \Phi_{i-1}) = \prod_{j=-1,1} \left[ \frac{\phi_j}{\psi_{ji}^s} \left( \frac{x_i}{\psi_{ji}^s} \right)^{\phi_j-1} \right]^{D_k(j)} \exp \left[ - \left( \frac{x_i}{\psi_{ji}^s} \right)^{\phi_j} \right], \quad k = -1, 1; s \in S. \quad (20)$$

Let  $N_d = S_d + B_d$  denote the number of trades on day  $d$ . The likelihood function is then given by

$$\prod_{d=1}^D \left( \sum_{s \in S} \pi_s \left( \prod_{i=1}^{N_d} p_{si}(w_i, t_i | \Phi_{i-1}) \right) \right). \quad (21)$$

Note that the term in the inner brackets of equation (21) is the likelihood function for day  $d$ , given that day  $d$  is in state  $s$  (the index  $d$  for the  $\{w_i, t_i\}$  data is suppressed).

In the PIN-EHO model, the Poisson assumption is adopted so that the hazard rate is constant and is equal to the reciprocal of the expected duration, which is used as a measure of the *intensity*. Under the Weibull assumption, however, the hazard rate is not constant, but is a decreasing function of the duration when the shape parameter is less than 1, as is the case empirically in this study.<sup>2</sup> Hence, we propose to use the reciprocal of the expected conditional duration as a measure of the intensity. Under the special case the shape parameters of the two latent processes are equal, PIN is simple to calculate. Let  $\lambda_{ji}^s = 1/\psi_{ji}^s$ . If the shape parameters are equal, PIN can be calculated as follows

$$\text{PIN} = \frac{\sum_{d=1}^D \sum_{i=1}^{N_d} (\pi_G \lambda_{1i}^G + \pi_B \lambda_{-1,i}^B)}{\sum_{d=1}^D \sum_{i=1}^{N_d} (\lambda_{-1,i}^N + \lambda_{1i}^N + \pi_G \lambda_{1i}^G + \pi_B \lambda_{-1,i}^B)}, \quad (22)$$

where again the index  $d$  for the intensities is suppressed.

Finally, we extend equation (14) to incorporate the influence of volume on the PIN estimates. In particular,  $w_{i-1} \ln s_{i-1}$  may be added to equation (14) to construct an augmented PIN-AACD model whose conditional scale parameters are based on modifying  $f_{ji}^s$  defined by

$$f_{ji}^s \equiv \nu_{j,-1} D_{-1}(y_{i-1}) + \nu_{j1} D_1(y_{i-1}) + \alpha_j \ln \psi_{j,i-1}^s + \beta_j \ln x_{i-1} + \varsigma_j y_{i-1} \ln s_{i-1}, \quad j = -1, 1. \quad (23)$$

We expect  $\varsigma_{-1} > 0$  and  $\varsigma_1 < 0$ , implying a large buy order induces a shorter expected duration for a subsequent buy order but a longer expected duration for a sell order, with a large sell

---

<sup>2</sup>The hazard function of a two-parameter Weibull random variable  $Y$  with shape parameter  $\phi$  and scale parameter  $\psi$  is  $\phi y^{\phi-1}/\psi^\phi$ , which is a decreasing function of  $y$  when  $\phi < 1$ .

order having the opposite effect.<sup>3</sup> Once the conditional scale parameters are estimated, PIN measures are calculated using equation (22).

## 4 Data

We apply the AACD model to intraday data of five NYSE companies: Boeing (BA), General Electric (GE), International Business Machines (IBM), Altria Group (formerly Philip Morris) (MO), and AT&T (T). The data are obtained from the TAQ database for July 1, 1994 to June 30, 1995.

We extract three variables on each stock: time of trade, transaction price, and signed volume inferred using the Lee and Ready (1991) algorithm. We also correct for the opening auction and for time-of-day effects, using procedures similar to those in Engle and Russell (1998). In particular, opening effects require the transactions occurring in the first 20 minutes of each day to be removed. The average duration for transactions over the following 10 minutes serves as the waiting time for the first trade after 10:00 a.m. (E.S.T.). All transactions recorded after 4:00 p.m. are also deleted. In some cases, the opening transaction occurred after the first 20 minutes. Also, on a few days there are insufficient transactions between 9:50 a.m. and 10:00 a.m. to obtain a meaningful average starting duration. Therefore, days with opening transactions after 9:50 a.m. and with less than three transactions over the next 10 minutes are also removed, along with November 25, 1994 due to an early “day after Thanksgiving” closing. Even after these deletions, a tremendous number of observations for each company remain, as documented in Table 1.

We estimate diurnal factors by applying a smoothing spline to the average duration at each time point with available data.<sup>4</sup> The diurnally adjusted durations are then formed by dividing each duration with the corresponding diurnal factor. For the remainder of this

---

<sup>3</sup>We may further allow variations in  $\zeta$  with respect to the information environment by supplementing the index  $s$  to obtain  $\zeta_j^s$  for  $s \in S$ . However, this may lead to over-parametrization and is not pursued in this paper.

<sup>4</sup>We used the MATLAB function *csaps.m* to compute the smoothing spline. The diurnal factor is adjusted to ensure the sample mean of the diurnally-adjusted durations is equal to the sample mean of the non-diurnally-adjusted data.

paper, durations  $x_i$  refer to mean-diurnally-adjusted durations. The diurnal factors for all five duration series are similar to those in Engle and Russell (1998). In particular, the diurnal factors initially increase, with the largest diurnal factor occurring at the middle of the day, before decreasing.

Some summary statistics of the data are given in Table 1. The number of observations available for BA is substantially lower than the other stocks, primarily due to less frequent trading as indicated by its average duration. The average number of trades per day varies from a low of 243.3 (BA) to a high of 677.9 (GE). The runs tests indicate that there is positive serial correlation in the directions of the trade. GE and T have more than 50% of buy trades, while the other three stocks have more sell trades than buy trades.

## 5 Empirical Results

Results of the PIN-EHO model are summarized in Table 2. The PIN estimates vary from the lowest value of 0.0879 for MO to the highest value of 0.1342 for BA. Note that BA is the stock with the lowest average daily trade in the sample. If we measure the relative intensity of informed traders versus uninformed traders by  $2\hat{\delta}/(\hat{\lambda}_{-1} + \hat{\lambda}_1)$ , the relative intensity of BA (0.883) is the highest in the sample and the relative intensity for GE (0.437) is the lowest. Thus, although BA has a lower probability of news than that of GE, it has a higher PIN. On the other hand, though MO has a higher relative intensity (0.718) than that of GE, since its probability of informed trading is very low, it has the lowest PIN in the sample.

The results of the PIN-AACD model are presented in Table 3. For each stock we estimate the model with equal and unequal shape parameters. The results show that the estimated shape parameters for the sell and buy orders are quite close. Furthermore, when the equality of the two parameters is imposed, the estimates of the other parameters remain quite stable. Thus, for further analysis we adopt the model with equal shape parameters.<sup>5</sup> We report some diagnostics of the PIN-AACD model. The diagnostics are based on the marginal distributions of the transaction duration and the trade direction. As the transaction duration

---

<sup>5</sup>We report that formal tests of the equality of the parameters reject the null at conventional levels of significance, which is due to the large sample size.

is a continuous random variable we use the empirical probability integral transform, which was proposed by Diebold, Gunther and Tay (1998) to assess density forecasts and adopted by Bauwens, Giot, Grammig and Veredas (2004) to analyze transaction-duration models. The method involves plotting the empirical probability integral transform of the transaction duration, which is uniformly distributed under the correct specification. As the marginal distribution of transaction duration is given in equation (10) under the model with equal shape parameters, the probability integral transform of the transaction duration  $x_i$  of the PIN-ACCD model, denoted by  $\hat{F}(x_i)$ , is given by

$$\hat{F}(x_i) = \sum_{s \in \mathcal{S}} \hat{\pi}_s \left( 1 - \exp \left[ - \left( \frac{x_i}{\hat{\psi}_i^s} \right)^{\hat{\phi}} \right] \right), \quad i = 1, \dots, N, \quad (24)$$

where  $\hat{\psi}_i^s$  is  $\psi_i$  in state  $s$  defined in equation (9).

As the direction of trade is a discrete random variable, we use the Brier score based on the absolute difference between the observed value of the trade direction and the probability of the trade direction. Specifically, we denote the predicted probability of a buy-trade in the  $i$ th transaction by  $\hat{y}_i$ , which is given by (see equation (11))

$$\hat{y}_i = \sum_{s \in \mathcal{S}} \hat{\pi}_s \left( \frac{\hat{\psi}_i^s}{\hat{\psi}_{1i}^s} \right)^{\hat{\phi}}, \quad i = 1, \dots, N. \quad (25)$$

Then the Brier score based on the absolute difference is calculated as

$$\frac{1}{N} \sum_{i=1}^N |\hat{y}_i - z_i|, \quad (26)$$

where  $z_i$  is the indicator variable which takes value 1 when  $w_i = 1$  and zero otherwise. When trade-direction forecast is perfect (i.e.,  $\hat{y}_i$  take the values of 1 or 0 that coincide with  $z_i$ ), the Brier score attains its minimum value of zero. The worst performance is when the Brier score is 1.

Figure 1 plots the empirical density functions of the estimated probability integral transforms of the transaction durations  $\hat{F}(x_i)$  of the five stocks. It can be seen that the distributions do not behave like a uniform distribution when the duration is very short. Otherwise, the uniform distribution appears to describe the transformation well in other parts of the distribution. This finding is similar to those of Bauwens, Giot, Grammig and Veredas (2004). We also compute the autocorrelations of  $\hat{F}(x_i)$  up to 30 lags. The autocorrelations at all

lags are statistically significant, except the first order autocorrelation for BA, MO, and T. All autocorrelations are, nonetheless, extremely small (less than 0.09 in absolute value). The Brier scores are reported in the last row of Table 3. It can be seen that all results are less than 0.5. The best performance is for the IBM data, with a Brier score of 0.3657, and the worst performance is for the MO data, with a Brier score of 0.4353. Overall, the diagnostics appear to support the AACD model.

Estimates of the ACCD model for trade direction exhibits a remarkable resemblance across the five firms. In particular, we observe the following. Firstly,  $\hat{\nu}_{-1,-1} < \hat{\nu}_{-1,1}$  and  $\hat{\nu}_{11} < \hat{\nu}_{1,-1}$ , implying that buy trades induce lower conditional expected duration of buy trades than sell trades, and sell trades induce lower conditional expected duration of sell trades than buy trades. This is consistent with positive serial correlation in trade direction. Secondly,  $\hat{\zeta}_{-1} > 0$  and  $\hat{\zeta}_1 < 0$ , implying large buy orders induce shorter conditional expected durations for subsequent buy orders but longer conditional expected durations for sell orders. The opposite is true for large sell orders. Thus, volume plays an explicit role in predicting trade directions. The estimates of  $\alpha_{-1} + \beta_{-1}$  and  $\alpha_1 + \beta_1$  for all stocks are generally comparable and are not close to one, indicating persistence is moderate.

The shape parameters across the stocks are quite similar and they are all less than 1. The two-parameter Weibull distribution nests the exponential distribution as a special case, and our results suggest that assuming Poisson arrival for the buy and sell orders may result in misspecification. We observe big drops in the PIN in the GE and IBM data when the PIN-AACD model is used versus the PIN-EHO model. In these two cases, the estimates of the probability of news are much reduced in the AACD model, causing the PIN estimates to drop. On the other hand, the PIN-ACCD of the BA data increases versus the PIN-EHO. Indeed, for the BA data, the estimate of the adjustment for information  $\mu$  is the highest among all stocks, suggesting the relative intensity of informed traders are high when there is news. Overall, the probability of news appears to be quite even across the stocks, with MO and T being the stocks with least news.



## 6 Conclusions

This paper implements an AACD model to explore the impact of previous trade variables such as direction, volume, duration and their interactions on subsequent arrival times and trade direction. The two-parameter Weibull model nests the exponential distribution as a special cases, and the use of transaction data allows us to model the interaction between buy and sell orders. In particular, independence between the buy and sell orders is not imposed. The AACD model provides forecasts of where the trade direction is heading by making use of transaction data.

Several extensions of the PIN-AACD model can be considered. First, we may endogenize the probability of no news, good news and bad news. This could be done by constructing models based on aggregate data of each trading day. This approach may result in PIN estimates on a daily basis. The estimates may then be applied to study the impact of various events on trading using an event-study method. Second, we may adopt a procedure of assigning or classifying trading days by category of news. This series may be used to define the ACD equations, perhaps using a dummy variable technique. Overall, the use of transaction data opens up other possibilities of estimating PIN by relaxing several restrictions in the EHO model. The estimates can be applied in the literature in studying the effects of asymmetric information on asset prices.

## References

- [1] Aslan, H., D. Easley, S. Hvidkjaer, and M. O'Hara, 2006, The determinants of informed trading: Implications for asset pricing, Working Paper, Cornell University.
- [2] Bauwens, L., and P. Giot, 2003, Asymmetric ACD models: Introducing price information in ACD models, *Empirical Economics* 28, 709-731.
- [3] Bauwens, L., P. Giot, J. Grammig, and D. Veredas, 2004, A comparison of financial duration models via density forecast, *International Journal of Forecasting* 20, 589-604.

- [4] Bauwens, L., and D. Veredas, 2004, The stochastic conditional duration model: A latent factor model for the analysis of financial durations, *Journal of Econometrics* 119, 381-412.
- [5] Benos, E., and M. Jochev, 2006, Testing the PIN variable, Working Paper, University of Illinois.
- [6] Chung, K. H., and M. Li, 2003, Adverse-selection costs and the probability of information-based trading, *Financial Review* 38, 257-272.
- [7] Diebold, F. X., T. A. Gunther, and A. S. Tay, 1998, Evaluating density forecasts with applications to financial risk management, *International Economic Review* 39, 863-883.
- [8] Dufour A., and R. Engle, 2000, Time and the price impact of a trade, *Journal of Finance* 55, 2467-2498.
- [9] Easley, D., S. Hvidkjaer, and M. O'Hara, 2002, Is information risk a determinant of asset returns?, *Journal of Finance* 57, 2185-2221.
- [10] Easley, D., M. O'Hara, and G. Saar, 2001, How stocks splits affect trading: A microstructure approach, *Journal of Financial and Quantitative Analysis* 36, 25-51.
- [11] Easley, D., N. Kiefer, and M. O'Hara, 1996, Cream-skimming or profit sharing? The curious role of purchased order flow, *Journal of Finance* 51, 811-833.
- [12] Easley, D., N. Kiefer, and M. O'Hara, 1997, The information content of the trading process, *Journal of Empirical Finance* 4, 159-186.
- [13] Easley, D., N. Kiefer, M. O'Hara, and J. B. Paperman, 1996, Liquidity, information, and infrequently traded stocks, *Journal of Finance* 51, 1405-1436.
- [14] Easley, D., and M. O'Hara, 1992, Time and the process of security price adjustment, *Journal of Finance* 47, 577-605.
- [15] Engle, R., 2000, The econometrics of ultra-high frequency data, *Econometrica* 68, 1-22.

- [16] Engle R., and J. Russell, 1998, Autoregressive conditional duration: A new model for irregularly spaced transaction data, *Econometrica* 66, 1127-1162.
- [17] Henry, T. R., 2006, Short selling, informed trading, and stock returns, Working Paper, University of Georgia.
- [18] Fernandes, M., and J. Grammig, 2005, Non-parametric specification tests for conditional duration models, *Journal of Econometrics* 127, 35-68.
- [19] Ghysels, E., C. Gouriéroux, and J. Jasiask, 2004, Stochastic volatility duration models, *Journal of Econometrics* 119, 413-435.
- [20] Grammig, J., and K. O. Maurer, 2000, Non-monotone hazard functions and the autoregressive conditional duration model, *Econometrics Journal* 3, 16-38.
- [21] Lee, C., and M. Ready, 1991, Inferring trade direction from intradaily data, *Journal of Finance* 46, 733-746.
- [22] Pacurar, M., 2006, Autoregressive conditional duration (ACD) models in finance: A survey of the theoretical and empirical literature, Working paper, Dalhousie University.
- [23] Zhang, M. Y., J. R. Russell, and R. S. Tsay, 2001, A nonlinear autoregressive conditional duration model with applications to financial transaction data, *Journal of Econometrics* 104, 179-201.

**Table 1.** Summary Statistics of Duration and Trade Direction

Statistics	Ticker Symbols				
	BA	GE	IBM	MO	T
	Average Diurnally Adjusted Duration (in seconds)				
All Trades $\bar{x}$	88.78	31.83	41.42	48.88	39.29
Buy-initiated Trades $\hat{\psi}_1$	197.86	55.23	86.31	110.29	72.29
Sell-initiated Trades $\hat{\psi}_{-1}$	161.04	75.12	79.64	87.79	86.07
	Order-Flow Statistics (volume in lots)				
Frequency of Buys (%)	44.87	57.63	47.99	44.32	54.35
Frequency of Sells (%)	55.13	42.37	52.01	55.68	45.65
Serial Correlation of Trade Direction	0.35	0.32	0.52	0.32	0.40
Runs Test of Trade Direction	-81.32	-132.56	-186.27	-105.77	-146.61
Average Volume (lot size)	27.80	19.91	30.83	31.48	25.31
Average Log Volume	1.97	1.70	2.36	2.13	1.61
Average Daily Number of Trades	243.30	677.90	521.10	442.30	549.10
Average Daily Number of Buy-Trades	109.17	390.67	250.08	196.03	298.44
Average Daily Number of Sell-Trades	134.13	287.23	271.02	246.27	250.66
Number of Observations in Sample	54,500	170,157	129,239	110,120	135,087

**Table 2.** Estimates of PIN-EHO Model

Variables	Parameters	Ticker Symbols				
		BA	GE	IBM	MO	T
Intensity for Sell-Initiated Trade	$\lambda_{-1}$	98.7929 (3.2689)	345.8723 (6.5457)	200.8416 (6.4490)	175.2851 (5.1569)	235.5643 (4.8911)
Intensity for Buy-Initiated Trade	$\lambda_1$	110.9957 (3.6987)	269.7381 (4.7169)	251.5821 (6.6138)	227.1735 (5.9444)	241.7717 (5.6469)
Adjustment for Information	$\delta$	92.6196 (6.8988)	134.4572 (8.6544)	148.5920 (8.9076)	144.5634 (15.4491)	200.0465 (17.5005)
Probability of News	$\theta_E$	0.3511 (0.0366)	0.4560 (0.0426)	0.4556 (0.0347)	0.2683 (0.0350)	0.3539 (0.0358)
Given News, Probability of Bad News	$\theta_B$	0.6859 (0.0758)	0.2870 (0.0611)	0.2858 (0.0684)	0.4443 (0.1010)	0.1433 (0.0555)
PIN		0.1342	0.0906	0.1302	0.0879	0.1292

Figures in parentheses are standard errors.

**Table 3.** Estimates of the PIN-AACD Model and PIN

Trade Variables	Parameters	Ticker Symbols									
		BA		GE		IBM		MO		T	
Sale after Sale	$v_{-1,-1}$	2.0089 (0.2186)	2.0364 (0.2150)	2.3578 (0.0831)	2.3375 (0.0737)	2.9924 (0.0949)	2.9425 (0.0866)	1.9095 (0.1365)	1.9049 (0.1376)	1.7145 (0.0866)	1.7101 (0.0871)
Sale after Buy	$v_{-1,1}$	2.3501 (0.2329)	2.3910 (0.2291)	2.7250 (0.0875)	2.7151 (0.0790)	3.9757 (0.1158)	3.9451 (0.1111)	2.1540 (0.1489)	2.1531 (0.1497)	2.2727 (0.1045)	2.2620 (0.1052)
Buy after Sale	$v_{1,-1}$	2.1333 (0.2230)	2.0615 (0.2352)	1.8331 (0.0703)	1.8447 (0.0608)	3.1139 (0.1451)	3.1335 (0.1321)	1.8436 (0.1521)	1.8364 (0.1609)	2.2709 (0.0850)	2.2775 (0.0851)
Buy after Buy	$v_{11}$	1.7404 (0.1980)	1.6929 (0.2095)	1.6868 (0.0694)	1.7018 (0.0603)	2.2876 (0.1298)	2.3329 (0.1080)	1.5357 (0.1409)	1.5357 (0.1490)	1.7295 (0.0713)	1.7315 (0.0713)
Conditional Duration for Sales	$\alpha_{-1}$	0.5322 (0.0389)	0.5259 (0.0381)	0.3872 (0.0157)	0.3898 (0.0149)	0.1784 (0.0171)	0.1825 (0.0170)	0.4789 (0.0273)	0.4787 (0.0273)	0.4773 (0.0191)	0.4786 (0.0193)
Lagged Duration for Sales	$\beta_{-1}$	0.0890 (0.0057)	0.0895 (0.0058)	0.0702 (0.0046)	0.0715 (0.0041)	0.1043 (0.0060)	0.1078 (0.0059)	0.1193 (0.0062)	0.1204 (0.0063)	0.1332 (0.0046)	0.1331 (0.0046)
Conditional Duration for Buys	$\alpha_1$	0.5511 (0.0352)	0.5617 (0.0374)	0.5331 (0.0140)	0.5314 (0.0130)	0.3388 (0.0202)	0.3369 (0.0197)	0.5452 (0.0251)	0.5464 (0.0265)	0.4847 (0.0156)	0.4839 (0.0157)
Lagged Duration for Buys	$\beta_1$	0.1276 (0.0073)	0.1267 (0.0072)	0.0714 (0.0039)	0.0701 (0.0033)	0.1417 (0.0074)	0.1379 (0.0068)	0.1529 (0.0078)	0.1517 (0.0079)	0.1209 (0.0042)	0.1210 (0.0042)
Adjustment for Information	$\mu$	0.5240 (0.0261)	0.5245 (0.0274)	0.2771 (0.0134)	0.2783 (0.0141)	0.3901 (0.0208)	0.3983 (0.0240)	0.4479 (0.0363)	0.4453 (0.0360)	0.4863 (0.0353)	0.4866 (0.0356)
Probability of News	$\theta_E$	0.3191 (0.0369)	0.3174 (0.0368)	0.3850 (0.0382)	0.3802 (0.0383)	0.3342 (0.0485)	0.3444 (0.0364)	0.2046 (0.0304)	0.2044 (0.0305)	0.2459 (0.0411)	0.2456 (0.0409)
Given News, Probability of Bad News	$\theta_B$	0.9515 (0.0359)	0.9457 (0.0404)	0.4751 (0.1890)	0.3902 (0.1243)	0.4206 (0.2715)	0.2676 (0.1163)	0.7361 (0.1058)	0.7101 (0.1287)	0.0378 (0.0527)	0.0375 (0.0430)
Volume - Direction for Sales	$\varsigma_{-1}$	0.0410 (0.0051)	0.0423 (0.0051)	0.1041 (0.0037)	0.1058 (0.0037)	0.0227 (0.0037)	0.0232 (0.0037)	0.0555 (0.0035)	0.0561 (0.0036)	0.0475 (0.0033)	0.0471 (0.0033)
Volume - Direction for Buys	$\varsigma_1$	-0.0565 (0.0055)	-0.0545 (0.0055)	-0.1079 (0.0039)	-0.1069 (0.0038)	-0.0445 (0.0041)	-0.0445 (0.0037)	-0.0700 (0.0042)	-0.0692 (0.0042)	-0.0427 (0.0031)	-0.0430 (0.0032)
Shape parameter of sell order	$\phi_{-1}$	0.9087 (0.0053)		0.8971 (0.0039)		0.9163 (0.0052)		0.9116 (0.0043)		0.8551 (0.0040)	
Shape parameter of buy order	$\phi_1$	0.8544 (0.0056)		0.8616 (0.0040)		0.8735 (0.0050)		0.8857 (0.0047)		0.8680 (0.0037)	
Shape parameter of sell and buy orders	$\phi$		0.8834 (0.0042)		0.8766 (0.0030)		0.8952 (0.0039)		0.8999 (0.0039)		0.8620 (0.0032)
PIN			0.1530		0.0444		0.0871		0.1068		0.1347
Brier Score			0.4304		0.4187		0.3657		0.4353		0.4097

For each stock models with equal and unequal shape parameters of the latent sell- and buy-order processes are estimated. Figures in parentheses are standard errors. The PIN estimates are based on the model with equal shape parameters.

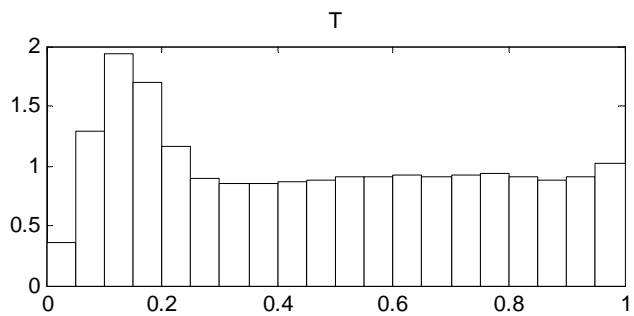
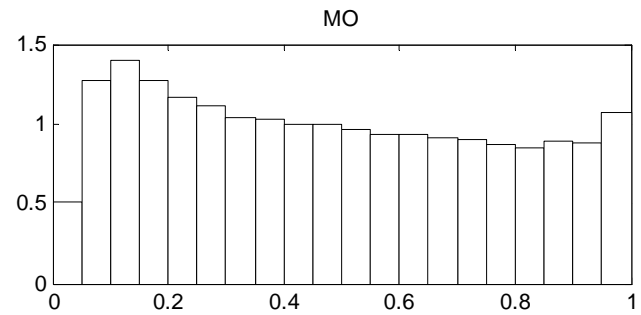
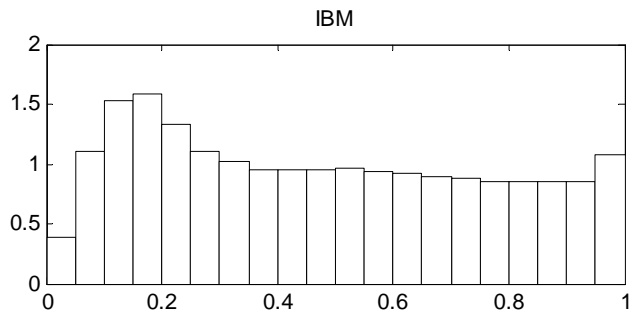
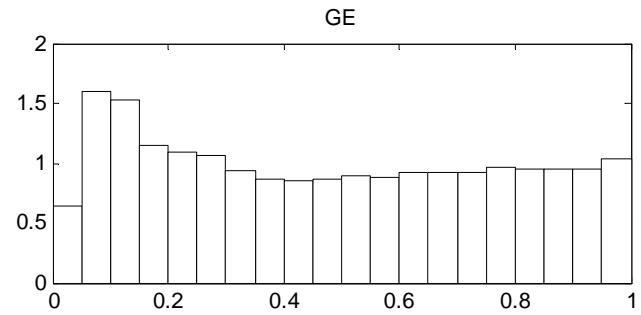
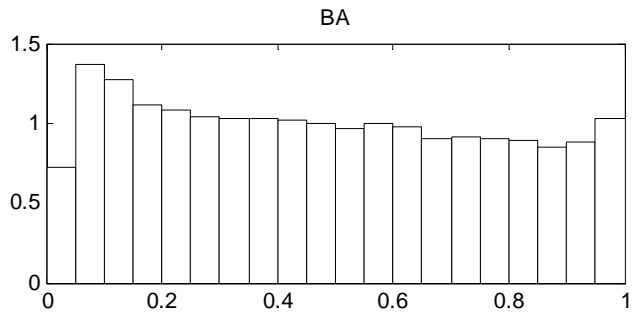


Figure 1: Empirical density function of the probability integral transform of the transaction duration of the PIN-ACCD model