JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. (2014)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2412

A HIDDEN MARKOV MODEL APPROACH TO INFORMATION-BASED TRADING: THEORY AND APPLICATIONS

XIANGKANG YIN AND JING ZHAO*
La Trobe University, Bundoora, Victoria, Australia

SUMMARY

This paper develops a novel approach to information-based securities trading by characterizing the hidden state of the market, which varies following a Markov process. Extensive simulation demonstrates that the approach can successfully identify market states and generate dynamic measures of information-based trading that outperform prevailing models. A sample of 120 NYSE stocks further verifies that it can better depict trading dynamics. With this sample, we characterize the features of information asymmetry and belief dispersion around earnings announcements. The sample is also applied to the study of the co-movements of trading activities due to private information or disputable public information. Copyright © 2014 John Wiley & Sons, Ltd.

Received 8 July 2013; Revised 30 May 2014

Supporting information may be found in the online version of this article.

1. INTRODUCTION

The impact of information on securities trading, and private information in particular, is a key concern of academic researchers and market practitioners alike. This paper develops a new approach to information-based trading of a risky financial asset, aiming to: well characterize the trading process in securities markets; accurately estimate the dynamic measures of trades stemming from information asymmetry and diverse opinions among investors; and successfully address financial issues associated with information-based trading. It applies the dynamic estimates of information-based trading generated from the new approach to study the information environment around earnings announcements of 120 New York Stock Exchange (NYSE) stocks. With this sample, it also investigates the co-movements of trading activities caused by different types of information.

Information, either private or public, is a driving force of trading activities. Our model proposes a latent/hidden process of market state to describe the evolution of a risky asset's information environment over time. The premise of our analysis is that the trading dynamics of a risky asset depend on its contemporaneous market state, which in turn can be characterized by the expected numbers of buyer-initiated orders and seller-initiated orders (hereafter, buy orders and sell orders) of the asset. Although the state is unobservable to financial analysts and econometricians, it can be inferred through observed trading data. We assume that there are three types of buy and sell orders submitted to the market. The first stems from the well-documented speculative investors who possess private information of the risky asset's value and use this information advantage to buy or sell the asset to maximize their profits. The second type of orders is caused by symmetric order-flow shocks (SOSs), as coined by Duarte and Young (2009). Although there is more than one possible source of SOSs, this paper inclines to the argument that different opinions and beliefs about public information events trigger SOSs, because a number of studies have found theoretical and empirical support for it (see Duarte and Young,

* Correspondence to: Jing Zhao, Department of Finance, La Trobe Business School, La Trobe University, Bundoora, Victoria 3086, Australia.
E-mail: j.zhao@latrobe.edu.au

Copyright © 2014 John Wiley & Sons, Ltd.

2009; Kandel and Pearson, 1995; Sarkar and Schwartz, 2009). We will use SOS orders/trading interchangeably with public information-induced orders/trading in this paper. The third type of trading orders is due to liquidity needs. Therefore, by identifying the state of a day we can detect the daily composition of the three types of trades as well as determine whether trading on that day has impounded private information and/or disputable public information.

Although the hidden Markov model (HMM) is arguably one of the most successful statistical modeling ideas arising in the last 50 years (Cappé *et al*., 2005), its applications to economics and finance are limited. To our knowledge, this paper is the first in the literature to develop an HMM of information-based trading. The most important advantage of the HMM approach is its ability to determine the daily composition of the three types of trades in a dynamic fashion. This in turn produces time-varying and relatively accurate measures of information-based trading. The PIN (Probability of INformed trading) measure developed by Easley *et al*. (1996) is arguably the most notable proxy for information asymmetry in the empirical literature. Duarte and Young (2009) extend the PIN analysis by introducing PSOS (Probability of Symmetric Order-flow Shock) to measure the relative intensity of trades due to different opinions on public information events. Our HMM approach adopts the same concepts of PIN and PSOS as proxies for measuring the proportions of trades induced by private information and disputable public information, respectively. However, it estimates all three types of trades based on a time-varying distribution of states to reflect the evolution of the information environment over time.

To evaluate the HMM approach, we first apply extensive Monte Carlo simulation experiments to investigate its effectiveness in identifying the most likely hidden state of each trading day. We find that the HMM can identify the daily hidden states with a very small misclassification rate. We then compare HMM's estimates of PIN and PSOS with those obtained from the three existing approaches, i.e. Easley *et al*. (2002, hereafter EHO), Duarte and Young (2009, hereafter DY), and Easley *et al*. (2008, hereafter EEOW).[1] The HMM approach performs better in estimating daily PIN and PSOS as well as PIN and PSOS over longer time intervals in all scenarios of the simulation. It also captures the volatility of daily PIN or PSOS with greater accuracy than the EEOW approach, while the EHO and DY approaches are static and simply pre-exclude such variations in their modeling.

The aforementioned four approaches are further evaluated by a sample of 120 randomly selected stocks traded on the NYSE in 2010 and 2011. Firstly, the likelihood ratio tests (Burnham and Anderson, 2002) show that the HMM significantly better describes the trading activities than the EHO and DY models for all 120 stocks. The Akaike information criterion (AIC) method (Burnham and Anderson, 2002) also demonstrates that the HMM approach is the best of the four in terms of minimizing information loss. Then, we explain the superior performance of the HMM by demonstrating its ability in capturing not only highly positive contemporaneous correlations between buy and sell order flows but also their serial correlations observed in transaction data. Thirdly, the HMM demonstrates that it can generate accurate out-of-sample forecasts of PIN, PSOS and order flow distributions, and its forecasting performance is superior to other models.

We apply the HMM estimates of PIN and PSOS to analyzing earnings announcements to shed insights on how information environments change before and after the announcements. The most notable findings include that on the announcement day there is a significant drop in PIN, consistent with the argument of Tetlock (2010) that public announcements can resolve information asymmetry. On the other hand, however, there is a substantial rise in PSOS, indicating diverse opinions among investors triggered by the released public information. After the announcement, PIN measure reverts back to its average level within 1 or 2 days, while PSOS remains high for a considerable period.

[1] Like the EEOW model, the asymmetric autoregressive conditional duration (AACD) model developed by Tay *et al*. (2009) and Preve and Tse (2013) can also generate time-varying measures of information-based trading. However, it requires data of the duration between two consecutive transactions and the volume of each transaction in addition to trade direction. It is applied to the estimation of PIN measure of five stocks and both PIN and PSOS measures of four stocks, respectively, in these two papers.

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

Another application of the HMM approach is to analyze co-movements in information-based trading. No substantial co-movement in PIN has been detected across sample stocks, while co-movements in PSOS are found to be extensive. In particular, when Standard & Poor's downgraded the US AAA credit rating to AA + on 5 August 2011, the PSOS measure of most sample stocks reached a very high level on that day and the following 2 days. Similarly, around the Flash Crash of 6 May 2010, co-movements in PSOS were also extremely high. These findings suggest that private information is more likely to trigger idiosyncratic trading shocks, while diverse opinions on public news could spread widely and have market-wide influences.

The remainder of the paper is organized as follows. In Section 2, we develop the HMM of information-based trading and outline the associated estimation methods. Section 3 evaluates the proposed HMM and dynamic PIN and PSOS measures by extensive Monte Carlo simulation experiments, while tables and figures reporting detailed simulation results are given as supporting information in the supplementary Appendix. Section 4 demonstrates the effectiveness of the HMM by matching the transaction data from a sample of 120 NYSE stocks. Earnings announcements and trading co-movement of the 120 NYSE stocks are analyzed in Section 5. Section 6 concludes the paper.

2. THE HMM APPROACH AND ITS LINK TO THE EXISTING MODELS

2.1. An HMM of Trading Dynamics

The proposed HMM consists of two parts: a two-dimensional unobservable stochastic process of state $\{H_t \equiv (H_{b;t}, H_{s;t})\colon\ t = 1, \cdots, T\}$, satisfying the Markov property; and a bivariate observable trading process $\{X_t \equiv (B_t, S_t) : t = 1, \cdots, T\}$, where T is the time horizon considered, H_t indicates the hidden information state on day t, and B_t and S_t represent the buyer- and seller-initiated orders of that day, respectively. Mathematically, this HMM is described by

$$\Pr\left(H_t|H_{(t-1)}\right) = \Pr(H_t|H_{t-1}) \text{ and } \Pr\left(X_t|X_{(t-1)}, H_{(t)}\right) = \Pr(X_t|H_t)$$

where $H_{(t)} \equiv (H_1, H_2, \ldots, H_t)$ and $X_{(t)} \equiv (X_1, X_2, \ldots, X_t)$. The evolution of hidden states can be characterized by the following transition matrix:

$$\Gamma = \begin{pmatrix} \gamma_{1,1;1,1} & \gamma_{1,1;1,2} & \cdots & \gamma_{1,1;m,n-1} & \gamma_{1,1;m,n} \\ \gamma_{1,2;1,1} & \gamma_{1,2;1,2} & & \gamma_{1,2;m,n-1} & \gamma_{1,2;m,n} \\ \vdots & & \ddots & & \vdots \\ \gamma_{m,n-1;1,1} & \gamma_{m,n-1;1,2} & \cdots & \gamma_{m,n-1;m,n-1} & \gamma_{m,n-1;m,n} \\ \gamma_{m,n;1,1} & \gamma_{m,n;1,2} & & \gamma_{m,n;m,n-1} & \gamma_{m,n;m,n} \end{pmatrix}$$

where $\gamma_{i,j;k,l} \equiv \Pr(H_{b;t+1} = k, H_{s;t+1} = l\,|H_{b;t} = i, H_{s;t} = j)$ is the probability of state being (k, l) on day $t + 1$ conditional on it being (i, j) on day t. Denote the unconditional probability of day t being in state (i, j) by $u_{i,j;t} \equiv \Pr(H_{b;t} = i, H_{s;t} = j)$, then the row vector $u_t \equiv (u_{1,1;t}, \ldots, u_{1,n;t}, \ldots, u_{m,1;t}, \ldots, u_{m,n;t})$ specifies the distribution of states on day t. We can therefore deduce the distribution of states on day $t + h$ through $u_{t+h} = u_t\Gamma^h$.

Given that the state of trading day t is (i, j), buy and sell order flows are assumed to arrive at the market according to a bivariate independent Poisson process, and the probability of observing b_t buy orders and s_t sell orders on that day is

$$p_{i,j}(X_t = x_t) = P_i(b_t)P_j(s_t)$$

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

where $P_i(b_t) = e^{-\lambda_{b;i}} \frac{(\lambda_{b;i})^{b_t}}{b_t!}$ and $P_j(s_t) = e^{-\lambda_{s;j}} \frac{(\lambda_{s;j})^{s_t}}{s_t!}$ and the arrival rates of buys and sells $\lambda_{b;i}$ and $\lambda_{s;j}$. Note that each pair of distributions as specified by $\lambda_{b;i}$ and $\lambda_{s;j}$ uniquely corresponds to a state of the market, (i, j). We use two subscripts, i and j, to index a state for the convenience of exposure. The unconditional probability of observing $x_t = (b_t, s_t)$ on day t can be calculated by

$$\Pr(X_t = x_t) = \sum_{i=1}^{m}\sum_{j=1}^{n} \Pr\left(x_t | H_{b;t} = i, H_{s;t} = j\right) \Pr\left(H_{b;t} = i, H_{s;t} = j\right) = u_t P(x_t)\mathbf{1}$$

where $mn \times mn$ -diagonal matrix $P(x_t)$ is defined by

$$P(x_t) \equiv \begin{pmatrix} P_1(b_t)P_1(s_t) & & 0 \\ & \ddots & \\ 0 & & P_m(b_t)P_n(s_t) \end{pmatrix} \text{ and } \mathbf{1} \equiv \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

The HMM also yields a probability distribution of states for each day, conditional on the history of observed trades:

$$\Pr\left(H_{b;t} = i, H_{s;t} = j | X_{(T)} = x_{(T)}\right) = \frac{\Pr\left(X_{(T)} = x_{(T)}, H_{1;t} = i, H_{2;t} = j\right)}{\Pr\left(X_{(T)} = x_{(T)}\right)}, \text{ for } t = 1, 2, \ldots, T \tag{1}$$

The parameters of the HMM include u_1, Γ, and $\lambda_{b;i}$ and $\lambda_{s;j}$ ($i = 1, 2, \ldots, m, j = 1, 2, \ldots, n$), which can be estimated by maximizing the following likelihood function:[2]

$$L\left(\mathbf{\Theta} | x_{(T)}\right) = u_1 P(x_1)\Gamma P(x_2)\Gamma P(x_3)\cdots\Gamma P(x_T)\mathbf{1} \tag{2}$$

The numbers of buy and sell states, m and n, are determined in model selection by certain information criterion, such as AIC or Bayesian information criterion (BIC).

2.2. Estimation of Trading Motives

After determining state by estimating $\lambda_{b;i}$ and $\lambda_{s;j}$, we need further to estimate the components of these order arrival rates. Since we consider three types of trading motives, the total expected numbers of buy and sell orders can be written as

$$\begin{aligned} \lambda_{b;i} &= \varepsilon_{b;i} + \mu_{b;i} + v_{b;i},\ i = 1,\ 2,\ \ldots, m, \\ \lambda_{s;j} &= \varepsilon_{s;j} + \mu_{s;j} + v_{s;j},\ j = 1,\ 2,\ \ldots,\ n \end{aligned} \tag{3}$$

where $\varepsilon_{b;i}$ and $\varepsilon_{s;j}$ are arrival rates of buy and sell orders from liquidity traders, $\mu_{b;i}$ and $\mu_{s;j}$ from privately informed traders, and $v_{b;i}$ and $v_{s;j}$ from public information (i.e. SOS) traders. We develop a two-step approach, based on the k-means clustering analysis together with the jump method of Sugar and James (2003), to identify these three types of trading for each information state.[3]

Step 1. Partitioning hidden states depending on whether they contain private information or not.

The insight shed by the EHO model is that private information leads to one-sided trading and substantial trade imbalance. Following this insight, we perform k-means clustering analysis on observed

[2] The proof of equation (2) and the parameter estimation details are given in the supplementary Appendix (S.1 and S.2).

[3] For a brief introduction of the k-mean clustering analysis, see Section S.3 of the supplementary Appendix.

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

trading imbalances over the whole estimation window, i.e. $|b_t - s_t|$ for $t = 1, 2, \ldots, T$, and determine the number of clusters by using the jump method of Sugar and James (2003). If there is only one cluster, we infer that observed trading imbalances are similar and there is no significant evidence for the existence of private information during the period. Therefore, we have $\mu_{b;i} = \mu_{s;j} = 0$ for all hidden states. The rationale behind this claim is the common 'wisdom' that trading due to liquidity needs or disputable public information is symmetric and only generates small trade imbalance, while privately informed trading yields substantial trade imbalance. Thus the daily trade imbalances cannot be consistent and similar over time if there is privately informed trading.

If the clustering analysis indicates multiple clusters of trade imbalances, the clusters with their centers no larger than that of the most frequent cluster are identified as the clusters without privately informed trading. The rationale of the classification once again is that private information induces profound trading imbalance. We choose the most frequent cluster for the cut-off point because it includes states which occur most often; and it is plausible to assume that most trading days do not have private information and/or disputable public information. In our simulation and sample data, the most frequent cluster in trade imbalances always turns out to be the one with the smallest center. We treat $|\lambda_{b;i} - \lambda_{s;j}|$ as an out-of-sample observation and assign it to the cluster whose center is the closest to it. If $|\lambda_{b;i} - \lambda_{s;j}|$ belongs to a cluster without privately informed trading, we have $\mu_{b;i} = \mu_{s;j} = 0$, and we use $\mathcal{S}$ to denote the set of such states. If state (i, j) does not belong to $\mathcal{S}$, it contains private information, and we have

$$\mu_{b;i} = |\lambda_{b;i} - \lambda_{s;j}| - |\lambda_{b;i^\#} - \lambda_{s;j^\#}| \text{ and } \mu_{s;j} = 0 \text{ if } \lambda_{b;i} > \lambda_{s;j}$$
$$\mu_{b;i} = 0 \text{ and } \mu_{s;j} = |\lambda_{b;i} - \lambda_{s;j}| - |\lambda_{b;i^\#} - \lambda_{s;j^\#}| \text{if } \lambda_{b;i} < \lambda_{s;j}$$

where $(i^\#, j^\#)$ is a matching state of state (i, j), which is a state in $\mathcal{S}$ with balanced trade that is the closest to the balanced trade of state (i, j).[4] The matching state is used to proxy the small trade imbalance caused by liquidity and/or SOS trading in state (i, j). We use $\mathcal{A}$ to denote the set of states with privately informed trading.

Step 2. Classifying hidden states into two sets depending on whether they contain disputable public information or not.

Liquidity trading exists on each trading day, which is symmetric in the sense that their average numbers of buys and sells are not considerably different. As argued by Duarte and Young (2009), if a public information event causes controversy and debate in its interpretation, both buys and sells surge, leading to a shock to balanced trading. Thus we conduct a k-means clustering analysis of Sugar and James (2003) on the observations of balanced trades $b_t + s_t - |b_t - s_t| (t = 1, 2, \ldots T)$ to separate states. If only one cluster exists, it implies that investors have very similar interpretations for public information in the market and all balanced orders are generated by liquidity traders. Therefore, we have $v_{b;i} = v_{s;j} = 0$ for all (i, j).

If more than one cluster is detected, we follow the similar rationale of state classification in Step 1 and treat the clusters with their centers larger than that of the most frequent cluster as the ones associated with disputable public information. For state (i, j), we take the expected number of balanced trades $\lambda_{b;i} + \lambda_{s;j} - |\lambda_{b;i} - \lambda_{s;j}|$ as an out-of-sample observation, and assign it to the cluster with the closest center to it. We use $\mathcal{P}$ to denote the set consisting of the states in the clusters with disputable public information, and the remaining states constitute set $\mathcal{L}$. If state (i, j) belongs to $\mathcal{L}$, we have $v_{b;i} = v_{s;j} = 0$. If state (i, j) belongs to $\mathcal{P}$

[4] Mathematically, $(i^\#, j^\#) = \operatorname{argmin}_{(i^*, j^*) \in \mathcal{S}} \left| \lambda_{b;i} + \lambda_{s;j} - |\lambda_{b;i} - \lambda_{s;j}| - \left(\lambda_{b;i^*} + \lambda_{s;j^*} - |\lambda_{b;i^*} - \lambda_{s;j^*}|\right) \right|$.

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

$$v_{b;i} = \lambda_{b;i} - \mu_{b;i} - \max_{(i^{\#},j^{\#})\in \mathcal{L}\cap\mathcal{S}} \{\lambda_{b;i^{\#}}\}$$

$$v_{s;j} = \lambda_{s;j} - \mu_{s;j} - \max_{(i^{\#},j^{\#})\in \mathcal{L}\cap\mathcal{S}} \{\lambda_{s;j^{\#}}\}$$

where $\mu_{b;i}$ and $\mu_{s;j}$ are obtained in the first step. The last terms in the above equations proxy $\varepsilon_{b;i}$, and $\varepsilon_{s;j}$ in equation (3), respectively. Set $\mathcal{L}$ includes both liquidity trading and privately informed trading, while set $\mathcal{S}$ includes both liquidity trading and public information trading. Their intersection, i.e. set $\mathcal{L}\cap\mathcal{S}$, includes states that involve only liquidity trading. We use the largest arrival rates of buy and sell orders in $\mathcal{L}\cap\mathcal{S}$ to subtract liquidity order arrival rates from the aggregate buy and sell order arrival rates to ensure that the arrival rates of buy and sell orders driven by disputable public information are not exaggerated.

2.3. PIN and PSOS from the HMM and their Links to the Existing Measures

The daily PIN is defined as the ratio of the expected number of buy and sell orders stemming from privately informed traders on day t to the expected total trades of that day:

$$\mathrm{PIN}_t^{\mathrm{HMM}} = \frac{\sum_{i,j}\left(\mu_{b;i} + \mu_{s;j}\right)\Pr\left(H_{b;t} = i, H_{s;t} = j \middle| x_{(T)}\right)}{\sum_{i,j}\left(\lambda_{b;i} + \lambda_{s;j}\right)\Pr\left(H_{b;t} = i, H_{s;t} = j \middle| x_{(T)}\right)} \tag{4}$$

Similarly, the probability of trades due to a shock to both buy and sell order flows (PSOS) is defined as the ratio of the expected number of trades from public information traders to the expected total number of trades of trading day t and it can be calculated by

$$\mathrm{PSOS}_t^{\mathrm{HMM}} = \frac{\sum_{i,j}\left(v_{b;i} + v_{s;j}\right)\Pr\left(H_{b;t} = i, H_{s;t} = j \middle| x_{(T)}\right)}{\sum_{i,j}\left(\lambda_{b;i} + \lambda_{s;j}\right)\Pr\left(H_{b;t} = i, H_{s;t} = j \middle| x_{(T)}\right)} \tag{5}$$

The estimates presented in equations (4) and (5) can be extended for measuring PIN and PSOS over multiple-day interval $[t_1, t_2]$:

$$\mathrm{PIN}_{[t_1,\ t_2]}^{\mathrm{HMM}} = \frac{\sum_{t=t_1}^{t_2}\sum_{i,j}\left(\mu_{b;i} + \mu_{s;j}\right)\Pr\left(H_{b;t} = i, H_{s;t} = j \middle| x_{(T)}\right)}{\sum_{t=t_1}^{t_2}\sum_{i,j}\left(\lambda_{b;i} + \lambda_{s;j}\right)\Pr\left(H_{b;t} = i, H_{s;t} = j \middle| x_{(T)}\right)}$$

$$\mathrm{PSOS}_{[t_1,,t_2]}^{\mathrm{HMM}} = \frac{\sum_{t=t_1}^{t_2}\sum_{i,j}\left(v_{b;i} + v_{s;j}\right)\Pr\left(H_{b;t} = i, H_{s;t} = j \middle| x_{(T)}\right)}{\sum_{t=t_1}^{t_2}\sum_{i,j}\left(\lambda_{b;i} + \lambda_{s;j}\right)\Pr\left(H_{b;t} = i, H_{s;t} = j \middle| x_{(T)}\right)}$$

The EHO and DY models can be considered special cases of the HMM. In the EHO model there are only two types of traders, with either private information or liquidity needs. A private information event occurs at the beginning of a trading day with probability α, and privately informed traders observe a signal of it. With probability $1 - \delta$ (δ), the signal is high (low) and privately informed traders buy (sell) the asset by submitting buy (sell) orders to the market with arrival rate μ. Therefore, private information leads to one-sided trading and considerable trade imbalance. Liquidity traders submit their buy and sell orders following Poisson processes with arrival rates ε_b and ε_s, respectively, which are

Copyright © 2014 John Wiley & Sons, Ltd.

independent of private information events. There are three possible information states in the EHO model and one can calculate PIN by

$$\mathrm{PIN}^{\mathrm{EHO}} = \frac{\alpha\mu}{\alpha\mu + \varepsilon_b + \varepsilon_s} \tag{6}$$

Duarte and Young (2009) extend the EHO model in two directions by allowing the arrival rates of buys and sells from privately informed traders to be different (i.e. using μ_b and μ_s to replace μ), and introducing SOS trading. Thus

$$\mathrm{PIN}^{\mathrm{DY}} = \frac{\alpha((1-\delta)\mu_b + \delta\mu_s)}{\alpha((1-\delta)\mu_b + \delta\mu_s) + (v_b + v_s)(\alpha\theta' + (1-\alpha)\theta) + \varepsilon_b + \varepsilon_s} \tag{7}$$

where θ' (θ) is the probability of SOS occurring conditional on a private information event occurring (not occurring). Moreover

$$\mathrm{PSOS}^{\mathrm{DY}} = \frac{(v_b + v_s)(\alpha\theta' + (1-\alpha)\theta)}{\alpha((1-\delta)\mu_b + \delta\mu_s) + (v_b + v_s)(\alpha\theta' + (1-\alpha)\theta) + \varepsilon_b + \varepsilon_s} \tag{8}$$

In the EHO and DY models, the probability distribution of states is static over the whole estimation horizon. This time-invariant distribution leads to constant PIN and PSOS measures over the whole estimation horizon as shown by equations (6)–(8). Both EHO and DY models can be treated as special cases of the proposed HMM, where the Markov transition matrix Γ is restricted to an identity matrix. In addition, these two popular models assume that all private signals or disputable public information events have the same impact on the market, characterized by the unique and constant order arrival rates μ_b (μ_s) or v_b (v_s). These assumptions are rather restrictive, as one may expect that information events are more likely to be heterogeneous than homogeneous.

Easley *et al*. (2008) extend the EHO model in another direction by introducing time-varying arrival rates of informed and uninformed trades but ignoring trades triggered by public information events. After assuming the arrival rates of liquidity buy and sell orders are equal on every day and other parameters such as α and δ are constant over time, they specify that the arrival rate forecasts follow a bivariate vector autoregressive process with predetermined forcing variables:[5]

$$\begin{pmatrix} \alpha\mu_t \\ 2\varepsilon_t \end{pmatrix} = \omega \odot e^{g(t-1)} + \Phi\left[\begin{pmatrix} \alpha\mu_{t-1} \\ 2\varepsilon_{t-1} \end{pmatrix} \odot e^{g(t-1)}\right] + \Psi\begin{pmatrix} |b_{t-1} - s_{t-1}| \\ b_{t-1} + s_{t-1} - |b_{t-1} - s_{t-1}| \end{pmatrix} \tag{9}$$

where $(\alpha\mu_t,\ 2\varepsilon_t)'$ contains the time-$(t-1)$ forecast of the arrival rates at time t, b_{t-1} and s_{t-1} are observed buy and sell orders at time $t-1$, $\odot$ denotes the Hadamard product, the vector $g = (g_1, g_2)'$ captures the growth rates of the two intensities, ω is a 2×1 parameter vector and Φ and Ψ are 2×2 parameter matrices. Thus, in the EEOW model, there are three possible information states but PIN is time varying, which can be calculatd by replacing μ and ε in equation (6) by μ_t and ε_t, respectively.

3. EFFECTIVENESS OF THE HMM, PIN$^{\mathrm{HMM}}$ AND PSOS$^{\mathrm{HMM}}$: A SIMULATION ANALYSIS

It is important to evaluate whether the proposed hidden Markov model and its estimation procedure work well. In particular, we want to investigate whether the most likely hidden state of a day is

[5] To reach this model, Easley *et al*. (2008) apply the approximation of $E(|S-B|) \doteq \alpha\mu$.

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

correctly identified by the HMM, and whether its PIN and PSOS estimates are accurate. To answer these questions, we conduct an extensive Monte Carlo simulation analysis, in which the series of hidden states and the associated PIN and PSOS are known and therefore the performance of a candidate approach can be evaluated. The HMM approach is evaluated by contrasting its performance with those of the EHO, DY and EEDW approaches. We report the main findings here, while the simulation results are tabulated and plotted in Section S.6 of the supplementary Appendix to control the length of the paper.

3.1. Experimental Design of Monte Carlo Simulation

We conduct an extensive Monte Carlo simulation analysis in which the series of hidden states and the associated PIN and PSOS are known. We create 16 different scenarios of trading processes as follows to generate hypothetical trading data. The parameters of each simulation scenario are detailed in Panel C of Table S-I in the supplementary Appendix. In Scenarios 1.1–1.4, the trading process is described by the EHO model. On a hypothetical trading day t, a draw from a Bernoulli distribution with parameter α determines whether private information arrives or not. If a private signal is observed, a draw from a Bernoulli distribution with probability δ indicates whether it is a high or low signal. In Scenarios 2.1–2.4, the trading process is described by the DY model. Compared with Scenarios 1.1–1.4, we need an additional draw from a Bernoulli distribution with parameter θ to reflect the occurrence of SOS events.

The trading process in Scenarios 3.1–3.4 is described by the EEOW model. The states are determined in the same way as Scenarios 1.1–1.4. The order arrival rate is time varying, however, and follows a bivariate vector autoregressive process. In order to have a time-varying counterpart of the DY model, we extend the EEOW model in Scenarios 4.1–4.4 with the arrival rate forecasts contained in a bivariate vector autoregressive process with predetermined forcing variables:[6]

$$\begin{pmatrix} \alpha\mu_t \\ 2\varepsilon_t + 2\theta\upsilon_t \end{pmatrix} = \omega \odot e^{gt} + \Phi\left[\begin{pmatrix} \alpha\mu_{t-1} \\ 2\varepsilon_{t-1} + 2\theta\upsilon_{t-1} \end{pmatrix} \odot e^{gt}\right] + \Psi\begin{pmatrix} |b_t - s_t| \\ b_t + s_t - |b_t - s_t| \end{pmatrix}$$

where the arrival rates of buys and sells due to SOSs are assumed to be equal. Since ε_t and υ_t are modeled in one quantity, one of them is assumed to be constant to make the model identifiable. Note that this model collapses to the original EEOW model if $\theta = 0$.

After the hidden state of hypothetical day t is obtained, the order arrival rates are determined so that the daily observations of buys and sells, b_t and s_t, can be generated by the draws from the corresponding Poisson distributions. With simulated data, the true daily PIN and PSOS are known. For instance, PIN_t in Scenarios 1.1–1.4 is zero if no private signal arrives and is equal to $\frac{\mu}{\varepsilon_b+\varepsilon_s+\mu}$ if a private information event occurs, while PSOS_t is always zero because the EHO model excludes SOSs. We deliberately use the three compared candidates of the HMM to generate simulation data so that their estimates in the corresponding scenarios have no model specification errors. This method of data generation can enhance our conclusions about how well the HMM approach functions when it outperforms other approaches.

3.2. Identification of Hidden States

Decoding or identifying the most likely hidden state is one of the main objectives in many applications. It also sets a test to examine the model's validity. We infer the most likely hidden state for each trading day by using the mode of its conditional distribution specified by equation (1) in Section 2. We then

[6] Similar to the EEOW model, this model is based on the approximation of $E(S+B-|S-B|) \doteq 2\varepsilon + 2\theta\upsilon$.

Copyright © 2014 John Wiley & Sons, Ltd.

check whether it matches with the day's true state implied by the simulation data. The misclassification rate is calculated as the percentage of trading days whose states are identified incorrectly. Panel A of Table S-I in the supplementary Appendix reports the average misclassification rate of the 100 replications in each scenario and shows that the most likely state of a day decoded by the HMM well coincides with the true state of that day. When the estimation window is longer and more data are available, the mean misclassification rate becomes even smaller. For example, if the estimation window is 252 days and the AIC is adopted, the largest misclassification rates of hidden states for the four sets of scenarios are 0.77%, 2.13%, 1.55% and 6.54%, respectively, which implies that the HMM approach can correctly identify the hidden state for more than 93% of the trading days over the whole estimation window. We find that the EM algorithm can correctly identify the initial hidden state in more than 92% of the replications for all the simulation scenarios considered. The details are available from the supplementary Appendix, i.e. Panel B of Table S-I.

3.3. Performance of Estimated PIN and PSOS

To evaluate the performance of the HMM approach further, we apply the four candidate approaches (HMM, EHO, DY and EEOW) to the simulated data used in Table S-I to compare the accuracies of their estimates of PIN and PSOS.

Hidden state is a random variable and thus the true daily PIN and PSOS are time varying. For a trading day without private information, its daily PIN is definitely zero. If observed data imply that a particular day has a high probability of privately informed trading, the associated daily PIN should be large. The same reasoning applies to PSOS. Figure S-3 in the supplementary Appendix plots the absolute errors of daily PIN and PSOS estimates of the four approaches in the first replication of Scenario 2.1 with an estimation window of 63 trading days. For PIN, the HMM approach provides the most accurate estimates among the four without any absolute error greater than 0.02. Owing to their static nature over the whole estimation window, the EHO and DY approaches fail to capture the daily variations in PIN. Although the EEOW approach allows for variations in PIN over time, its estimates are far from sufficient to accommodate the daily changes of privately informed trading. Since Scenario 2.1 is generated by the DY model, SOSs may exist within the estimation window. However, the EHO and EEOW approaches assign a zero PSOS to all trading days and thus cannot accommodate a non-zero PSOS on days 12, 18, 39, 45 and 61, as shown in the lower part of Figure S-3. The DY approach uses the constant estimate of PSOS for the whole estimation window and is unable to capture the daily variation of PSOS. The HMM approach is the only one of the four candidates that can better identify the occurrence of SOSs for each trading day and give us a daily estimate of PSOS virtually without error.

For 100 replications of each scenario, we consider the overall performance of daily estimates in terms of mean absolute error (MAE):

$$\mathrm{MAE}^{(r)}_{\mathrm{PIN}_t^{[0,T]}} = \frac{1}{T}\sum\nolimits_{t=1}^{T}\left|\widehat{\mathrm{PIN}}_t^{(r)} - \mathrm{PIN}_t^{(r)}\right|, \mathrm{MAE}^{(r)}_{\mathrm{PSOS}_t^{[0,T]}} = \frac{1}{T}\sum\nolimits_{t=1}^{T}\left|\widehat{\mathrm{PSOS}}_t^{(r)} - \mathrm{PSOS}_t^{(r)}\right|$$

where $\mathrm{PIN}_t^{(r)}$ and $\mathrm{PSOS}_t^{(r)}$ denote the true PIN and PSOS of the rth replication on day t, and $\widehat{\mathrm{PIN}}_t^{(r)}$ and $\widehat{\mathrm{PSOS}}_t^{(r)}$ are the daily estimates obtained by a candidate approach. Figure S-4 plots the MAE of daily PIN and PSOS estimates for each replication of Scenario 2.1 over an estimation window of 63 days. We find that the HMM approach provides the most accurate daily estimates for all replications. The simulation results of daily PIN and PSOS estimates for the 16 simulation scenarios are documented in Table S-II. Several observations are worth noting. First, consistent with the findings from Figure S-4, the HMM generates the smallest mean of MAEs of PIN estimates for all 16 scenarios and for both estimation windows of 63 and 252 trading days. Its mean MAE is negligible and about 1/10 to 1/100 of the other

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

approaches' mean MAE. Second, in Scenarios 1.1–1.4 and 3.1–3.4 the true value of PSOS is zero for all trading days and the HMM approach's mean *MAEs* of PSOS are equal to zero or negligibly small. Therefore, the HMM approach can effectively exclude the type of orders not existing in the market.

In the meantime, the PSOS estimated by the DY approach is also quite accurate, though it is outperformed by the HMM approach in general. In Scenarios 2.1–2.4 and 4.1–4.4, the true value of daily PSOS on some days is positive, and the HMM approach is able to identify the occurrence of SOSs on a daily level, as evidenced by its negligible mean MAEs of PSOS. Its improvements in accuracy compared with the other three candidate approaches range from about three times to 60 times. Third, with a longer estimation window, i.e. from 63 trading days to 252 trading days, more data are available to build up the likelihood of state and therefore the accuracy of the daily estimates of PIN and PSOS are improved in all scenarios irrespective of the candidate approach used.

To have a more complete picture, we further examine the volatilities (standard deviations) of daily PIN and PSOS. Their means from the four candidate approaches over the 100 replications in each scenario are documented in the last four columns of Table S-II. The EHO and DY approaches generate constant estimates of PIN and PSOS with zero volatility. The standard deviation of daily PIN estimated by the EEOW approach is at least three times less than the true standard deviation. In contrast, the volatility of the HMM's PIN estimates is much closer to the true value. Meanwhile, of the four candidates, only the HMM approach can generate time-varying estimates of daily PSOS and its variation is very close to the true one. Furthermore, we take Scenario 2.1 again as an example and Figure S-5 illustrates the absolute errors of the standard deviations of the daily PIN and PSOS estimated by the HMM and EEOW approaches in each replication. It can be seen that the HMM approach captures the variations more accurately than the EEOW approach.

We have also evaluated the performance of PIN and PSOS estimates over a specific time interval. Figure S-6 takes the first replication of Scenario 2.1 as an example again and depicts the estimated PIN and PSOS over the interval [0, *t*] in the upper and lower charts, respectively. For the EHO, DY and EEOW approaches, the short-term deviations of their PIN measures from the true values are significant. For instance, the absolute error of their 2-day or 3-day PIN is larger than 0.06, i.e. at least 6% of trades are misattributed on average. As the length of time interval increases, the DY approach can generate a PIN measure approaching the true value, but errors in the PIN measures of the EHO or EEOW approach remain significant. In contrast, the PIN measure of the HMM approach almost coincides with the true one irrespective of the length of time interval considered. In the lower part of Figure S-6, we can see that whenever an SOS event occurs in the market the HMM approach can capture the changes in PSOS contemporaneously, evidenced by its estimates of PSOS overlapping with the true values. Both the EHO and EEOW approaches generate a zero PSOS and thus fail to capture the occurrence of SOS events. PSOS^{DY} is constant over the whole estimation window, which can fairly capture the overall SOS trading but cannot accommodate its short-term variations. The results for the 16 simulation scenarios are compiled in Table S-III and we summarize here the main findings. First, PIN^{HMM} and PSOS^{HMM} are close to the true values in all scenarios and lengths of time periods considered. Second, although the EHO, DY and EEOW approaches manage to generate estimates of PIN and/or PSOS of the whole estimation window (e.g. 252 trading days) with satisfactory accuracy in some scenarios, their estimates over sub-periods (e.g. 5 trading days) suffer from significant errors. Third, when the hypothetical trading data not only incorporate SOS events but also time-varying order arrival rates (i.e. in Scenarios 4.1–4.4), the HMM approach's performance is always the best of the four candidates. Arguably, this set of scenarios in all four sets is the closest to the real status of the market.

4. EMPIRICAL EFFECTIVENESS AND IMPLICATIONS OF THE HMM

In addition to evaluating the validity of the HMM by simulation experiments, this section uses observed transaction data to examine the effectiveness of the HMM in describing trading activities

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

relative to the alternative models. The characteristics of the hidden states are reported in Section S.5 of the supplementary Appendix to control the length of the paper.

4.1. Sample Data

Our dataset is a sample of 120 stocks that were traded on the NYSE in 2010 and 2011. In order to choose representative stocks from a variety of industries and market capitalizations, we randomly select 40 stocks each from the S&P 500 Index, S&P MidCap 400 Index and S&P SmallCap 600 Index. The ticker symbols of these sample stocks are listed in panel A of Table I. Transaction data of these stocks are taken from the Thomas Reuters Tick History (TRTH) transaction database from 1 January 2010 to 31 December 2011. We exclude transactions and quotes that occur before and at the open, as well as those at and after the close. Quotes with zero bid or ask prices, quotes for which the bid–ask spread is greater than 50% of the price, and transactions with zero prices are also excluded to eliminate possible data errors. Data for 26 November 2010 and 25 November 2011 are removed because of an early 'day after thanksgiving' closing. The Lee–Ready (1991) algorithm is applied to the TRTH transaction data to determine the daily numbers of buys and sells. Although the algorithm is imperfect, Chakrabarty *et al*. (2012) report that its misclassification rates are near zero at the daily aggregate level, because buy-side errors and sell-side errors offset each other in the process of daily aggregation. Given that estimations of the HMM and the alternative models use daily aggregated observations of buy and sell orders, the effect of the classification errors on the estimation of PIN and PSOS is likely to be acceptable. From the Center for Research in Security Prices (CRSP), we obtain data on the daily return of each sample stock.

4.2. Model Selection

For each stock in the sample, we estimate parameters of the four candidate models first. Because the DY model extends the EHO model, and the EHO and DY models are special cases of the HMM, we use the likelihood ratio test (see Burnham and Anderson, 2002) to compare the fit of a pair of candidate models for each stock in the sample. The EHO model is rejected at the 5% level in favor of the DY model, but the latter is rejected at the 5% level in favor of the HMM for all stocks in the sample.

Although the HMM is preferred over the EHO and DY models according to the likelihood ratio test, this test is inapplicable for comparison with the EEOW model because it is not nested with the other three models. Therefore, we further employ the AIC method introduced by Burnham and Anderson (2002) to compare the relative goodness of fit of candidate models by examining the estimated information loss. Of the four candidate models, the HMM generates the minimum information loss for all stocks in the sample, and the other three models are almost zero times as probable as the HMM to minimize the information loss, which suggests selecting the HMM over the other three models for all the sample stocks.

Finally, assume N_i^* is the total number of states of the sample stock i, endogenously determined by the HMM. We can use the likelihood ratio test to compare the fit of the N_i^* -state HMM with that of the $\left(N_i^* - 1\right)$ -state HMM. The $\left(N_i^* - 1\right)$ -state HMM is rejected at the 1% level in favor of the N_i^* -state HMM for all the stocks in the sample. It further assures that the parameters determined by the HMM reflect the true dynamics of the order arrivals.

4.3. Empirical and Model Implied Moments

Whether model implied moments are consistent with empirically observed moments is another test for model validity. A key criticism of Duarte and Young (2009) concerning the PIN model of Easley *et al.*

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

Table I. Ticker symbols of the sample stocks and summary statistics of the moments of buy and sell orders implied by empirical data and the four candidate models

Panel A: Ticker symbols of the 120 sample stocks

S&P 500 constituents				S&P MidCap 400 constituents				S&P SmallCap 600 constituents			
ACE	CLX	DNB	FMC	AAP	BGC	CVD	FLO	ABM	AZZ	CCC	DRH
ICE	MKC	SCG	TAP	HNI	MAN	SM	URS	FOR	KWR	ORB	SNX
BCR	CMG	DO	GME	ADS	BKH	DCI	FRT	AHS	BDC	CRY	DW
IGT	PCP	SHW	TWC	IEX	MLM	SON	VCI	FUL	LDL	POL	SUP
BWA	COL	DOV	GPC	AJG	CBT	DLX	GES	AIT	BGS	CUZ	EIG
LUK	PLL	SNI	TXT	JLL	NYT	SXT	WRB	HHS	LTC	PPS	UBA
CB	DD	EW	HAR	AYI	CRL	DRC	GGG	ALE	BMI	DAR	EXP
MA	PWR	SWK	VMC	KEX	PNM	TKR	WSO	IVC	MED	RT	UNF
CCE	DGX	FII	HSP	BCO	CSL	ESI	GHL	AXE	CBR	DIN	FIX
MCD	ROP	SWY	ZMH	LNT	ROL	TRN	XEC	KRG	ONB	SCL	ZEP

Panel B: Summary statistics of empirical and model implied contemporaneous correlations between buy and sell orders

	Sample data	HMM	EHO model	DY model	EEOW model
Mean	0.838	0.8584	−0.1404	0.2679	−0.2470
(Mean AE)		(0.0195)	(0.9793)	(0.5708)	(1.0858)
Median	0.8571	0.8840	−0.1420	0.2672	−0.2124
(Median AE)		(0.0180)	(0.9945)	(0.5868)	(1.0529)
SD	0.0890	0.0860	0.0363	0.0381	0.1164
(SD of AE)		(0.0074)	(0.0787)	(0.0748)	(0.1546)
Minimum	0.5679	0.5878	−0.2156	0.1783	−0.6241
(Minimum AE)		(0.0088)	(0.7037)	(0.3557)	(0.7308)
Maximum	0.9574	0.9662	−0.0414	0.3580	−0.0577
(Maximum AE)		(0.0423)	(1.1179)	(0.7157)	(1.5516)

Panel C: Summary statistics of empirical and model implied serial correlations of buy orders and sell orders

	Sample data		HMM		EEOW model	
	$\mathrm{ACF}_1(B_t)$	$\mathrm{ACF}_1(S_t)$	$\mathrm{ACF}_1(B_t)$	$\mathrm{ACF}_1(S_t)$	$\mathrm{ACF}_1(B_t)$	$\mathrm{ACF}_1(S_t)$
Mean	0.5365	0.5459	0.5227	0.5167	0.5227	0.5167
(Mean AE)			(0.0121)	(0.0113)	(0.1830)	(0.1845)
Median	0.5406	0.5501	0.5211	0.5307	0.5211	0.5302
(Median AE)			(0.0096)	(0.0094)	(0.1454)	(0.1583)
SD	0.0864	0.0829	0.2341	0.2363	0.2341	0.2363
(SD of AE)			(0.0113)	(0.0091)	(0.1366)	(0.1359)
Minimum	0.2791	0.3036	−0.0404	−0.0522	−0.0404	−0.0522
(Minimum AE)			(0.0000)	(0.0001)	(0.0006)	(0.0064)
Maximum	0.7082	0.7129	0.9387	0.9402	0.9387	0.9402
(Maximum AE)			(0.0671)	(0.0451)	(0.6059)	(0.6118)

Note: The sample includes 120 stocks that traded on the NYSE in 2010 and 2011. They are randomly selected from the S&P 500 Index, S&P MidCap 400 Index and S&P SmallCap 600 Index, with 40 stocks from each index. Panel A lists ticker symbols of the sample stocks. Panel B reports summary statistics of the contemporaneous correlation between buy and sell orders implied by sample data, the HMM, EHO, DY and EEOW models. Panel C reports the summary statistics of the sample autocorrelation function of lag 1 of buy order flows ($\mathrm{ACF}_1(B_t)$) and sell order flows ($\mathrm{ACF}_1(S_t)$) and those implied by the HMM and EEOW models. The summary statistics of the absolute error (AE) of the contemporaneous correlation or the first-order autocorrelation implied by the candidate models in comparison with those of sample data are reported in parentheses.

(1996) and EHO is that it generates a negative contemporaneous correlation between the numbers of buys and sells, while the correlation implied by trading data is positive. Panel B of Table I presents the summary statistics of the sample contemporaneous correlations between buys and sells and their counterparts implied by the four candidate models for the whole sample.[7] We can see that buys and

[7] The derivation of contemporaneous correlations implied by the EHO and DY models follows Duarte and Young (2009). For the EEOW model, replacing the static order arrival rates in the EHO model with the time-varying counterparts leads to an explicit formula of contemporaneous correlation. The HMM has a closed-form expression for contemporaneous correlation and is available upon request. As the EEOW model and the HMM produce time-varying contemporaneous correlations between buys and sells, the sample means are calculated for each stock.

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

sells are strongly and positively correlated for all the sample stocks, with contemporaneous correlations ranging from 0.5679 to 0.9574. However, correlations of buys and sells calculated based on the EHO and EEOW models are always negative, the reason being that these two models only consider two trading motives by ignoring SOS trading. As a result, buys and sells driven by private signals arrive on different days, creating a negative correlation between buys and sells. Although the DY model allows for positive correlations between buys and sells, it implies contemporaneous correlations ranging from 0.1783 to 0.358—significantly smaller than the ones implied by empirical data. In contrast, the contemporaneous correlations implied by the HMM is highly consistent with those of empirical data. Absolute error (AE) in panel B of Table I reports the difference between sample correlation and those implied by the candidate models. Apparently, the HMM yields the smallest absolute errors relative to the other three models.

Panel C of Table I presents summary statistics of the order flows' autocorrelation functions of lag 1 (i.e. ACF_1), implied by empirical data, the HMM and the EEOW model, respectively. For all sample stocks, strong and positive serial correlations are observed in both buy and sell order flows of empirical data with the sample ACF_1 ranging from 0.2791 to 0.7082. The ACF_1 implied by the HMM are all significantly positive and consistent with the sample ones, evidenced by the negligible absolute errors. The EEOW model can readily capture the serial dependence for some stocks in the sample. However, its absolute errors are still substantially larger than the HMM's. Both the EHO and DY models treat trading activities across different days as independent events and hence cannot generate serially dependent order flows.

Overall, only the HMM approach can accommodate both the positive contemporaneous correlation between buys and sells and the serial dependence of order flows with high accuracy.

4.4. Number of States

The number of states for each stock in the sample can be determined based on certain information criteria, such as AIC, and their summary statistics are depicted in panel A of Table II for the whole sample and the three size-based subsamples. Obviously, empirical data implies much more hidden states than the simulation scenarios. Comparing the results for the three sized-based subsamples, it is obvious that larger firms tend to have more hidden states in their stock trading. Furthermore, the difference between the means of any two subsamples is statistically significant at the 1% level, as shown by the *p*-values of the *t*-tests. Our estimation (not tabulated in Table II) shows that the correlation between the number of states and firm size is 0.5538. Since a larger firm is more likely to be a conglomerate and a smaller firm is more likely to be a standalone company, our result supports Cohen and Lou (2012), who argue that the standalone firm has straightforward information processing, whereas the information processing of a conglomerate firm tends to be complicated.

Share turnover can be used as a proxy for the liquidity and visibility of a stock (see Gervais *et al.*, 2001). To see its link to the number of states, we first sort the stocks into quartiles according to their turnovers over the sample period, and then analyze the cross-sectional differences in the number of states. Panel B of Table II shows that higher-turnover stocks are associated with a larger number of states. There are, on average, 27.87 states for high-turnover stocks in Q_4, in comparison to 23.63 states for low-turnover stocks in Q_1, and the difference is significant at the 1% level. We further perform *t*-tests for the differences between the means of any two quartiles. As shown by panel B, the mean of Q_1 is significantly different from that of Q_2 at the 5% level and that of Q_3 at the 1% level. This finding is consistent with the intuition that higher turnover stocks, being more liquid and visible, provide more potential for investors to learn about the firm's prospects and are more likely to be followed by investors. Therefore, investors more actively dig for those firms' private information, while their public information releases can induce a greater degree of belief updating by investors, which complicates the information structure. Panel B of Table II

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

Table II. The numbers of hidden states

Panel A: Number of states for the whole sample and the three size-based subsamples

	Mean	SD	25th percentile	Median	75th percentile
Whole sample	26.27	5.81	22	26	30
Large stock group	29.72	6.13	25.5	30	32
Medium stock group	26.47	4.38	24	26	29
Small stock group	22.62	4.52	19	22	26

t-test for the difference between the means of two subsamples

	Large – small	Large – medium	Medium – small
p-value	<0.0001	0.0043	<0.0001

Panel B: Number of states for the four subsamples sorted by share turnover or illiquidity measure

	Mean	SD	25th percentile	Median	75th percentile
Subsamples sorted by share turnover					
Q_1 (low turnover)	23.63	3.81	22	24	26
Q_2	26.16	3.85	24	25	30
Q_3	27.43	3.39	23	27	30
Q_4 (high turnover)	27.87	5.98	25	27.5	31
Subsamples sorted by illiquidity measure					
Q_1 (low illiquid)	28.1	5.92	23	28	31
Q_2	26.73	3.72	24	26.5	30
Q_3	26.56	3.56	24	26.5	29
Q_4 (highly illiquid)	23.7	3.74	22	24	27

t-test for the difference between the means of two quartiles

	$Q_4 - Q_1$	$Q_4 - Q_2$	$Q_4 - Q_3$	$Q_3 - Q_1$	$Q_3 - Q_2$	$Q_2 - Q_1$
p-value between subsamples sorted by share turnover	<0.0001	0.2592	0.7427	0.0012	0.188	0.0122
p-value between subsamples sorted by illiquidity measure	<0.0001	<0.0001	0.0038	0.2548	0.8343	0.2476

Panel C: Summary statistics of the number of hidden states belonging to a particular type

	Mean	SD	25th percentile	Median	75th percentile
liquidity information state	3.42	1.21	3	3	4
private information state	8.72	2.95	7	8	10
public information state	3	1.49	2	3	4
private and public information state	11.13	2.29	10	11	13

Panel D: Number of states for the whole sample estimated over different horizons

	Mean	SD	25th percentile	Median	75th percentile
Number of states estimated over different horizons					
2010	22.403	5.175	19.5	22	24.5
2011	22.312	4.458	19	23	25
2010 – 2011	26.27	5.81	22	26	30
Stationary probability of extreme states					
2010 vs. 2010–2011	0.0275	0.0305	0	0.0141	0.0492

Note: For each stock in the sample, its number of states is determined by AIC in the HMM. Panel A presents the summary statistics of the number of states for the whole sample and the three size-based subsamples, and the outcomes of *t*-tests for the difference between the means of two subsamples. Panel B documents the summary statistics of the number of states for the quartiles when stocks are sorted according to their share turnover or Amihud (2002) illiquidity measure. It also documents the results of mean difference tests of any two quartiles. For a particular stock, its hidden states can be classified into four types: liquidity state, private information state, public information state, and state with both private and public information. Panel C reports the summary statistics of the number of states for each type over the whole sample. Panel D reports the summary statistics of the number of states for the whole sample estimated over 2010, 2011 and 2010 – 2011. States detected in the 2-year horizon but not detected over 2010 are treated as extreme states. The last row of panel D reports the summary statistics of the stationary probability of extreme states.

Copyright © 2014 John Wiley & Sons, Ltd.

also reports the summary statistics of the number of states in the four subsamples sorted by the average daily Amihud (2002) illiquidity measure, with the p-values of testing the differences between the means of any two subsamples. More liquid stocks are associated with a larger number of states on average, consistent with the results implied by share turnover, and highly illiquid stocks in Q_4 have significantly less states than stocks in other three quartiles.

As mentioned previously, the type of hidden states can be identified after its associated expected numbers of buy and sell orders are decomposed into a maximum of three components (see equation (3)). Therefore, we can form four types of states: namely, liquidity state, private information state, public information state, and private and public information state.[8] Let us take the sample stock BCO as an example. It has 25 states in total. Among them, two states are liquidity states, eight states are associated only with privately informed trading, three states are associated only with public information trading, and the remaining 12 states include trading based on both private and disputable public information. The two liquidity states imply that variation in market condition is not always accompanied by the occurrence of information events. The market regime as characterized by expected numbers of trade imbalances and balanced trades can shift from one to another solely because of time-varying liquidity needs. The eight private information states of the BCO market demonstrate that private information can have a wide-ranging impact. The complexity of private information in terms of its varying content, dissemination channel and coverage leads to different mean levels of trade imbalances over time. In contrast, the pattern of the effects of disputable public information alone is relatively simple, evidenced by the fact that the number of public information states is only three. However, trading due to disputable public information often interacts with that due to private information, illustrated by 12 relevant hidden states in the market of BCO. As long as investors differ in their ability to interpret and process public news and differ in their priors, order flows still contain information that is not common knowledge.

Panel C of Table II reports the summary statistics of the numbers concerning the four types of states for the whole sample. For most sample stocks, the number of liquidity states is larger than one, reflecting sizable variation of liquidity needs over time. A significant portion of states are classified as states with private information but without disputable public information. The nature of a firm's private signals tends to be more complicated than that of public information. For instance, a private signal can be positive or negative. Informed investors may receive an extremely good signal of a company or it is just slightly better than expected. The private signal can be observed by either a very limited number of investors or many investors. This variation and complexity induce more private information states than public information states. The mean number of states with both private and public information is 11.13, which is consistent with Kim and Verreccia's (1994) argument that some informed market participants (e.g. financial analysts, large shareholders) can efficiently process the released public information into private information. Overall, the analysis of state numbers shows that the information environment of securities trading is complex and cannot be fully described by three states, as suggested by the EHO model, or six states, as suggested by the DY model.

To test the robustness of the parameter estimates, we split the estimation period of two years ($2010-2011$) into two horizons of 1 year and panel D of Table II reports the summary statistics of the number of states for each stock estimated over the three horizons.[9] On average, more states are detected in the extended period. For the case of extending the estimation window from 2010 to 2010 and 2011, more states are detected for 80 sample stocks, while the numbers of states of the rest 40 stocks remain unchanged. Although extending the estimation horizon may introduce additional

[8] Since liquidity trading exists in all states, state (i, j) is called liquidity state if $\mu_{b;i}=\mu_{s;j}=\nu_{b;i}=\nu_{s;j}=0$, private information state if $\mu_{b;i}+\mu_{s;j}\neq 0$ and $\nu_{b;i}=\nu_{s;j}=0$, public information state if $\mu_{b;i}=\mu_{s;j}=0$ and $\nu_{b;i}+\nu_{s;j}\neq 0$, private and public information state if $\mu_{b;i}+\mu_{s;j}\neq 0$ and $\nu_{b;i}+\nu_{s;j}\neq 0$.

[9] Limited by the scope, the detailed results of robustness checks are not reported. They and the confidence intervals of parameter estimates through parameter bootstrap (see Efron and Tibshirani, 1993) are available upon request.

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

noises, the union set of order arrival rates estimated by the two 1-year horizons largely overlaps the order arrival rates estimated by the 2-year horizon. This suggests that the state structure of the HMM is somewhat sensitive to the coverage of the estimation horizon and, when the length of the horizon extends, the HMM is flexible enough to accommodate the regime changes by introducing new states. Let us call the states estimated over the 2-year horizon (2010 – 2011) extreme states if their order arrival rates are outside those of states estimated over 2010. The last row of panel D documents the statistics of the stationary probabilities associated with those extreme states. The mean stationary probability is 0.0275, which is significant at the 1% level. This demonstrates that the new states introduced by the HMM over the extended estimation horizon are non-trivial.

4.5. Out-of-Sample Performance of the HMM

The HMM has demonstrated its outstanding performance through in-sample analysis. This subsection adopts two ways to evaluate its out-of-sample forecasting performance relative to the other three models. First, we examine the accuracy of forecast PIN and PSOS. Since the true daily measures of PIN and PSOS are not observable, we treat the in-sample measures estimated by the HMM over the sample period of 2010–2011 as the benchmark measures. We choose this benchmark for two reasons. First, the HMM’s superior in-sample performance has been well demonstrated in simulation studies. Second, of the four candidate models, only HMM can accommodate both contemporaneous correlations and serial correlations between empirical order flows, as discussed in Section 4.3. For each trading day in 2011, we estimate each of the four candidate models using the preceding year’s trading data and then forecast the out-of-sample PIN and PSOS measures on that day. Therefore, we have 251 1-day forecasts of PIN and PSOS from each candidate model for each sample stock. From panel A of Table III, we can see that the HMM results in the highest correlation between the 1-day forecasts and benchmark measures, across all the summary statistics for both PIN and PSOS. Panel B of Table III further counts the number of stocks with a positive and significant correlation coefficient between the 1-day forecast and benchmark measure; it also demonstrates again that the HMM performs the best. For instance, at the 10% significance level, the PIN forecast by the HMM is positively correlated to the benchmark measure for 99 out of the 120 sample stocks, while the other three models achieve this significance level for fewer than 30 stocks. The 1-day forecast of PSOS from the HMM is positively correlated to the benchmark measure at the 1% significance level for 114 out of 120 sample stocks, while the DY model achieves such performance only for 17 stocks. Furthermore, from panel C of Table III we can see that the absolute error of HMM’s 1-day forecast is the smallest of the candidate models for both PIN and PSOS across all summary statistics.

One potential criticism of the above performance evaluation is its usage of the HMM in generating in-sample PIN and PSOS. To address this concern, we use the forecast pseudo-residuals of buys and sells to evaluate the out-of-sample performance of the candidate models following Dunn and Smyth (1996) and Zucchini and MacDonald (2009). The detailed construction of the forecast pseudo-residual segments is shown in Section S.4 of the supplementary Appendix. If the segment on trading day t is extreme, say lying entirely within the top or bottom 0.5% of the standard normal distribution, the observation $X_t = x_t$ is an outlier or the candidate model no longer adequately describes the underlying time series. For each sample stock, we first calculate its normal forecast pseudo-residual for buys and sells respectively on each trading day of 2011, using each of four candidate models that are fitted by trading data from the preceding year. We then count the number of trading days whose forecast pseudo-residual segment lies entirely within the top or bottom 0.5% of the standard normal distribution. The summary statistics reported in panel D of Table III demonstrate that the HMM outperforms the other models across all summary statistics. Let us take the ’Median’ statistics as an example. Over 251 trading days of the forecasting period, the HMM produces extreme forecast pseudo-residuals on

Copyright © 2014 John Wiley & Sons, Ltd.

Table III. Out-of-sample forecasting performance of the four candidate models

Panel A: Summary statistics of correlations between the in-sample and out-of-sample daily measures

	5th percentile	25th percentile	Median	75th percentile	95th percentile
$\mathrm{corr}\left(\mathrm{PIN}_{i,t}^{\mathrm{IS}}, \mathrm{PIN}_{i,t}^{\mathrm{OS}}\right)$					
HMM	−0.016	0.109	0.239	0.368	0.498
EHO model	−0.208	−0.084	−0.017	0.064	0.201
DY model	−0.163	−0.038	0.021	0.080	0.180
EEOW model	−0.138	−0.063	−0.008	0.046	0.148
$\mathrm{corr}\left(\mathrm{PSOS}_{i,t}^{\mathrm{IS}}, \mathrm{PSOS}_{i,t}^{\mathrm{OS}}\right)$					
HMM	0.535	0.711	0.784	0.824	0.881
DY model	−0.419	−0.193	−0.076	0.057	0.301

Panel B: Number of sample stocks with positive correlation between daily forecasts and in-sample measures which are significant at various levels

	$\mathrm{corr}\left(\mathrm{PIN}_{i,t}^{\mathrm{IS}}, \mathrm{PIN}_{i,t}^{OS}\right) > 0$			$\mathrm{corr}\left(\mathrm{PSOS}_{i,t}^{\mathrm{IS}}, \mathrm{PSOS}_{i,t}^{\mathrm{OS}}\right) > 0$		
	10% level	5% level	1% level	10% level	5% level	1% level
HMM	99	92	85	114	114	114
EHO model	25	20	12			
DY model	29	21	11	26	24	17
EEOW model	18	11	7			

Panel C: Absolute errors of the daily forecasts with respective to the in-sample measures

	5th percentile	25th percentile	Median	75th percentile	95th percentile
$\left\lvert\mathrm{PIN}_{i,t}^{\mathrm{IS}} - \mathrm{PIN}_{i,t}^{\mathrm{OS}}\right\rvert$					
HMM	0.010	0.016	0.020	0.027	0.045
EHO model	0.027	0.044	0.054	0.066	0.092
DY model	0.018	0.030	0.037	0.046	0.061
EEOW model	0.053	0.078	0.094	0.107	0.159
$\left\lvert\mathrm{PSOS}_{i,t}^{\mathrm{IS}} - \mathrm{PSOS}_{i,t}^{\mathrm{OS}}\right\rvert$					
HMM	0.032	0.042	0.053	0.065	0.158
DY model	0.088	0.106	0.129	0.159	0.257

Panel D: Summary statistics of the number of trading days with extreme forecast pseudo-residuals in buys and sells

	5th percentile	25th percentile	Median	75th percentile	95th percentile
Buys					
HMM	10.5	19	24	30	41.5
EHO model	51	78	96.5	117.5	137
DY model	19.5	33	43	58	81
EEOW model	41	73	89.5	101.5	116
Sells					
HMM	10.5	19	23	29.5	40.5
EHO model	53	75	90	107.5	133.5
DY model	20	32.5	39	48.5	77.5
EEOW model	40.5	76	94	105	116.5

Note: This table reports the out-of-sample forecasting performance of the four candidate models. $\mathrm{PIN}_{i,t}^{\mathrm{IS}}$ and $\mathrm{PSOS}_{i,t}^{\mathrm{IS}}$ denote the in-sample (IS) PIN and PSOS on trading day *t* for sample stock *i*, estimated by the HMM over the sample period of 2010–2011. $\mathrm{PIN}_{i,t}^{\mathrm{OS}}$ and $\mathrm{PSOS}_{i,t}^{\mathrm{OS}}$ denote the out-of-sample (OS) 1-day forecasts of PIN and PSOS predicted by one of the four candidate models using trading data of the preceding year. The forecasting range covers all trading days in 2011. For each sample stock and each candidate model, the correlation coefficients between in-sample and out-of-sample measures are denoted by $\mathrm{corr}\left(\mathrm{PIN}_{i,t}^{\mathrm{IS}}, \mathrm{PIN}_{i,t}^{\mathrm{OS}}\right)$ and $\mathrm{corr}\left(\mathrm{PSOS}_{i,t}^{\mathrm{IS}}, \mathrm{PSOS}_{i,t}^{\mathrm{OS}}\right)$. The summary statistics of the correlations are reported in panel A. Panel B counts the number of sample stocks with positive correlation significant at the 10%, 5% and 1% levels, respectively. Panel C documents summary statistics of absolute errors of daily forecasts. For each sample stock, the forecast pseudo-residual segments of buys and sells are calculated according to the supplementary Appendix for each trading day in 2011. The number of trading days with segments lying outside the bands of 0.5% and 99.5% of the standard normal distribution is counted, of which the summary statistics are reported in panel D.

24 (23) trading days for buy (sell) order flows, while the other three models produce much more extreme forecast pseudo-residuals, ranging from 39 to 96.5 trading days. In the out-of-sample period of 1 year, the HMM reasonably predicts order flow distributions on more than 90% of the trading days,

Copyright © 2014 John Wiley & Sons, Ltd.

and produces the smallest number of extreme order flows among the four candidate models. Thus we can be more comfortable with the HMM for describing the trading dynamics.

5. FURTHER APPLICATIONS OF THE HMM APPROACH

5.1. Information-Based Trading around Earnings Announcements

We identify quarterly earnings announcements using the announcement dates and times recorded in the Thomas Reuters I/B/E/S database. Over the 2 years (2010 and 2011) there are 960 earnings announcements for the 120 sample companies we mentioned in the previous section. Announcements occurring at or after 4:00 p.m. are relabeled with the following day's date to ensure that the event day is the day on which investors and stock prices have time to react to the earnings announcement. For Earnings Surprise (ES) measure we require that there is at least one observation in the I/B/E/S database for calculating the mean of analyst forecasts prior to an earnings announcement. This screens out five announcements and leads to a final sample with 955 earnings announcements. We consider an event window of 21 days,[10] with day 0 denoting the announcement day and 10 trading days on each side. ES is defined as the difference between actual and expected earnings per share and it is then scaled by the stock price of 15 days before the announcement. The consensus analyst earnings forecast just prior to the announcement proxies the market's expectation because Brown and Caylor (2005) find that this earnings benchmark is most strongly associated with the valuation premium of earnings surprises.

We regress daily PIN and PSOS on event day dummies through panel regressions similar to Patton and Verardo (2012):

$$\mathrm{PIN}_{i,t} = \sum\nolimits_{k=-10}^{10} D_k^A I_i(t-k) + \delta_i^A + \varepsilon_{i,t},\ \mathrm{PSOS}_{i,t} = \sum\nolimits_{k=-10}^{10} D_k^S I_i(t-k) + \delta_i^S + \varepsilon_{i,t} \qquad (10)$$

where $I_i(t-k)$ is an event day dummy variable of firm i with $I_i(t-k) = 1$ if $t-k=0$ and $I_i(t-k) = 0$ if $t-k \neq 0$. The levels of average information-based trading outside of the event window are captured by the firm fixed effects δ_i^A and δ_i^S. The relative measures of PIN and PSOS around earnings announcements can be detected respectively by coefficients D_k^A and D_k^S. The regressions specified by equation (10) are estimated for the whole sample, and the subsample with the top 10% positive ES. Table IV and the upper diagram in Figure 1 show that on day 0 PIN experiences a plunge of 0.0098 for the whole sample. Noting that average δ_i^A is 0.0846, the plunge thus represents a more than 10% reduction of information asymmetry on the announcement day. Since earnings announcements cannot effectively reduce the portion of information asymmetry irrelevant to earnings news, PIN reverts back from day 1 and differs insignificantly from its average level outside the event window. This pattern of PIN is consistent with the argument that public news can resolve information asymmetry in securities trading (e.g. Tetlock, 2010). However, the effect may be short lived. Some investors more able to assess the firm's performance on the basis of the announcements may quickly regain information advantage, as predicted by Kim and Verrecchia (1994).

In addition to easing information asymmetry, an earnings announcement is likely to be interpreted differently by investors and thus significantly affect belief dispersion among investors. Table IV and the lower diagram in Figure 1 show that for the whole sample PSOS experiences a sharp increase of 0.3136 on the event day but an immediate drop on day 1, and then remains significantly high over the post-announcement period with a downward trend towards δ_i^S. Noting that average δ_i^S over the whole sample is 0.1246, the surge of PSOS on day 0 makes it more than double the average level of δ_i^S.

When compared to the results from the whole sample, Figure 1 and Table IV indicate that the variations in PIN and PSOS are greater for the subsample with the top 10% positive ES. This implies that

[10] For robustness checks, we have considered different lengths of the event window and the results are similar.

Copyright © 2014 John Wiley & Sons, Ltd.

Table IV. Information-based trading around earnings announcements

Panel A: Estimated relative measures of information-based trading

	Whole sample				Subsample of the top 10% positive ES			
Day (k)	$\widehat{D_k^A}$	(t-stat.)	$\widehat{D_k^S}$	(t-stat.)	$\widehat{D_k^A}$	(t-stat.)	$\widehat{D_k^S}$	(t-stat.)
−10	−0.01%	(−0.031)	−1.79%***	(−3.535)	1.52%**	(2.338)	−1.95%*	(−1.700)
−9	−0.04%	(−0.150)	−2.63%***	(−5.219)	−1.22%*	(−1.839)	−1.40%	(−1.312)
−8	−0.40%	(−1.488)	−2.99%***	(−6.088)	−1.00%	(−0.879)	−0.38%	(−0.244)
−7	−0.01%	(−0.043)	−2.22%***	(−4.461)	−0.26%	(−0.225)	−2.22%	(−1.807)
−6	0.04%	(0.146)	−1.89%***	(−3.569)	0.98%	(1.262)	−2.36%*	(−1.704)
−5	0.10%	(0.316)	−0.01%	(−0.008)	1.90%**	(2.511)	0.29%	(0.183)
−4	0.05%	(0.154)	−0.75%	(−1.515)	0.68%	(0.626)	0.87%	(0.564)
−3	0.04%	(0.120)	−0.22%	(0.399)	−0.35%	(−0.308)	0.58%	(0.345)
−2	0.26%	(0.915)	1.56%***	(2.699)	0.87%	(0.843)	4.92%***	(2.892)
−1	−0.12%	(−0.465)	7.55%***	(10.946)	0.48%	(0.529)	11.65%***	(6.226)
0	−0.98%***	(−3.354)	31.36%***	(22.940)	−1.49%**	(−2.439)	36.12%***	(12.053)
1	−0.10%	(−0.327)	17.19%***	(21.431)	−0.26%	(−0.242)	20.59%***	(9.475)
2	0.61%*	(1.801)	11.23%***	(15.936)	1.01%*	(1.920)	13.56%***	(6.359)
3	0.28%	(0.921)	8.24%***	(12.271)	1.63%	(1.444)	8.81%***	(4.678)
4	0.33%	(1.006)	7.03%***	(12.270)	−0.50%	(−0.521)	8.14%***	(3.917)
5	0.08%	(0.259)	6.86%***	(12.029)	0.32%	(0.268)	8.73%***	(4.533)
6	0.08%	(0.246)	5.87%***	(9.165)	0.58%	(0.649)	7.92%***	(4.295)
7	0.52%*	(1.695)	4.58%***	(8.016)	0.60%	(0.617)	3.67%**	(2.210)
8	0.46%	(1.348)	4.41%***	(7.738)	−0.05%	(−0.049)	5.99%***	(3.015)
9	0.41%	(1.183)	4.46%***	(8.646)	1.12%*	(1.973)	3.84%***	(2.686)
10	0.35%	(0.960)	3.41%***	(6.483)	0.72%	(0.703)	2.73%*	(1.786)

Panel B: Average measures of information-based trading outside of the event window

	Whole sample	Large stock group	Medium stock group	Small stock group
Ave. δ_i^A	8.46%	6.85%	8.13%	10.39%
Ave. δ_i^S	12.46%	10.19%	13.52%	13.67%

Panel C: hypothesis tests on the equality of pre-event and post-event relative measures

H_0: pre-event relative PIN is equal to post-event relative PIN, i.e. ($\sum_{k=-10}^{-1} D_k^A = \sum_{k=1}^{10} D_k^A$).

p-value is equal to 0.3905 for the whole sample, and 0.0561 for the subsample of the top 10% positive ES.

H_0: pre-event relative PSOS is equal to post-event relative PSOS, i.e. ($\sum_{k=-10}^{-1} D_k^S = \sum_{k=1}^{10} D_k^S$).

p-value is less than 0.0001 for both the whole sample and the subsample of the top 10% positive ES.

Note: Panel A presents the estimated daily measures of information-based trading for the 21 trading days around quarterly earnings announcements, computed as the difference with respect to the average measures outside of the event window. In particular, the estimates are obtained from panel regressions, $\text{PIN}_{i,t} = \sum_{k=-10}^{10} D_k^A I_i(t-k) + \delta_i^A + \varepsilon_{i,t}$ and $\text{PSOS}_{i,t} = \sum_{k=-10}^{10} D_k^S I_i(t-k) + \delta_i^S + \varepsilon_{i,t}$, where $I_i(t-k)$ is an event day dummy variable of firm i with $I_i(t-k) = 1$ if $t-k=0$ and $I_i(t-k) = 0$ if $t-k \neq 0$. Therefore, $\widehat{D_k^A}$ and $\widehat{D_k^S}$ denote the relative measures of PIN and PSOS, respectively. The left part of panel A is of the whole sample, while the right part is of the subsample with the top 10% positive Earnings Surprise (ES), defined as the difference between actual quarterly earnings and the consensus of analyst forecasts, is scaled by the stock price. Asterisks indicate significance at the *10%, **5% and ***1% levels. Panel B reports the average measures of information-based trading outside of the event window for the whole sample and the three size-based subsamples. Panel C reports the p-value of the hypothesis tests on the equality of pre-event and post-event relative PIN (PSOS) for the whole sample and the subsample of the top 10% positive ES.

larger earnings surprises are more likely to be preceded with pre-event privately informed trading and to cause investor disagreement on and after the event day.

In order to examine the overall effect of earnings announcements, we perform hypothesis tests on the equality of pre- and post-announcement relative measures of PIN and PSOS, of which the p-values are reported in panel C of Table IV. Although for the whole sample the pre-event PIN is not significantly different from its post-event counterpart, the p-value of the same test for the subsample with the top 10% positive ES is 0.0561, which demonstrates the effect of ES on changes in information asymmetry. Since information asymmetry in the pre-event period is highly likely to be relevant to the earnings news, the announcements with a larger ES tend to be more effective in reducing information

Copyright © 2014 John Wiley & Sons, Ltd.

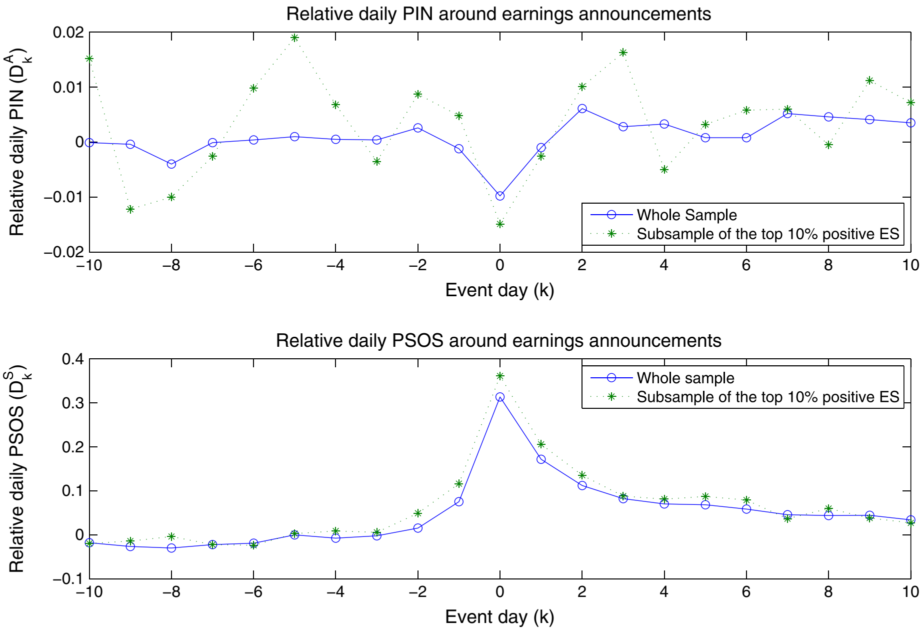

Figure 1. PIN and PSOS around earnings announcements. This figure presents the cross-sectional averages of relative PIN and PSOS over the 21 days around earnings announcements for the whole sample and the subsample with the top 10% positive earnings surprises (ES). Relative PIN and PSOS are defined by the estimated coefficients of the dummies, D_k^A and D_k^S, from the two panel regressions: $\text{PIN}_{i,t} = \sum_{k=-10}^{10} D_k^A I_i(t-k) + \delta_i^A + \varepsilon_{i,t}$ and $\text{PSOS}_{i,t} = \sum_{k=-10}^{10} D_k^S I_i(t-k) + \delta_i^S + \varepsilon_{i,t}$, where $I_i(t-k)$ is an event day dummy variable of firm i with $I_i(t-k) = 1$ if $t-k = 0$ and $I_i(t-k) = 0$ if $t-k \neq 0$

asymmetry. The equality of pre-event and post-event relative PSOS is rejected at the 1% level for the whole sample and the subsample considered. It is evident that a substantial effect of earnings announcements will trigger disagreement among investors of all stocks.

To further examine the association of PIN and PSOS with ES, we sort earnings announcements into quintiles according to ES. Panel A of Table V shows that the majority of earnings announcements in our sample exceed the consensus earnings forecast and the second quintile (Q_2) is identified as the one with nearly no earnings surprises.

In the analysis, we consider the standardized measures of PIN and PSOS defined as $\text{PIN}_{i,[t_1,,t_2]}/\text{PIN}_{i,\text{non}}$ and $\text{PSOS}_{i,[t_1,,t_2]}/\text{PSOS}_{i,\text{non}}$, where subscript $[t_1, t_2]$ indicates days related to an earnings announcement and 'non' indicates the period outside the 21-day event window. As reported in panel B of Table V, the mean standardized PIN for $[-10, -1]$ is 1.064 for the whole sample, which is significantly larger than 1. This implies that information asymmetry in stock trading increases by 6.4% prior to earnings announcements. In other words, the PIN measure estimated by the HMM approach does not experience the anomaly documented by Benos and Jochec (2007) that the PIN of EHO in the pre-announcement period is smaller than its post-announcement counterpart. Since it is possible that pre-announcement information about earnings news can be leaked, we expect that stocks in the extreme ES quintiles have higher pre-event standardized PIN because using private information to speculate these stocks is more profitable. Panel B of Table V shows that this is the case, although only the pre-event standardized PIN with extreme positive surprises (Q_5) is significantly greater than 1. The insignificance of standardized PIN for Q_1 (extreme negative ES) is probably because short selling constraints restrict some privately informed traders to sell stocks

Table V. Standardized PIN and PSOS around earnings announcements

Panel A: Earnings surprises of the whole sample and the five quintiles

	Whole sample	Q_1	Q_2	Q_3	Q_4	Q_5
Mean	0.13%***	-0.45%***	0.00%	0.08%***	0.20%***	0.63%***
(t-stat.)	(4.320)	(-7.654)	(-1.288)	(41.791)	(62.970)	(19.691)

Panel B: Standardized PIN around earnings announcements

				Time period $[t_1, t_2]$		
		[−10, −1]	[−5, −1]	Event day	[1, 5]	[1, 10]
Whole sample	Mean	1.064*	1.066	0.874***	1.012	1.024
	(*t*-stat.)	(1.890)	(1.296)	(−4.251)	(0.651)	(1.455)
Q_1, negative ES	Mean	1.042	1.033	0.848***	1.048	1.063*
	(*t*-stat.)	(1.506)	(0.895)	(−2.416)	(1.238)	(1.889)
Q_2, nearly no ES	Mean	1.016	0.991	0.930	1.016	1.030
	(*t*-stat.)	(0.530)	(-0.264)	(−1.087)	(0.461)	(1.047)
Q_3	Mean	0.986	1.003	0.873**	0.979	0.987
	(*t*-stat.)	(-0.604)	(0.100)	(−2.081)	(-0.637)	(-0.521)
Q_4	Mean	1.000	1.014	0.909**	0.993	0.999
	(*t*-stat.)	(-0.007)	(0.404)	(−2.400)	(-0.187)	(-0.022)
Q_5, positive ES	Mean	1.276*	1.293**	0.804**	1.035	1.042
	(*t*-stat.)	(1.973)	(2.173)	(−2.528)	(0.588)	(0.451)

Panel C: Standardized PSOS around earnings announcements

				Time periods $[t_1, t_2]$		
		[−10,−1]	[−5,−1]	Event day	[1,5]	[1,10]
Whole sample	Mean	0.824***	0.894***	2.183***	1.354***	1.245***
	(*t*-stat.)	(−8.328)	(−4.358)	(24.937)	(11.928)	(8.999)
Q_1, negative ES	Mean	0.841***	0.917	2.236***	1.533***	1.381***
	(*t*-stat.)	(−3.303)	(−1.503)	(11.332)	(7.105)	(5.686)
Q_2, nearly no ES	Mean	0.796***	0.855**	2.061***	1.249***	1.202***
	(*t*-stat.)	(−4.091)	(−2.563)	(10.621)	(3.627)	(3.160)
Q_3	Mean	0.864***	0.914	2.124***	1.285***	1.177***
	(*t*-stat.)	(−3.000)	(−1.626)	(12.267)	(5.261)	(3.378)
Q_4	Mean	0.806***	0.888**	2.138***	1.301***	1.192***
	(*t*-stat.)	(−4.545)	(−2.244)	(9.673)	(4.986)	(3.258)
Q_5, positive ES	Mean	0.813***	0.897*	2.376***	1.404***	1.279***
	(*t*-stat.)	(−3.606)	(−1.714)	(12.385)	(5.745)	(4.565)

Note: Earnings announcements are sorted into quintiles according to the associated Earnings Surprise (ES), defined as the difference between the actual earnings and the consensus of analyst forecasts, scaled by the stock price of 15 days prior to the announcement. Panel A presents the sample means of ESs for the whole sample and the quintiles. Panel B (panel C) reports the sample means of standardized PINs (PSOSs), which is defined as $\mathrm{PIN}_{i,[t_1,,t_2]}/\mathrm{PIN}_{i,\mathrm{non}}\left(\mathrm{PSOS}_{i,[t_1,,t_2]}/\mathrm{PSOS}_{i,\mathrm{non}}\right)$, where subscript $[t_1, t_2]$ indicates days relative to an earnings announcement and 'non' indicates the days outside the 21-day event window of the announcement. Asterisks indicate the value being significantly different from zero in panel A and from one (unit) in panels B and C at the *10%, **5% and ***1% levels.

short with forthcoming bad news.[11] This conjunction is also supported by the finding of Engelberg *et al.* (2012) that short sellers generally do not anticipate public news announcements.

On the event day, the standardized PIN is less than 1 but remains larger than 0.8 for all ES quintiles. The public disclosure cannot eliminate all information asymmetry probably because earnings announcements can only remove private information related to earnings news. For the post-event period, stocks with extreme earnings surprises (Q_1 and Q_5) have a larger standardized PIN than others. Note that PIN is not exclusively an insider trading measure since it also captures informed trading by investors who are particularly skillful in analyzing public news. Thus our finding is likely to be the result of the increased incentives to search for private information and gain superior information for greater ES, which is consistent with the view of Barron *et al.* (2008). In contrast to the case of pre-announcement period, however, it

[11] For instance, Almazan *et al.* (2004) document that two-thirds of US equity mutual funds restrict their managers from selling stock short and only 3% of all mutual funds actually engage in short selling.

is the post-event standardized PIN for Q_1 that tends to be significantly greater than 1 and larger than that for Q_5. We conjecture that market participants may lose confidence in the management of negatively surprising companies and increase their digging of private information more actively, which consequently leads to stocks in Q_1 having more privately informed trading over the post-event period. As such information searching takes time we notice that standardized PIN for Q_1 is larger and more significant over a longer period [1, 10] than that over a shorter period [1, 5].

Bamber (1987) argues that more surprising or informative announcements are likely to spawn a wide variety of interpretations. This is supported by the standardized PSOS on and after announcement days documented in panel C of Table V. The five ES quintiles show a general pattern of a greater divergence of beliefs among investors following a greater ES.

5.2. Co-movements in Information-Based Trading

The co-movements in information-based trading are closely related to systematic risk but the relationship is difficult to explore unless dynamic measures of information-based trading are employed. Co-movement in our analysis is characterized by the co-exceedance measure proposed by Bae *et al.* (2003). We adopt co-exceedance instead of correlation because the latter gives an equal weight to all trading activities, irrespective of their magnitudes, which is not appropriate for evaluating differential impacts owing to large variations in information-based trading. Co-exceedance is a measure of contagion capturing the coincidence of extreme shocks across stocks. Let us focus for a moment on

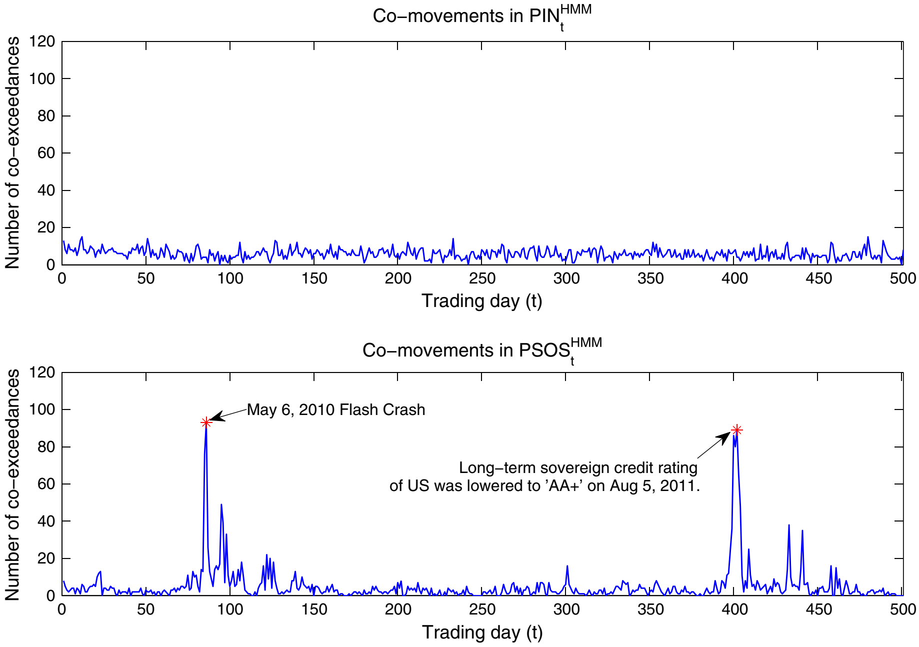

Figure 2. The numbers of co-exceedances in daily measures of information-based trading in 2010 and 2011. For stock *i*, the occurrence of extreme measure of $\mathrm{PIN}_{i,t}^{\mathrm{HMM}}$ ($\mathrm{PSOS}_{i,t}^{\mathrm{HMM}}$) is defined as an exceedance if it lies above the 95th percentile of its distribution. Co-movement is measured by the joint occurrences of exceedances, i.e. the number of co-exceedances of $\mathrm{PIN}_{\cdot,t}^{\mathrm{HMM}}$ ($\mathrm{PSOS}_{\cdot,t}^{\mathrm{HMM}}$), in sample stocks on a particular day. The upper and lower diagrams plot the time series of the numbers of co-exceedances in $\mathrm{PIN}_{i,t}^{\mathrm{HMM}}$ and $\mathrm{PSOS}_{i,t}^{\mathrm{HMM}}$, respectively, over the 120 sample stocks for 2010 and 2011. The Flash Crash on 6 May 2010 featured the biggest 1-day point decline (998.5 point) in the history of the Dow Jones Industrial Average. On 5 August 2011, Standard & Poor's lowered the long-term sovereign credit rating of the US to AA+

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

the occurrence of extreme measures of privately informed trading. For stock i, an extreme PIN occurs on day t if $\mathrm{PIN}_{i,t}^{\mathrm{HMM}}$ lies above the 95th percentile of its distribution. We then count the number of stocks with joint occurrences of extreme $\mathrm{PIN}_{\cdot,t}^{\mathrm{HMM}}$, or co-exceedances, on day t. The upper diagram in Figure 2 plots the time series of the number of co-exceedances in $\mathrm{PIN}_{i,t}^{\mathrm{HMM}}$ in 2010 and 2011. There does not exist any trading day with the number of co-exceedances larger than 20. Since there are 120 stocks in the sample, the possibility of market-wide co-movements in PIN seems very low. In other words, we cannot see significant market-wide trading shocks caused by private information according to the measure of co-exceedances.

In contrast to PIN, some trading days have a quite large number of co-exceedances in PSOS. The lower diagram in Figure 2 plots the number of co-exceedances in $\mathrm{PSOS}_{i,t}^{\mathrm{HMM}}$, which is much more volatile than that in $\mathrm{PIN}_{i,t}^{\mathrm{HMM}}$. There are six trading days with the number of co-exceedances in $\mathrm{PSOS}_{i,t}^{\mathrm{HMM}}$ larger than 60, i.e. half of the number of sample stocks. On 5 August 2011, Standard & Poor's lowered the long-term sovereign credit rating of the US to AA + from AAA. On that day and the following 2 days, the numbers of co-exceedances in $\mathrm{PSOS}_{i,t}^{\mathrm{HMM}}$ were extremely large, representing the 2nd, 4th and 3rd highest co-exceedances in the 2-year sampling period, with an average number of daily co-exceedances of 81.67. The Flash Crash on 6 May 2010 featured the biggest 1-day point decline (998.5 points) in the history of the Dow Jones Industrial Average. The number of co-exceedances on that day was 73, ranking 5th over the sampling period. The day following the Flash Crash experienced the largest number of co-exceedances in $\mathrm{PSOS}_{i,t}^{\mathrm{HMM}}$ and reached 88. These widespread co-movements in trading activities by and large were triggered by hugely different opinions about the public news events. On days after public news releases, more two-sided markets were observed, especially when the news surprises were profound and controversial.

6. CONCLUDING REMARKS

This study presents a novel hidden Markov model of information-based trading, incorporating both asymmetric information and symmetric order-flow shocks. The hidden state reflects the fundamentals of trading activities and the information environment of these activities. Thus the trading dynamics can be completely characterized by a time-varying distribution of states. The states can be classified into different types according to whether they feature privately informed trading and/or trading due to disputable public information. The model is sufficiently flexible to cope with the variations in information asymmetry and belief heterogeneity over time and across companies.

We first use extensive Monte Carlo simulation experiments to demonstrate the superior performance of HMM in identifying hidden states and measuring information-based trading. Using a sample of 120 NYSE stocks, we further show the impressive effectiveness of the HMM approach and its dynamic PIN and PSOS in describing and characterizing the trading process. By examining the behavior of PIN and PSOS around the earnings announcements of stocks, we provide insights on the market's information environment and its variation before and after the announcements. The most recognizable findings include that PIN plunges on the announcement day but quickly reverts to its average level, while PSOS surges on the announcement day and then gradually trends down. Interestingly, the PIN obtained from the HMM approach does not suffer the anomaly of pre-announcement PIN being greater than post-announcement PIN as documented in previous studies. More importantly, we find that earnings announcements do cause considerable disagreement among market participants.

Turning to trading co-movement, there is no substantial evidence showing trading activities due to private information synchronizing across stocks. However, co-movements of trading due to disputable public information are evident and often peaked around the time when a critical and economy-wide event occurs.

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

The hidden Markov model approach is likely to provide an effective and flexible platform for the analysis of information-based trading, as demonstrated by our studies of earnings announcements and trading co-movements. Other applications of the HMM approach, particularly in event studies, are of great potential and remain to be explored further.

ACKNOWLEDGEMENTS

The authors thank constructive comments and suggestions from two anonymous referees, Baibing Li, Ming Fang, Nagaratnam Jeyasreedharan, Paul Kim, Lily Nguyen, Sandy Suardi, Terry Walter, and participants of seminars at the University of Sydney, the University of Melbourne, Monash University, the University of Adelaide, University of Technology Sydney and University of Tasmania, the 7th World Congress of the Bachelier Finance Society and 2012 FIRN Conference. The project is partly supported by the Discovery Projects funding from Australian Research Council (DP140100113).

REFERENCES

Almazan A, Brown KC, Carlson M, Chapman DA. 2004. Why constrain your mutual fund manager? *Journal of Financial Economics* **73**: 289–321.

Amihud Y. 2002. Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets* **5**: 31–56.

Bae K, Karolyi GA, Stulz RM. 2003. A new approach to measuring financial contagion. *Review of Financial Studies* **16**: 717–763.

Bamber L. 1987. Unexpected earnings, firm size and trading volume around quarterly earnings announcements. *Accounting Review* **62**: 510–32.

Barron OE, Byard D, Yu Y. 2008. Earnings surprises that motivate analysts to reduce average forecast error. *Accounting Review* **83**: 303–326.

Benos E, Jochec M. 2007. Testing the Pin variable. Working paper, University of Illinois at Urbana-Champaign.

Brown L, Caylor M. 2005. A temporal analysis of thresholds: propensities and valuation consequences. *Accounting Review* **80**: 423–440.

Burnham K, Anderson D. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer: New York.

Cappé O, Moulines O, Rydén T. 2005. *Inference in Hidden Markov Models*. Springer: New York.

Chakrabarty B, Moulton PC, Shkilko A. 2012. Short sales, long sales, and the Lee–Ready trade classification algorithm revisited. *Journal of Financial Markets* **15**: 467–491.

Cohen L, Lou D. 2012. Complicated firms. *Journal of Financial Economics* **104**: 383–400.

Duarte J, Young L. 2009. Why is PIN priced? *Journal of Financial Economics* **91**: 119–138.

Dunn P, Smyth G. 1996. Randomized quantile residuals. *Journal of Computational and Graphical Statistics* **5**: 236–244.

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

Easley D, Hvidkjaer S, O'Hara M. 2002. Is information risk a determinant of asset returns? *Journal of Finance* **57**: 2185–2221.

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

Efron B, Tibshirani R. 1993. *An Introduction to the Bootstrap*. Chapman & Hall: New York.

Engelberg E, Adam R, Ringgenberg M. 2012. How are shorts informed? Short sellers, news, and information processing. *Journal of Financial Economics* **105**: 260–278.

Gervais S, Kaniel R, Mingelgrin D. 2001. The high-volume return premium. *Journal of Finance* **56**: 877–919.

Kandel E, Pearson N. 1995. Differential interpretation of public signals and trade in speculative markets. *Journal of Political Economy* **103**: 831–872.

Kim O, Verrecchia R. 1994. Market liquidity and volume around earnings announcements. *Journal of Accounting and Economics* **17**: 41–67.

Lee C, Ready M. 1991. Inferring trade direction from intraday data. *Journal of Finance* **46**: 733–746.

Copyright © 2014 John Wiley & Sons, Ltd.

Patton A, Verardo M. 2012. Does beta move with News? Firm-specific information flows and learning about profitability. *Review of Financial Studies* **25**: 2789–2839.

Preve D, Tse Y. 2013. Estimation of time-varying adjusted probability of informed trading and probability of symmetric order-flow shock. *Journal of Applied Econometrics* **28**: 1138–1152.

Sarkar A, Schwartz R. 2009. Market sidedness: Insights into motives for trade initiation. *Journal of Finance* **64**: 375–423.

Sugar C, James G. 2003. Finding the number of clusters in a dataset. *Journal of the American Statistical Association* **98**: 750–763.

Tay A, Ting C, Tse Y, Warachka M. 2009. Using high-frequency transaction data to estimate the probability of informed trading. *Journal of Financial Econometrics* **7**: 288–311.

Tetlock P. 2010. Does public financial news resolve asymmetric information? *Review of Financial Studies* **23**: 3520–3557.

Zucchini W, MacDonald I. 2009. *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC Press: Boca Raton, FL.

Copyright © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/jae