取引開始前の気配更新と価格発見*

太田　亘
大阪大学大学院経済学研究科

本稿では，unbiasedness regression により，東京証券取引所の取引開始前に配信される気配の情報効率性について分析した．他の取引所と同様，東京証券取引所の取引開始前の気配もノイズではなく新情報が反映しており，投資家は気配を通じた学習を行っていると考えられる．取引開始前の気配の更新は，価格が大きく動く日ほど活発である．また，価格変化の大きさや気配更新の活発さにかかわらず，取引開始直後にはほぼ情報効率的な価格形成がなされている．以上より，価格変化が大きく価格発見が困難であるときには，気配を活発に更新しながら価格発見を行っている可能性がある．

情報効率性，価格発見，寄前気配，unbiasedness regression.

1　はじめに

情報は価格にどのように織り込まれていくだろうか．また，本源的価値の探索を価格発見というが，この価格発見はどのように行われるだろうか．証券取引所の取引時間中は，取引を通じて情報が価格に織り込まれていくと考えられる．例えばKyle (1985) のモデルでは，一般の投資家に加えて情報上優位な投資家が注文を出し，一般の投資家も価格や注文情報から自分の知らない情報を推測したうえで取引する過程を通じ価格発見が行われ，情報が価格に反映して行く．では，取引をせず，投資家の注文を反映した気配 (indicative quote) を単に提示しているだけの場合，情報が気配に反映されるだろうか．取引前の注文をコストなくキャンセルすることができ，しかも取引開始前に出された注文に対して注文間の優先ルールである時間優先が適用されない場合には，投資家が後からキャンセルするつもりで適当な注文を出したり，価格操作を狙った注文を出したりすることにより，気配に新情報が反映しない可能性がある．

この問題に関し，Biais et al. (1999) は，unbiasedness regression とよばれる手法を用い，パリ証券取引所の取引開始前に配信されている気配には新情報が反映しており，それを通じて投資家間で学習が行われている，という実証結果を示している．その他の取引所についても，Cao et al. (2000), Madhavan and Panchapagesan (2000), Barclay et al. (2003, 2008), Comerton-Forde and Rydge (2006), Chen et al.(2009) などの研究が，取引開始前の気配に新情報が織り込まれている，という結果を報告している．本稿では，東京証券取引所 (東証) について Biais et al. (1999) と同様の方法を用い，取引開始前の気配を通じた価格発見について分析を行う．

本稿では，取引開始前の気配について，次の3つの仮説を検証する．仮説1は「取引開始前の気配はノイズではなく，新情報が反映している」である．この仮説は既に Biais et al. (1999) 等の先行研究が他の市場について検証しているが，本稿でも東証について検証する．先行研究は，取引開始前の気配の水準のみを問題にしているが，実際には取引開始前の気配が活発に更新される日とそう

*本稿は「取引開始前および取引開始後の価格発見」の一部を改訂したものです．宇野淳氏および大庭昭彦氏，日本ファイナンス学会第18回大会，応用経済時系列研究会第27回研究報告会，大阪大学金融・保険教育研究センター中之島ワークショップの参加者からの貴重なコメントに感謝します．なお，本研究は文部科学省科学研究費補助金 (課題番号 21243019) の援助を受けています．

でない日とがある．これに関し，取引開始前の気配の更新頻度が価格発見と何らかの関係があるのか，例えば気配の頻繁な更新は価格操作によるものなのか，については必ずしも明らかになっていない．そこで本稿独自の分析として，仮説2「価格変化が大きな日に，取引開始前に気配が頻繁に更新される」および仮説3「取引開始時の価格形成は，価格変化・気配の更新頻度の影響を受けない」を検証する．

取引価格ではなく気配を通じた価格発見または学習についての理論研究には，Kobayashi (1977), Jordan (1982), Vives (1995) などがある．この中で Kobayashi (1977) は，証券市場における合理的期待均衡を探索するにあたり，投資家の数が多いとき，均衡価格を見つけるまでに気配の更新をより多くの回数行う必要がある可能性を示している．この結論は，情報が偏在しており価格発見が難しいと考えられるとき，例えば前日取引終了後の重大なニュースにより本源的価値が大きく変化したと投資家が予想するとき，新しい均衡価格を発見するにあたり，気配を頻繁に更新しながら学習や情報交換を行い，その結果，情報効率的な価格形成を行うことのできる可能性があることを示唆している．本源的価値の変化は観察不可能であるため，本稿では取引価格の変化を本源的価値の変化または価格発見の難しさの代理変数とする．そのうえで，仮説2のように，前日から価格が大きく変化したときに取引開始前の気配の更新頻度が高いか，また仮説3のように気配の更新頻度が高いときにも通常と同様な価格形成が行われているか，を検証する．

気配更新と価格変化との関係は，例えば価格操作や取引所外での情報共有の影響も受ける．価格操作が情報非効率な価格形成を狙う活動であるとすると，価格操作者が取引開始前に発注・キャンセルを繰り返えすことで気配の更新が活発になるとき，それによって価格変動が大きくなり，同時に取引開始後の取引価格の情報効率性が損なわれると考えられる．この場合，仮説2と整合的であるが，仮説3と不整合な実証結果が得られる．また価格操作者が，自分の操作と反対方向に価格が動くリスクを避けるために，本源的価値の変化が小さいと予想されるときに価格操作を行おうとするのであれば，価格変化が小さいときに気配の更新頻度が高くなる．このような価格操作が頻繁に行われている場合には，仮説2と不整合な実証結果が得られる．一方，公表情報などにより取引開始前の気配更新を必要とすることなく価格発見が行われる場合にも，仮説2と不整合な可能性が高くなる．逆に，仮説2および仮説3と整合的な結果が得られるとき，取引開始前の気配が頻繁に更新されるのは，価格操作などのためというよりも，本源的価値を発見するためである可能性が高い．

以上3つの仮説を検証するため，2004年1月から2005年12月までの2年間において，日経平均採用銘柄中170銘柄の1日の取引開始前の気配について分析を行った．分析結果は，仮説1，2，3のいずれとも整合的であった．まず仮説1について，先行研究が分析した他の市場と同様，東証においても取引前の気配に新情報が反映されている．仮説2の価格変化の大きさと取引開始前の気配の更新について，価格変化が大きかった日ほど，気配が活発に更新されている．仮説3について，価格変化が大きかった日には，取引開始前の気配の調整が相対的に遅い傾向がある．しかし，市場全体の共通要因に対する価格付けについても，各銘柄の固有要因に対する価格付けについても，取引開始直後には情報効率的な価格形成がなされている．また，気配の更新頻度が高いときに，気配が過剰または過小に反応する傾向は観察されなかった．これは，取引開始前の気配について，価格操作が頻繁に行われているわけではないことを示唆している．以上より，取引開始前の気配には新情報が反映しているとともに，特に本源的価値が大きく変動したと考えられる価格変動が大きいときには，取引開始前に気配を頻繁に更新しながら価格発見を行っているといえる．なお，東証は，特別気配の更新値幅という価格制限ルールにより，取引時間中の価格変動を抑えているが，このルールは取引開始前には適用されない．しかし取引開始前において，取引開始直後に価格制限ルールの発動が予想される場合には，気配の過剰な変動が観察される．

以下ではまず，分析期間中に東証が配信していた気配情報について説明し，次に分析手法として用いる unbiasedness regression について説明する．その後，サンプルについて説明し，第5節で基

本モデルにより仮説1を検証したうえで，第6節で取引開始前の気配の更新頻度に関し仮説2を検証する．さらに第7節において気配の更新頻度が価格形成に与える影響についての仮説3に関する実証結果をまとめ，最後に結論を述べる．

2　東証の配信情報

本節では，東証の取引ルールおよび公表情報のうち，本稿の分析において重要となる寄前気配と特別気配に関連するルールとともに，東証の配信情報を簡潔に説明する．

2.1　取引ルールと気配配信

東証の取引時間 (立会時間) は，9時から11時までの前場と，12時30分から15時の後場からなる．東証は，証券会社等の取引参加者からの注文を8時から11時および12時5分から15時まで受け付ける．取引所が受け取った注文は，最初の約定時点まではすべて未執行であるが，未執行の注文は板に記録される．東証は前場開始前8時20分から9時直前までと，後場開始前12時5分から12時30分直前まで，板の情報のうち一部を寄前気配として配信している．また取引時間中にも類似の板情報を配信している．これらの情報は，取引所に注文が入り板が変化するのに応じて更新される．

取引開始後最初の約定には，板寄せ方式とよばれるコール・オークションが用いられ，最初の約定の時点を寄付きとよぶ．東証は，特別気配の更新値幅とよばれる価格制限ルールにより，直近の取引価格に比べて特別気配の更新値幅を超える大きな価格変化が起こる場合，即時の約定を中止して特別気配を配信し，その後一定の条件が満たされたときに約定を行う．特別気配として配信されるのは，その時点で価格制限にかからない限界の価格，すなわちその時点で即時に約定を行うことのできる価格である．取引開始直後に特別気配が配信されない場合には，取引開始前に寄前気配が配信され，取引開始の9時0分または12時30分に寄付いて最初の約定が行われ，その後，最良売り気配と最良買い気配を含む上下5本気配とよばれる板情報が配信される．一方，前日終値に比べて特別気配の更新値幅を超える価格変化が発生する場合は，取引開始前に寄前気配が配信され，9時0分または12時30分から特別気配が配信され，板寄せ方式および価格変化の条件が満たされたときに寄付いて最初の約定が行われ，その後に上下5本気配が配信される．すなわち，取引開始の9時0分または12時30分以降に特別気配が配信される場合があり，その特別気配はルールにより価格変化が抑えられている．

特別気配の更新値幅による価格制限は，5分ごとに緩和される．そのため，取引開始時から特別気配が配信されている場合，9時5分や9時10分に価格制限の制約が緩和されることで寄付き，特別気配の配信から上下5本気配の配信に移行する可能性が高い．

2.2　寄前気配

取引開始前に配信される寄前気配は，基本的に，売り買いの累積数量が逆転する上側の価格を最良売り気配，下側の価格を最良買い気配として情報が配信され，さらにその気配での累積数量も公表される．この時点で仮想的に約定を行うとき，価格は最良売り気配か最良買い気配となる．そのため，寄前気配は，1日の最初の取引価格がどのような水準になるかの情報を投資家に与える可能性がある．また最良気配に追加して，最良売り気配から高い方向に4つの売り気配，最良買い気配から低い方向に4つの買い気配，およびそれらの気配で待っている指値注文の合計数量も配信される．売り注文と買い注文がクロスしておらず，その時点で仮想的に板寄せ方式を行った場合に約定

表 1: 寄前気配

価格 (円)	売り注文 数量	買い注文 数量	売り注文 累積数量	買い注文 累積数量	売り 寄前気配	買い 寄前気配
89	1		6	0	1	
88	1		5	0	1	
87	1		4	0	1	
86	1		3	0	1	
85	1	1	2	1	2	
84	1	1	1	2		2
83		1	0	3		1
82		1	0	4		1
81		1	0	5		1
80		1	0	6		1

寄前気配の配信方法の具体例を示している．売り (買い) 注文数量は，80 円から 89 円の価格に出された指値注文の数量である．成行注文はない．表は，この状況における売り買いの累積数量および寄前気配を示している．

数量がゼロとなる場合には，板上の買い指値注文のうち最も高い価格を最良買い気配，板上の売り指値注文のうち最も低い価格を最良売り気配として寄前気配が配信される．

例として，80 円から 85 円までそれぞれ 1 単位の買い指値注文が出ており，同時に 84 円から 89 円までそれぞれ 1 単位の売り指値注文が出ている状況を考えよう．成行注文はないものとする．表 1 は，この状況を示している．価格優先の原則により，売り (買い) 注文は価格の低い (高い) 注文が優先されるので，売り (買い) 注文については価格の低い (高い) 方から累積数量を計算する．売りと買いの累積数量が逆転するのは，84 円と 85 円の間である．よってこの例の場合，寄前気配として，85 円で 2 単位の最良売り気配，86 円から 89 円まで 1 単位の売り気配，84 円で 2 単位の最良買い気配，80 円から 83 円まで 1 単位の買い気配が配信される．

取引開始後は特別気配の更新値幅の制約を受けるが，取引開始前にはそのような制約はない．そのため，寄前気配は基本的に制約なく自由に変動する．但し，制限値幅とよばれる一日の価格制限ルールがあり，これにより投資家の出すことのできる指値注文の価格帯が制限され，よって寄前気配も同様に制限値幅の制約を受けることになる．

東証では，注文をいつでも自由にキャンセルすることができるとともに，取引開始時の板寄せ方式の注文優先順序として時間優先を用いていない．そのため，投資家にとって，取引開始前の早い段階から特定価格にコミットする指値注文を出すインセンティブは小さい．また，価格操作を狙うなどの動機で，寄前気配を大きく動かすような注文を出し，取引開始直前にキャンセルする，という発注行動をコストゼロで行うことが可能である．このように，東証のルールの下では，寄前気配に新情報が反映されにくいと考えられる．例えば Biais et al. (2009) は，寄付き前の注文をキャンセルできるルールの場合に比べ，キャンセルできないルールの場合には，より適切な価格形成が行われる，という実験結果を示している．キャンセルや時間優先に関するルール変更が東証の寄前気配の形成にどのような影響を与えるかについての分析は，今後の課題として残されている．

3 分析方法

本稿で用いる unbiasedness regression は，証券の本源的価値を v，それ以前のある時点 0 での本源的価値の推定値を $E(v|I_0)$，時点 0 と本源的価値 v が明らかになる時点の中間時点 τ における価格を P_τ としたとき，

$$v - E(v|I_0) = \alpha + \beta(P_\tau - E(v|I_0)) + \epsilon \tag{1}$$

とする回帰分析である．α と β は回帰係数，ϵ は誤差項であり，以下でも同様とする．取引日 d の終値を $P_{c,d}$ とおき，$v = P_{c,d}$，$E(v|I_0) = P_{c,d-1}$，取引日 d の時点 τ における価格を $P_{\tau,d}$ としよう．

Biais et al. (1999) は，複数取引日のデータを用いて，取引開始前および開始後の各時点 τ ごとに，

$$\frac{P_{c,d}-P_{c,d-1}}{P_{c,d-1}}=\alpha+\beta\frac{P_{\tau,d}-P_{c,d-1}}{P_{c,d-1}}+\epsilon_d \tag{2}$$

とする回帰式を推計している．このとき

$$\beta=\frac{Cov(P_{c,d}-P_{c,d-1},\ P_{\tau,d}-P_{c,d-1})}{Var(P_{\tau,d}-P_{c,d-1})}=1+\frac{Cov(P_{c,d}-P_{\tau,d},\ P_{\tau,d}-P_{c,d-1})}{Var(P_{\tau,d}-P_{c,d-1})} \tag{3}$$

である．時点 τ の価格 $P_{\tau,d}$ に，時点 0 以降に発生した新情報が全く反映されておらず，$Cov(P_{c,d}-P_{c,d-1},e_{\tau,d})=0$ となる $e_{\tau,d}$ に対して $P_{\tau,d}=P_{c,d-1}+e_{\tau,d}$ であれば，$\beta=0$ となる．それに対して新情報が反映し価格発見が行われていれば，$\beta>0$ となる．一方，価格が情報効率的でランダムウォークしていれば，時点 τ にかかわらず $Cov(P_{c,d}-P_{\tau,d},\ P_{\tau,d}-P_{c,d-1})=0$ であり，$\beta=1$ となる．一方，$Cov(P_{c,d}-P_{\tau,d},\ P_{\tau,d}-P_{c,d-1})>0$ のとき $\beta>1$ であり，このとき情報効率的な価格に比べて過小反応が起こっている，と解釈することができる．逆に，$\beta<1$ となるのは $Cov(P_{c,d}-P_{\tau,d},\ P_{\tau,d}-P_{c,d-1})<0$ のときであり，この場合には情報効率的な価格に比べて過剰反応が起こっている，と解釈することができる．また，価格制限ルールにより特別気配が提示されている場合，価格変化がルールによって過小となるので，β の推計値が大きくなると予想される．

Biais et al. (1999) は，パリ証券取引所に上場されている 39 銘柄の 19 日分のデータを用い，それら銘柄から計算した指数および個別銘柄について，unbiasedness regression を行っている．指数については，取引開始前の β の推計値は 1 未満の正の値をとり，取引開始前の気配に新情報が反映しており，また取引開始が近づくほど β の推計値は 1 に向けて大きくなり，取引開始時には推計値がほぼ 1 となって情報効率的な価格形成が行われている，という実証結果を示している．一方，個別銘柄については，取引開始前から気配に新情報が反映しているものの，$\beta=$ の推計値がほぼ 1 となるのは取引開始から 30 分程度経過した後である，と報告している．

4 サンプル

本稿で用いるデータは，日経メディアマーケティング提供のティックデータおよびポートフォリオ・マスターである．本節では，分析対象期間および銘柄，分析対象日，データ加工方法等を説明する．

4.1 分析対象期間および銘柄

分析対象期間は，2004 年および 2005 年の 2 年間である．分析対象銘柄は，期間中の最低価格が 200 円を超えている銘柄の中で，東証に継続して上場されているとともに日経平均株価の算出に継続して採用されており，さらに American Depository Receipt (ADR, 米国預託証券) がニューヨーク証券取引所または NASDAQ に上場されていない銘柄である．この基準を満たす銘柄は 170 ある．以下では銘柄を $i=1,\cdots,170$ で表す．

東証における価格の最小の変化幅である呼値の刻みは，2000 円以下の価格帯では 1 円である．価格が 200 円を下回る場合，呼値の刻みの価格に対する比率が 0.5%以上となり，価格が離散であることから発生する誤差が大きくなる．そのため分析対象外とする．

日経平均先物取引において指数に関する価格発見が行われているときには，日経平均株価算出に採用されている銘柄と採用されていない銘柄とでは，寄前気配の更新や価格発見に差がある可能性が発生する．そのため，本稿の分析では，日経平均株価に採用されている銘柄に分析対象を限定した．日経平均採用・非採用が価格発見に与える影響に関する分析は，今後の課題である．

また ADR が取引されている銘柄では価格発見の一部が外国市場で行われ，ADR が取引されていない銘柄と取引開始前の価格発見において差が発生する可能性がある．その差の影響を避けるため，本研究では，ADR が取引されていない銘柄に分析対象を限定する．指数先物取引の影響と同様，ADR 取引の影響に関する分析も，今後の課題である．

4.2 分析対象日

本稿の unbiasedness regression では，1 日の取引開始時前後の価格形成を分析するため，前日価格および当日価格を同時に用いる．そのため，前日から当日にかけて価格に断絶があると考えられる以下の取引日を分析対象から除外する．まず午後の取引が行われない各年最初の取引日の翌日，システムトラブルのあった 2005 年 11 月 1 日およびその翌日を分析対象外とする．また当日の始値形成を分析するため，派生証券の清算値段決定の影響がある各月第 2 金曜日 (休日の場合は前日) を分析対象外とする．各銘柄の権利落ち日には，価格に理論上ジャンプが発生するため，多くの銘柄で権利落ち日となる 2004 年 3 月 26 日，2004 年 9 月 27 日，2005 年 3 月 28 日，2005 年 9 月 27 日を分析対象外とする．最後に，午後の取引が行われない各年最初と最後の取引日も分析対象から外す．以上を除外した取引日数は 455 日である．以下では $D = 455$ とし，取引日を $d = 1, \cdots, D$ により表す．

さらに個別銘柄について，各銘柄の権利落ち日を分析対象外とする．具体的には，ある銘柄の権利落ち日が先に除いた 2004 年 3 月 26 日，2004 年 9 月 27 日，2005 年 3 月 28 日，2005 年 9 月 27 日以外の場合は，その銘柄についてその日を欠損値とする．

4.3 気配の計測

分析対象とする気配は次の通りである．取引開始前も開始後も，最良売り気配と最良買い気配の仲値 (平均) を気配とする．特別気配が配信されている場合は，特別気配を気配とする．また，最良売り (買い) 気配しか配信されていない場合には，最良売り (買い) 気配を気配とする．最良売り気配も最良買い気配も配信されていない場合は，欠損値とする．

東証の配信データは，時刻が分単位で記録されている．そのため本稿では，時刻を分単位で計測し，各時刻の最後，すなわち次の時刻の直前の気配を分析対象とする．例えば，8 時 59 分の気配は 9 時 0 分直前の気配，9 時 0 分の気配は取引開始後 1 分経過した 9 時 1 分直前の気配である．以下では，時刻を t により表し，例えば $t = 8:59$ は 8 時 59 分を意味するものとする．

4.4 本源的価値の代理変数

式 (1) の unbiasedness regression では，本源的価値 v の代理変数およびそれ以前における本源的価値の推定値 $E(v|I_0)$ の代理変数を選択する必要がある．本稿では，本源的価値 v に当日 10 時 55 分の最良気配の仲値，本源的価値の推定値 $E(v|I_0)$ に前日 14 時 55 分の最良気配の仲値を用いる．Amihud and Mendelson (1991) により，東証の前場終値は，少なくとも前場始値に比べて情報効率的であることが知られている．また前場終値に対して後場終値は午後の価格変化も含むため，前場開始前の価格形成の分析にあたり推計誤差が大きくなる可能性がある．そのため，本源的価値に当日の終値ではなく，前場の終値付近の気配を用いる．Comerton-Forde and Rydge (2006) も同様に，本源的価値の代理変数に 11 時の最良気配の仲値を用いている．また終値形成は，Madhavan, Richardson, and Roomans (1997)，Cushing and Madhavan (2000)，Hillion and Suominen (2004),

Chang et al. (2008) などが議論しているように，ディーラーの在庫費用の増加，機関投資家の大口注文，価格操作等の影響を受ける可能性がある．そのため取引終了時ではなく，取引終了 5 分前である 10 時 55 分および 14 時 55 分の最良気配の仲値を，本源的価値 v および本源的価値の推定値 $E(v|I_0)$ の代理変数として用いる．

Barclay and Hendershott (2008) が議論しているように，観察される価格が，本源的価値にノイズののった値である場合，unbiasedness regression の推計値はバイアスをもつ．しかし，どのようにバイアスが発生するかは，ノイズの構造に依存する．より適切な推計のための変数選択は，今後の課題として残されている．

4.5 指数および市場リターン

各銘柄の各時刻の気配から時価総額を計算し，分析対象 170 銘柄の時価総額の合計を指数とする．但し，個別銘柄の分析では，第 4.3 節で説明したように，売り気配および買い気配の提示がない場合には，当該時刻を欠損値とするが，指数の計算においては欠損値にせず，直近の気配を指数の計算に用いる．この指数は，株式分割や売買単位変更等の影響は受けないが，配当落ちにより誤差が発生する可能性があるとともに，特別気配の影響を受ける．配当落ちからの誤差を少なくするため，第 4.2 節で説明したように，配当落ちが集中する 3 月末および 9 月末を分析対象から除外している．また，一部の銘柄で特別気配が提示されるか提示が予想されるときには，指数にバイアスが発生する可能性がある．そのため，第 4.1 節で説明したように，日経平均株価の算出に採用され，売買が活発で特別気配が提示される可能性の少ない銘柄に分析対象を限定している．

指数の変化率を市場リターンとする．取引日 d において，前日 14 時 55 分から当日時刻 t までの指数の変化率を $\Delta p^M_{t,d}$ とし，特に前日 14 時 55 分から当日 10 時 55 分までの指数の変化率を $\Delta p^M_{c,d}$ とする．10 時 55 分または 14 時 55 分に特別気配が提示されている銘柄が多い場合，市場リターンにバイアスの発生する可能性が大きくなる．しかし，指数を計算する銘柄×日のうち，10 時 55 分において特別気配が配信されていたのは 0.003%，14 時 55 分において特別気配が配信されていたのは 0.004%である．そのため，本源的価値計測時点である 10 時 55 分および 14 時 55 分における指数に対する特別気配の直接的影響は小さいと考えられる．

4.6 トータルリターンおよび固有リターン

個別銘柄の収益率をトータルリターンとよび，銘柄 i 取引日 d について，前日 14 時 55 分から当日時刻 t までのトータルリターンを $\Delta p^{Ti}_{t,d}$ とする．特に前日 14 時 55 分から当日 10 時 55 分までのトータルリターンを $\Delta p^{Ti}_{c,d}$ とする．

トータルリターンを，市場全体にかかわる共通要因による変動と，それ以外の銘柄固有要因による変動とに分解することができる．前節で説明した市場リターンは，共通要因による価格変動を表す．例えば Tse (1999) は，指数に関する価格発見は，指数先物取引においても行われている，という実証結果を報告している．そのため，指数先物取引の影響を受ける共通要因の価格変動を取り除き，各銘柄の固有要因に関する価格発見に関しても分析する．固有要因による価格変動を固有リターンとよぶが，本稿では，トータルリターンを市場リターンに回帰するモデルであるインデックス・モデルを推計し，その残差を固有リターンとする．そして，市場リターン，固有リターン，トータルリターンのそれぞれの価格発見を分析する．Biais et al. (1999) も同様に，市場リターン，各銘柄のトータルリータンおよび固有リターンに対して分析を行っている．

具体的に，固有リターンを次のように推計する．銘柄 i 取引日 d について，前日 14 時 55 分から当日時刻 t までの固有リターン $\Delta p^{Ii}_{t,d}$ は，銘柄 i 時刻 t ごとのインデックス・モデルの切片の推計値を

$\hat{\alpha}_t^i$，傾きの推計値を $\hat{\beta}_t^i$ としたとき，$\Delta p_{t,d}^{Ii} = \Delta p_{t,d}^{Ti} - \hat{\alpha}_t^i - \hat{\beta}_t^i \Delta p_{t,d}^M$ である．特に前日 14 時 55 分から当日 10 時 55 分までの固有リターンを $\Delta p_{c,d}^{Ii}$ とする．すなわち，$\Delta p_{c,d}^{Ii} = \Delta p_{c,d}^{Ti} - \hat{\alpha}_{10:55}^i - \hat{\beta}_{10:55}^i \Delta p_{c,d}^M$ である．

固有リターンの計測にあたり，インデックス・モデルの傾きの推計誤差が結果に影響を与える可能性がある．そのため仮説 3 の検証では，固有リターンを，トータルリターンと市場リターンの差とする場合の推計結果についても検討する．

本稿では，各銘柄のトータルリターンに影響を与える共通要因として，市場リターンのみを考えるが，これ以外の要因を考えることも可能である．例えば Fama and French (1993) は，市場リターン以外に，小型株と大型株のリターンの差であるサイズ・ファクターと，簿価/時価比率の高い銘柄と低い銘柄のリターンの差であるバリュー・ファクターを共通要因として考えている．しかし，サイズ・ファクターやバリュー・ファクターを 1 分間隔で計測するにあたり，より多くの銘柄の気配を用いる必要があり，それら銘柄は売買がより不活発で価格制限の影響をより強く受けるため，推計誤差がより大きくなると考えられる．そのため本稿では，共通要因として市場リターンのみを考える．市場リターン以外の共通要因についての価格発見，および市場リターン以外の共通要因を考慮した後の固有リターンについての価格発見に関する分析は，今後の課題である．

4.7 気配更新頻度

寄前気配の更新頻度が価格形成に与える影響を分析するため，各銘柄各時刻分単位で，寄前気配のうち最良気配が更新された回数を計測し，気配更新頻度とする．さらに市場リターン，固有リターン，トータルリターンと同様に，市場更新頻度，固有更新頻度，総更新頻度の 3 種類を考える．取引日 d 時刻 t の市場更新頻度 $q_{t,d}^M$ を，各銘柄について寄前気配の配信開始から時刻 t までの累積の気配更新頻度を全銘柄について合計した値の対数値とする．銘柄 i 取引日 d 時刻 t の総更新頻度 $q_{t,d}^{Ti}$ を，その銘柄の寄前気配の配信開始から時刻 t までの累積の気配更新頻度の対数値とする．銘柄 i 取引日 d 時刻 t の固有更新頻度 $q_{t,d}^{Ii}$ は，銘柄 i 時刻 t ごとに被説明変数を $q_{t,d}^{Ti}$，説明変数を $q_{t,d}^M$ とし，$d = 1, \ldots, D$ の観測値を用いて推計した回帰モデルの残差とする．すなわち切片の推計値を $\bar{\alpha}_t^i$，傾きの推計値を $\bar{\beta}_t^i$ としたとき，$q_{t,d}^{Ii} = q_{t,d}^{Ti} - \bar{\alpha}_t^i - \bar{\beta}_t^i q_{t,d}^M$ である．銘柄ごとの回帰モデルについて，説明変数の係数の推計値 $\bar{\beta}_t^i$ の平均は 0.926，決定係数の平均は 0.379 である．市場更新頻度，固有更新頻度，総更新頻度のいずれも 8 時 59 分まで計測されるが，9 時 0 分以降は 8 時 59 分と同じ値をとるものとする．

寄前気配として最良気配の外側の気配も配信されているが，本稿では最良気配の更新のみを分析対象とする．寄前気配のうち最良気配は，仮想的取引価格に影響を与えるような注文が出た場合に更新されるのに対し，最良気配の外側の気配は，仮想的取引価格に直接影響を与えない注文が出た場合に更新される．ここでは，価格形成においてより重要であると考えられる最良気配に焦点を絞る．例えば Cao et al. (2009) は，取引時間中の板について，最良気配の外側の効果を含めて分析し，最良気配の情報が最も重要であると報告している．寄前気配における最良気配の外側および数量の情報についての分析は，今後の課題である．

4.8 特別気配ダミー

東証の価格制限ルールにより，取引開始後に配信される特別気配は，価格変化が抑えられた気配となっている．一方，価格制限ルールの発動が予想されるとき，投資家の戦略的発注行動により，取引開始前の気配も価格制限の影響を受ける可能性がある．例えば Subrahmanyam (1997) は，取引停

止が予想される場合の戦略的発注行動についてを理論分析を行っており，一方，George and Hwang (1995) は，東証における取引停止について実証分析を行っている．

このような特別気配の影響をコントロールするため，特別気配ダミーを用いた分析を行う．銘柄 i 取引日 d の特別気配ダミー L_d^i を，9 時 1 分直前までに最初の約定があった日にゼロ，それ以外の日に 1 をとるダミー変数とする．言い換えると，L_d^i は，寄付きが 1 分以上遅延し，9 時 1 分に特別気配が配信されていた日に 1 をとるダミー変数である．

市場全体についての特別気配ダミー L_d^M を，多くの銘柄で寄付きが 1 分以上遅れたことを表すダミー変数とする．各取引日について 9 時 1 分直前までに最初の約定があった銘柄を数え，その平均をとると，148.5 銘柄である．これを利用し，L_d^M を，9 時 1 分までに最初の約定があった銘柄が 170 銘柄中 148 銘柄以下である日に 1，それ以外の日にゼロをとるダミー変数とする．

ダミー変数 L_d^i および L_d^M の値は，9 時 1 分直前を基準とし，同一銘柄同一日においてすべての時刻で共通とする．ここで投資家は，9 時 0 分前において，9 時 1 分までに約定が開始されるかを完全予見できると仮定している．取引開始前の気配の水準により，特別気配の配信をある程度予想できるのであれば，特に 8 時 59 分において，1 分後に寄付いているかを完全予見できるとする仮定は，それほど強い仮定ではないと考えられる．

5 基本モデルの推計

本節では，仮説 1「取引開始前の気配はノイズではなく，新情報が反映している」について，分析方法を説明した後，市場リターン，固有リターン，トータルリターンのそれぞれについての unbiasedness regression の推計結果を報告する．

5.1 基本モデル

推計にあたり，東証の価格制限ルールを考慮して，特別気配ダミーとリターンとのクロス項を含める．市場リターンについては，時刻 t ごとに，取引日 $d = 1, \ldots, D$ の観測値を用いて，

$$\Delta p_{c,d}^M = \alpha + \beta \Delta p_{t,d}^M + \gamma L_d^M \Delta p_{t,d}^M + \epsilon_d \tag{4}$$

を推計する．α と β に加え γ も回帰係数であり，以下でも同様とする．β の推計値が気配の情報効率性の指標であり，0 よりも大きければノイズでなく新情報を含んでおり，1 であれば情報効率的，0 超 1 未満であれば過剰反応，1 超であれば過小反応であることを意味する．また γ の推計値が 0 以外のとき，取引開始直後の特別気配の配信が，価格形成に影響を与えている．unbiasedness regression に特別気配ダミーとリターンとのクロス項を含むことについて，例えば Comerton-Forde and Rydge (2006) も同様に，ダミー変数とリターンとのクロス項を含めた unbiasedness regression を推計している．

固有リターンについては，銘柄 i 時刻 t ごとに，取引日 $d = 1, \ldots, D$ の観測値を用いて，

$$\Delta p_{c,d}^{Ii} = \alpha + \beta \Delta p_{t,d}^{Ii} + \gamma L_d^i \Delta p_{t,d}^{Ii} + \epsilon_d \tag{5}$$

を推計する．トータルリターンについても同様に，銘柄 i 時刻 t ごとに，取引日 $d = 1, \ldots, D$ の観測値を用いて，

$$\Delta p_{c,d}^{Ti} = \alpha + \beta \Delta p_{t,d}^{Ti} + \gamma L_d^i \Delta p_{t,d}^{Ti} + \epsilon_d \tag{6}$$

を推計する．

仮説1の検証において，帰無仮説は，8時59分以前の推計において，リターンの係数βがゼロ，である．一方，対立仮説は，リターンの係数βが正，である．

固有リターンおよびトータルリターンの銘柄ごとの推計では，以下の観測値を外れ値として推計から外す．東証では未執行の注文をペナルティなく自由にキャンセルすることができ，寄前気配は寄付き直前のキャンセルを意図した注文の影響を受ける可能性がある．すなわち，価格操作を狙うなど売買を意図しない注文により寄前気配が大きく変動し，それに応じて一部の観測値が，回帰分析における外れ値になる可能性がある．どれを外れ値と考えるかは，実際に寄付きで取引しようとする投資家にとって重要な問題であるとともに，推計結果にバイアスをもたらす可能性がある点で価格形成を評価するうえでも重要である．外れ値の識別にどのような方法を用いるのが適切であるかは，本稿の分析を超える問題であるが，ここでは，固有リターンの各銘柄各時刻の推計において，被説明変数$\Delta p^{Ii}_{c,d}$および説明変数$\Delta p^{Ii}_{t,d}$のそれぞれの1パーセンタイルと99パーセンタイルの外側にある観測値を外れ値として除外する．サンプルを統一するため，外れ値とされた同じタイミングの観測値を，トータルリターンを用いる推計においても除外する．

銘柄ごとの推計により，各時刻において170銘柄の説明変数の係数の推計値が得られるが，これに対してt検定を行い係数の有意性を判断する．但し銘柄間での残差の相関の影響を考慮するため，Chordia et al. (2008)と同様，各銘柄の残差の分散と銘柄間での残差の相関が銘柄に関わらず同一であると仮定したもとでt値を修正する．なお自由度170のt分布の上側確率0.025を与える確率点は1.974，上側確率0.01を与える確率点は2.348である．

特別気配ダミーとリターンとのクロス項の係数γは，次のような符号になると予想される．価格制限は5分ごとに緩和されるため，9時0分以降9時4分までは，価格制限により気配の変化が過小となり，γの推計値は正になると予想される．8時59分以前における推計では，γの推計値の符号は予めわからない．但し，板寄せ方式では価格優先の原則が適用され，売り注文であれば価格のより低い注文，買い注文であれば価格のより高い注文が約定において優先されるが，投資家が執行を確実にするためにこのような注文を競って出すのであれば，寄前気配が情報効率的な水準を超えて大きく変化することで過剰反応となり，γの推計値は負になる．

5.2 推計結果

表2は，市場リターン，固有リターン，トータルリターンのそれぞれについて，8時55分から9時5分の各時刻ごとに，unbiasedness regressionの推計結果をまとめたものである．固有リターンとトータルリターンについては，各時刻各銘柄ごと回帰を行い，各時刻について170銘柄の説明変数の係数の推計値の平均および決定係数の平均を表に示している．括弧内はt値である．( )の場合は説明変数の係数がゼロと有意に異なるか，[ ]の場合は説明変数の係数が1と有意に異なるか，を検定するためのt値を示している．

8時55分から8時59分の推計において，リターンの係数は，市場リターン，固有リターン，トータルリターンのいずれにおいても正であり，また有意水準1%で係数ゼロの帰無仮説は棄却される．これは，仮説1と整合的な結果であり，先行研究と同様，取引を伴わない気配に新情報が反映していることを示している．係数の大きさは，9時前は1よりも小さく過剰反応が観察されるが，9時に近づくほど1に近づいており，取引開始がせまるほど気配の過剰な変動が抑制されている，と解釈することができる．

取引開始後9時0分から9時5分について，市場リターン，固有リターン，トータルリターンのいずれの係数の推計値も約1で，1と有意に異ならない．これは，取引開始後1分以内に情報効率的な価格形成が行われていることを示唆している．

表 2: 基本モデルの推計結果

パネル A: 被説明変数:市場リターン ($\Delta p^M_{c,d}$)								
時刻 t	定数項		$\Delta p^M_{t,d}$			$L^M_d \Delta p^M_{t,d}$		R^2
8:55	0.000	(1.69)	0.460	(7.03)	[-8.24]	0.121	(1.64)	0.438
8:56	0.000	(1.55)	0.497	(7.52)	[-7.60]	0.103	(1.40)	0.462
8:57	0.000	(1.54)	0.521	(7.70)	[-7.08]	0.095	(1.26)	0.472
8:58	0.000	(1.25)	0.556	(8.00)	[-6.38]	0.093	(1.20)	0.492
8:59	-0.000	(-0.10)	0.809	(10.8)	[-2.55]	0.030	(0.37)	0.612
9:00	-0.000	(-1.74)	1.037	(11.2)	[0.40]	0.318	(3.01)	0.621
9:01	-0.000	(-1.25)	0.990	(11.3)	[-0.12]	0.324	(3.20)	0.621
9:02	-0.000	(-0.86)	0.985	(11.4)	[-0.18]	0.303	(3.04)	0.623
9:03	-0.000	(-0.61)	0.977	(11.6)	[-0.27]	0.289	(2.96)	0.627
9:04	-0.000	(-0.40)	0.977	(11.7)	[-0.28]	0.283	(2.92)	0.631
9:05	-0.000	(-0.28)	0.956	(12.2)	[-0.57]	0.115	(1.32)	0.649
パネル B: 被説明変数:固有リターン ($\Delta p^{Ii}_{c,d}$)								
時刻 t	定数項		$\Delta p^{Ii}_{t,d}$			$L^i_d \Delta p^{Ii}_{t,d}$		R^2
8:55	-0.000	(-6.30)	0.479	(20.1)	[-21.9]	0.039	(1.87)	0.233
8:56	-0.000	(-6.36)	0.510	(20.6)	[-19.8]	0.024	(1.25)	0.249
8:57	-0.000	(-5.69)	0.548	(20.9)	[-17.3]	0.003	(0.17)	0.267
8:58	-0.000	(-4.91)	0.603	(21.2)	[-14.0]	-0.032	(-1.48)	0.293
8:59	-0.000	(-2.98)	0.751	(31.1)	[-10.3]	-0.133	(-5.93)	0.336
9:00	-0.000	(-6.28)	1.008	(52.6)	[0.40]	0.167	(5.00)	0.353
9:01	-0.000	(-6.39)	1.000	(55.4)	[-0.01]	0.212	(6.10)	0.370
9:02	-0.000	(-6.50)	0.998	(58.7)	[-0.12]	0.233	(6.82)	0.384
9:03	-0.000	(-6.45)	0.999	(59.0)	[-0.06]	0.228	(6.78)	0.395
9:04	-0.000	(-6.74)	1.009	(62.6)	[0.53]	0.221	(6.20)	0.407
9:05	-0.000	(-5.60)	1.003	(67.1)	[0.18]	-0.024	(-1.03)	0.447
パネル C: 被説明変数: トータルリターン ($\Delta p^{Ti}_{c,d}$)								
時刻 t	定数項		$\Delta p^{Ti}_{t,d}$			$L^i_d \Delta p^{Ti}_{c,d}$		R^2
8:55	0.000	(0.44)	0.524	(5.06)	[-4.60]	0.116	(1.64)	0.313
8:56	0.000	(0.32)	0.56	(5.43)	[-4.28]	0.086	(1.24)	0.330
8:57	0.000	(0.25)	0.601	(5.71)	[-3.79]	0.055	(0.77)	0.346
8:58	0.000	(0.01)	0.672	(6.39)	[-3.11]	-0.001	(-0.02)	0.372
8:59	-0.000	(-0.57)	0.864	(12.2)	[-1.91]	-0.144	(-2.09)	0.420
9:00	-0.000	(-0.77)	1.054	(19.2)	[0.98]	0.270	(2.52)	0.448
9:01	-0.000	(-0.58)	1.029	(19.0)	[0.53]	0.307	(2.78)	0.462
9:02	-0.000	(-0.34)	1.022	(20.0)	[0.43]	0.322	(2.88)	0.476
9:03	-0.000	(-0.24)	1.015	(20.5)	[0.30]	0.323	(2.96)	0.484
9:04	-0.000	(-0.15)	1.017	(22.1)	[0.38]	0.319	(2.83)	0.495
9:05	-0.000	(-0.18)	1.014	(22.6)	[0.32]	0.022	(0.28)	0.518

2004 年から 2005 年において，継続的に日経平均採用銘柄であり ADR が取引されていない 170 銘柄について，unbiasedness regression の推計結果を示している．市場リターンは 170 銘柄の時価総額の変化率，トータルリターンは個別銘柄の収益率，固有リターンは個別銘柄のインデックス・モデルの残差である．( ) 内は係数がゼロと有意に異なるかを検定するための t 値，[ ] 内は係数が 1 と有意に異なるかを検定するための t 値である．固有リターンおよびトータルリターンの場合は銘柄ごとに推計し，係数の平均および R^2 の平均，残差相関修正後の t 値を報告している．

取引開始前における価格制限の影響について，固有リターンとトータルリターンでは，リターンと特別気配ダミーとのクロス項の係数の推計値は，8時55分から8時58分まではゼロと有意に異ならないが，8時59分には負となり，しかもゼロと有意に異なる．この結果は，価格制限が予想されるとき，取引開始直前に気配が過剰に変動する，または気配が過剰に変動するとき価格制限にかかりやすい，ということを示唆している．市場リターンについては，取引開始前のクロス項の係数の推計値は正でゼロと有意に異ならず，過剰反応は観察されない．但し，この結果は，市場リターンについての特別気配ダミーの定義に依存している．表2パネルAでは，市場全体についての特別気配ダミー L_d^M は，第4.8節で定義したように，9時1分までに最初の約定があった銘柄が170銘柄中148銘柄以下である日に1をとるダミー変数としている．これに対して例えば，148銘柄ではなく100銘柄以下である日に1をとると特別気配ダミーの定義を変更した場合，8時59分におけるリターンと特別気配ダミーとのクロス項の係数の推計値は，有意ではないものの負となる．

取引開始後の価格制限の影響について，リターンと特別気配ダミーとのクロス項の係数の推計値は，市場リターン・固有リターン・トータルリターンのいずれの場合も，9時0分から9時4分まで正でゼロと有意に異なり，価格制限が緩和される9時5分にゼロと有意に異ならない値にまで低下している．これは取引ルールから予想される結果である．

6 取引開始前の気配更新

本節では，取引開始前の気配更新に関する仮説2「価格変化が大きい日に，取引開始前に気配が頻繁に更新される」についての分析を行う．本稿では，価格変化が大きな日を価格発見が困難である日とするが，第6.1節でこの理由を説明し，また具体的に変数の定義を行う．そのうえで，寄前気配更新の平均的な頻度を示し，さらに価格変化と気配更新についての回帰分析の推計結果を報告する．

6.1 価格発見の困難さの指標

価格発見が困難であるとき，寄前気配を通じた価格発見活動がより重要となる可能性がある．そのため，仮説2のように，価格発見が困難であるとき寄前気配が活発に更新されるかを検証する．同時に仮説3のように，価格発見が困難であるときであっても，寄前気配を通じた価格発見を行うことで，価格発見が容易であるときと同様に充分な価格発見が行われているかを検証する．価格発見は，本源的価値の変化が大きいときにより難しいと考えられる．また仮説1の検証結果より，9時0分の気配は平均的に情報効率的であり，本源的価値に近いと考えられる．そのため本稿では，前日14時55分から取引開始9時0分までの市場リターンおよび固有リターンを，寄付きにおける価格発見の難しさの代理変数とする．8時59分以前には，前日から当日9時0分までの価格変化率が不明であるもとで価格発見を行う，または不明であるので価格発見を行うが，事後的に価格変化が大きかった日を価格発見が困難であった日と考え，その日にどのように価格発見を行ったかを分析することになる．

価格発見の難しさの代理変数として，前日14時55分から当日9時0分までのリターンを用いるのは，以下の4つの理由による．(1) 前節の分析から，特別気配が配信されていなければ，9時0分の価格は情報効率的である．(2) 寄付きの価格発見の分析であるので，寄付きから大きく時間が経過しない方が望ましい．(3)8時59分以前までのリターンを用いる場合には，寄前気配に含まれるノイズの影響を受けてしまう．また，(4) 分析の主な対象は8時58分や8時59分であり，9時0分という2分先または1分先までの価格変化に対して完全予見を仮定することに重大な問題はないと考えられる．

表 3: 価格変化と累積気配更新頻度

	V_d^M	0	1	0	1
時刻 t	V_d^{Ii}	0	0	1	1
8:55		49.0	53.2	54.8	57.9
8:56		52.1	56.8	58.6	62.0
8:57		55.7	60.8	62.8	66.4
8:58		60.0	65.6	67.8	71.7
8:59		67.3	72.9	75.5	79.4

2004 年から 2005 年において，継続的に日経平均採用銘柄であり ADR が取引されていない 170 銘柄の，寄前気配配信開始から各時刻までの累積の気配更新頻度を示している．市場リターンの大小および固有リターンの大小により条件付けているが，V_d^M は前日 14 時 55 分から当日 9 時 0 分までの市場リターンが大きな日に 1 をとるダミー変数，V_d^{Ii} は銘柄 i の前日 14 時 55 分から当日 9 時 0 分までの固有リターンが大きな日に 1 をとるダミー変数である．リターンの大小により取引日をグループ分けしたうえで，各銘柄各時刻ごとに累積の気配更新頻度の平均を求め，さらに銘柄間における平均値を示している．

価格変動と気配更新との関係を調べるために，1 日の価格変化の大きさによって取引日を 4 つにグループ分けしたうえで，8 時 55 分から 8 時 59 分までの気配更新頻度の平均を計測する．グループ分けは，市場リターンの絶対値が大きい日に 1 をとるダミー変数 V_d^M，および銘柄 i の固有リターンの絶対値が大きい日に 1 をとるダミー変数 V_d^{Ii} を用いる．V_d^M は，前日 14 時 55 分から当日 9 時 0 分までの市場リターン $\Delta p_{9:00,d}^M$ を用い，$d = 1, \cdots, D$ の観測値より平均 $\mu_{9:00}^{pM}$ と標準偏差 $\sigma_{9:00}^{pM}$ を計算し，平均からの乖離の絶対値が標準偏差の半分を超える $|\Delta p_{9:00,d}^M - \mu_{9:00}^{pM}| > \sigma_{9:00}^{pM}/2$ である日に 1，それ以外の日にゼロをとるダミー変数である．V_d^{Ii} は，銘柄 i について前日 14 時 55 分から当日 9 時 0 分までの固有リターン $\Delta p_{9:00,d}^{Ii}$ を用い，まず $d = 1, \cdots, D$ の観測値より標準偏差 $\sigma_{9:00}^{pi}$ を計算し，固有リターンの絶対値がその標準偏差の半分を超える $|\Delta p_{9:00,d}^{Ii}| > \sigma_{9:00}^{pi}/2$ である日に 1，それ以外の日にゼロをとるダミー変数である．さらに (V_d^M, V_d^{Ii}) により，取引日を $(0,0)$，$(1,0)$，$(0,1)$，$(1,1)$ の 4 つに分類し，共通要因により価格が大きく動いたか，固有要因で価格が大きく動いたか，を区別する．

本稿では，当日 9 時 0 分までのリターンを価格発見の困難さの代理変数とするが，仮説 3 の分析ではリータンを直接用いず，前段で説明したダミー変数 V_d^M および V_d^{Ii} を用いる．リターンの大小と本源的価値の変化率の大小が多くの日で逆転するほどダミー変数の誤差が大きくなるが，基本モデルの推計結果より 9 時 0 分の気配は平均的に情報効率的であるため，誤差が非常に大きな訳ではないと予想される．また価格制限により，本源的価値が大きく変化するときにリターンが過小になるが，価格制限にかからない日のリターンよりもリターンを小さくする訳ではないので，V_d^M および V_d^{Ii} は，価格制限から大きなバイアスを受けない．以上のように，特に仮定 3 の検証では，価格変化の大きさを価格発見の難しさの代理変数とすることから発生する問題は，より小さいと考えられる．

価格発見の困難さに関し，例えば前日取引終了後から当日取引開始前までに何らかの情報の公表がある場合に価格発見が困難である，として分析する方法が考えられる．例えば取引開始直前の企業情報や政府統計の公表は，価格発見の困難さに影響を与える可能性がある．しかし，特に銘柄固有のリータンについて，分析のために十分な数の情報公開を特定することは難しいと考えられる．そのため本稿では，価格の変化率を価格発見の困難さの代理変数とする．企業の情報公開や政府統計の公表が，寄付きの価格形成にどのような影響を与えるかについての分析は，今後の課題である．

6.2 気配更新頻度

表 3 は，(V_d^M, V_d^{Ii}) を用いて取引日を分類したうえで，寄前気配の配信開始から時刻 t までの累積の気配更新頻度の平均を銘柄ごと求め，さらに銘柄間の平均を計算した値を示している．表より，

取引開始が近づくほど投資家が活発に注文を出し，気配が頻繁に更新されていることがわかる．例えば，価格変化の小さな $(V_d^M, V_d^{Ii}) = (0,0)$ の日には，8 時 55 分までの累積の気配更新頻度は平均 49.0 回あったところ，8 時 56 分では 52.1 回であり，8 時 56 分台の 1 分間に最良気配を動かす注文が 3 回程度発注されていることがわかる．これが 8 時 59 分台には 7 回程度に増加する．

価格変化と気配更新の活発さの関連について，共通要因により価格が大きく変化する日も，銘柄固有要因により価格が大きく変化する日も，そうでない日に対して気配が活発に更新されている．具体的には，$V_d^M = 0$ または $V_d^M = 1$ のもとで，いずれの時刻においても $V_d^{Ii} = 0$ より $V_d^{Ii} = 1$ の場合に気配の更新頻度が高い．同様に $V_d^{Ii} = 0$ または $V_d^{Ii} = 1$ のもとで，$V_d^M = 0$ より $V_d^M = 1$ の場合に気配の更新頻度が高い．この結果は，仮説 2 と整合的である．

6.3 価格変化と気配更新

仮説 2 を検証するため，被説明変数を市場更新頻度，固有更新頻度または総更新頻度とし，説明変数を価格変化の指標とした回帰分析を行った．本来の説明変数は，価格発見の難しさの指標であるが，代理変数として，前日 14 時 55 分から当日 9 時 0 分の市場リターンの絶対値 $|\Delta p^M_{9:00,d}|$ および前日 14 時 55 分から当日 9 時 0 分の固有リターンの絶対値 $|\Delta p^{Ii}_{9:00,d}|$ を用いる．これらリターンは，V_d^M および V_d^i の定義に用いたリターンである．

市場更新頻度 $q^M_{t,d}$ と価格変化との関係を調べるため，取引日 $d = 1, \ldots, D$ の観測値を用いて，

$$q^M_{8:59,d} = \alpha + \gamma |\Delta p^M_{9:00,d}| + \epsilon_d \tag{7}$$

を推計する．また固有更新頻度 $q^{Ii}_{t,d}$ と総更新頻度 $q^{Ti}_{t,d}$ について，銘柄 i ごと，取引日 $d = 1, \ldots, D$ の観測値を用いて，

$$q^{Ii}_{8:59,d} = \alpha + \beta |\Delta p^{Ii}_{9:00,d}| + \gamma |\Delta p^M_{9:00,d}| + \epsilon_d \tag{8}$$

$$q^{Ti}_{8:59,d} = \alpha + \beta |\Delta p^{Ii}_{9:00,d}| + \gamma |\Delta p^M_{9:00,d}| + \epsilon_d \tag{9}$$

を推計する．

仮説 2 の検証において，市場更新頻度の推計における帰無仮説は，市場リターンの絶対値の係数 γ がゼロ，対立仮説は係数が正，である．固有更新頻度の推計における帰無仮説は，固有リターンの絶対値の係数 β がゼロ，対立仮説は係数が正，である．総更新頻度の推計における帰無仮説は，固有リターンの絶対値の係数 β および市場リターンの絶対値の係数 γ がゼロ，対立仮説はいずれかの係数が正，である．

表 4 が推計結果を示している．括弧内の数値は t 値である．固有更新頻度および総更新頻度の推計では，各銘柄ごとに推計を行い，係数の推計値の平均および決定係数の平均を示しており，また括弧内には残差相関修正後の t 値を示している．仮説 2 の帰無仮説は，市場更新頻度および固有更新頻度では有意水準 1%で棄却される．総更新頻度では，固有リターンの絶対値の係数ゼロの帰無仮説は有意水準 1%で棄却され，市場リターンの絶対値の係数ゼロの帰無仮説も有意水準 1%で棄却される．2 つの係数がともにゼロであるという帰無仮説について各銘柄ごと F 検定を行うと，F 値の平均は 16.3 であり，また 92.9%の銘柄で帰無仮説は有意水準 1%で棄却される．以上より，気配の更新は仮説 2 と整合的であり，市場リターンの絶対値が大きいときに市場全体で気配の更新が活発になっており，固有リターンの絶対値が大きいときに銘柄独自に気配の更新が活発になっている，といえる．

表 4: 価格変化と寄前気配の更新

被説明変数	定数項		固有リターンの絶対値 ($\lvert\Delta p^{Ii}_{9:00,d}\rvert$)		市場リターンの絶対値 ($\lvert\Delta p^{M}_{9:00,d}\rvert$)		R^2
市場更新頻度 ($q^{M}_{8:59,d}$)	9.370	(567.3)			19.17	(5.60)	0.065
固有更新頻度 ($q^{Ii}_{8:59,d}$)	-0.071	(-8.45)	14.38	(15.6)	-0.405	(-0.25)	0.054
総更新頻度 ($q^{Ti}_{8:59,d}$)	4.033	(17.25)	16.93	(3.90)	17.47	(3.09)	0.068

2004 年から 2005 年において，継続的に日経平均採用銘柄であり ADR が取引されていない 170 銘柄について，8 時 59 分までの気配更新を被説明変数，前日 14 時 55 分から当日 9 時 0 分までの価格変化を説明変数とした回帰モデルの推計結果を示している．括弧内は t 値である．被説明変数が固有更新頻度および総更新頻度の場合，銘柄ごと推計し，係数の平均および R^2 の平均，残差相関修正後の t 値を報告している．

7 気配更新が価格発見に与える影響

本節では，仮説 3「取引開始時の価格形成は，価格変化・気配の更新頻度の影響を受けない」についての分析を行う．まず回帰モデルを説明したうえで推計結果を報告し，その後に頑健性について議論する．

7.1 回帰モデル

第 5 節の基本モデルでは，前日からのリターン，および特別気配ダミーとリターンとのクロス項を説明変数とした．本節における unbiasedness regression では，さらに，価格変化および気配更新の活発さを表すダミー変数とリターンとのクロス項を説明変数に追加し，推計を行う．

市場リターンの推計では，ダミー変数 $H^M_{t,d}$，V^M_d，$H^M_{t,d}V^M_d$ を用いる．V^M_d は，第 4.8 節で定義したように，市場リターンの絶対値が大きい日に 1 をとるダミー変数である．$H^M_{t,d}$ は，時刻 t までの市場全体の気配更新が通常の日よりも多いことを表すダミー変数である．具体的には，市場更新頻度 $q^M_{t,d}$ を用い，各時刻 t ごと，$d = 1, \cdots, D$ の観測値を用いて $q^M_{t,d}$ の平均 μ^{qM}_t と標準偏差 σ^{qM}_t を計算し，$q^M_{t,d}$ がその平均プラス標準偏差の半分を超えた $q^M_{t,d} > \mu^{qM}_t + \sigma^{qM}_t/2$ のときに 1，それ以外はゼロをとるダミー変数とする．推計するモデルは，

$$\Delta p^M_{c,d} = \alpha + \beta_1 \Delta p^M_{t,d} + \beta_2 H^M_{t,d}\Delta p^M_{t,d} + \beta_3 V^M_d \Delta p^M_{t,d} + \beta_4 H^M_{t,d}V^M_d \Delta p^M_{t,d} + \gamma L^M_d \Delta p^M_{t,d} + \epsilon_d \quad (10)$$

である．

固有リターンの推計では，ダミー変数 $H^{Ii}_{t,d}$，V^i_d，$H^{Ii}_{t,d}V^i_d$ を用いる．V^{Ii}_d は，第 4.8 節で定義したように，各銘柄の固有リターンの絶対値が大きい日に 1 をとるダミー変数である．$H^{Ii}_{t,d}$ は，銘柄固有の要因により時刻 t までの気配更新が活発であるときに 1 をとるダミー変数である．具体的には，固有更新頻度 $q^{Ii}_{t,d}$ を用い，銘柄 i 時刻 t のそれぞれについて，各日の観測値を用いて $q^{Ii}_{t,d}$ の標準偏差 σ^{qIi}_t を求め，$q^{Ii}_{t,d}$ が標準偏差の半分を超えた $q^{Ii}_{t,d} > \sigma^{qIi}_t/2$ のときに 1，そうでないときにゼロをとるダミー変数とする．推計するモデルは，

$$\Delta p^{Ii}_{c,d} = \alpha + \beta_1 \Delta p^{Ii}_{t,d} + \beta_2 H^{Ii}_{t,d}\Delta p^{Ii}_{t,d} + \beta_3 V^{Ii}_d \Delta p^{Ii}_{t,d} + \beta_4 H^{Ii}_{t,d}V^{Ii}_d \Delta p^{Ii}_{t,d} + \gamma L^i_d \Delta p^{Ii}_{t,d} + \epsilon_d \quad (11)$$

である．

トータルリターンの推計では，ダミー変数 $H^{Ti}_{t,d}$，V^{Ti}_d，$H^{Ti}_{t,d}V^{Ti}_d$ を用いる．$H^{Ti}_{t,d}$ は，各銘柄の気配更新が通常よりも頻繁であるとき 1 をとるダミー変数である．このダミー変数は，銘柄 i 時刻 t ごとに，総更新頻度 $q^{Ti}_{t,d}$ の各日の観測値を用いて平均 μ^{qTi}_t および標準偏差 σ^{qTi}_t を計算し，$q^{Ti}_{t,d}$ が平均プラス標準偏差の半分を超えた $q^{Ti}_{t,d} > \mu^{qTi}_t + \sigma^{qTi}_t/2$ のときに 1，それ以外のときにゼロをとるダミー変数である．V^{Ti}_d は市場リターンと固有リターンの絶対値のいずれかが大きいときに 1，それ

以外はゼロをとるダミー変数であり，$V_d^{Ti} = 1 - (1 - V_d^M)(1 - V_d^{Ii})$ と定義する．これらダミー変数を用いて，

$$\Delta p_{c,d}^{Ti} = \alpha + \beta_1 \Delta p_{t,d}^{Ti} + \beta_2 H_{t,d}^{Ti} \Delta p_{t,d}^{Ti} + \beta_3 V_d^{Ti} \Delta p_{t,d}^{Ti} + \beta_4 H_{t,d}^{Ti} V_d^M \Delta p_{t,d}^{Ti} + \gamma L_d^i \Delta p_{t,d}^{Ti} + \epsilon_d \quad (12)$$

を推計する．

仮説3の検証における帰無仮説は次のようになる．表2の基本モデルの推計結果より，$j \in \{M, Ii, Ti\}$ のそれぞれについて，取引開始直後の9時0分における $\Delta p_{t,d}^j$ の係数の推計値は1と有意に異ならない．この推計値を基準にし，価格変化の大きい場合や気配の更新が活発な場合にも，$\Delta p_{t,d}^j$ の係数の推計値が1と有意に異ならないかを検証する．帰無仮説は，9時0分において，$\Delta p_{t,d}^j$ の係数が1，かつ $H_{t,d}^j \Delta p_{t,d}^j$ の係数がゼロ，かつ $V_d^j \Delta p_{t,d}^j$ の係数がゼロ，かつ $H_{t,d}^j V_d^j \Delta p_{t,d}^j$ の係数がゼロであり，対立仮説はいずれかの係数が上記と異なる，である．もし unbiasedness regression の係数の推計値にバイアスがあったとしても，価格変化の大きな日や気配更新が活発な日にも同様にバイアスが発生するのであれば，ダミー変数とリターンとのクロス項の係数がゼロであるかを検定することで，日によって気配の情報効率性に相違があるかを検証することができる．

7.2 推計結果

推計結果を表5にまとめている．係数の推計値の下の ( ) 内には推計値がゼロと異なるかを検定するためのt値を示しており，[ ] 内には推計値が1と異なるかを検定するためのt値を示している．F値は，仮説3についてF検定を行った場合の値である．固有リターンおよびトータルリターンの推計では，各銘柄ごと推計し，係数の推計値の平均，決定係数の平均，および仮説3についてのF値の平均を報告している．t値は，残差相関修正後の値である．F値の下の <> 内には，仮説3に関するF検定において，帰無仮説を有意水準5%で棄却した銘柄の比率を示している．

市場リターンについて，表5パネルAが示すように，8時59分前は，$V_d^M \Delta p_{t,d}^M$ の係数の推計値が正でゼロと有意に異なる．これは，価格変化が大きな日において，取引開始前の気配は，寄付きにおける大きな価格変化を十分に読み込んでおらず，相対的に気配の調整に遅延が発生していることを示している．しかし8時59分から9時1分にかけて，$\Delta p_{t,d}^M$ の係数の推計値は1と有意に異ならず，また $\Delta p_{t,d}^M$ とダミー変数とのクロス項の係数の推計値は，特別気配ダミー L_d^M とのクロス項を除いて，それぞれゼロと有意に異ならない．仮説3についてのF検定では，9時0分に有意水準5%で帰無仮説を棄却しない．この結果は，価格変化の大きさおよび気配の更新頻度にかかわらず，9時0分には共通要因についての価格発見が完了していることを示唆しており，仮説3と整合的である．

固有リターンについての表5パネルBの結果より，8時59分の $\Delta p_{t,d}^{Ii}$ の係数の推計値は0.496で1と有意に異なり，$V_d^{Ii} \Delta p_{t,d}^{Ii}$ の係数の推計値は0.278でゼロと有意に異なる．また80%以上の銘柄で仮説3についての帰無仮説が有意水準5%で棄却される．この結果は，取引開始直前までに固有要因に関する価格発見は完了せず，特に価格変化が大きいときに気配がゆっくりと反応する傾向があることを示している．一方，9時0分および9時1分における $\Delta p_{t,d}^{Ii}$ の係数の推計値は1と有意に異ならず，$\Delta p_{t,d}^{Ii}$ とダミー変数とのクロス項の係数の推計値は，特別気配ダミー L_d^i とのクロス項を除いて，それぞれゼロと有意に異ならない．仮説3についてのF検定において，有意水準5%で帰無仮説を棄却する銘柄の比率は，8時59分において85%であるのに対し，9時0分には約25%であり，大幅な低下が観察される．以上より，固有要因の価格発見は9時0分にはほぼ完了している，と考えられる．これは，仮説3と整合的な結果である．パネルCのトータルリターンに関する推計結果は，市場リターンの結果と固有リターンの結果の中間であり，トータルリータンについても仮説3と整合的である．

表 5: 価格変化と気配更新に関する推計結果

パネル A: 被説明変数:市場リターン ($\Delta p^M_{c,d}$)								
時刻 t	定数項	$\Delta p^M_{t,d}$	$H^M_{t,d}\Delta p^M_{t,d}$	$V^M_d\Delta p^M_{t,d}$	$H^M_{t,d}V^M_d\Delta p^M_{t,d}$	$L^M_d\Delta p^M_{t,d}$	R^2	F 値
8:57	0.001	0.253	0.163	0.470	-0.271	-0.057	0.490	16.7
	(2.27)	[-7.35]	(0.52)	(3.71)	(-0.86)	(-0.64)		
8:58	0.001	0.304	0.134	0.442	-0.253	-0.041	0.507	13.8
	(1.96)	[-6.60]	(0.42)	(3.38)	(-0.78)	(-0.45)		
8:59	0.000	0.801	-0.208	0.039	0.106	0.040	0.615	2.33
	(0.24)	[-1.23]	(-0.62)	(0.22)	(0.31)	(0.46)		
9:00	-0.000	1.335	-0.532	-0.299	0.378	0.382	0.625	1.24
	(-1.73)	[1.39]	(-1.24)	(-1.17)	(0.86)	(3.35)		
9:01	-0.000	1.12	-0.402	-0.109	0.252	0.363	0.624	0.84
	(-1.02)	[0.54]	(-0.96)	(-0.45)	(0.58)	(3.32)		
パネル B: 被説明変数:固有リターン ($\Delta p^{Ii}_{c,d}$)								
時刻 t	定数項	$\Delta p^{Ii}_{t,d}$	$H^{Ii}_{t,d}\Delta p^{Ii}_{t,d}$	$V^{Ii}_d\Delta p^{Ii}_{t,d}$	$H^{Ii}_{t,d}V^{Ii}_d\Delta p^{Ii}_{t,d}$	$L^i_d\Delta p^{Ii}_{t,d}$	R^2	F 値
8:57	-0.000	0.255	0.015	0.346	0.023	-0.032	0.287	33.8
	(-5.58)	[-31.7]	(0.32)	(12.5)	(0.45)	(-1.56)		<99.4>
8:58	-0.000	0.319	0.025	0.333	0.006	-0.064	0.311	26.8
	(-4.93)	[-25.8]	(0.48)	(11.1)	(0.10)	(-3.04)		<97.1>
8:59	-0.000	0.496	0.042	0.278	-0.007	-0.147	0.348	9.08
	(-3.22)	[-15.3]	(0.61)	(7.76)	(-0.09)	(-6.54)		<84.1>
9:00	-0.000	1.077	0.133	-0.105	-0.042	0.157	0.358	1.69
	(-6.62)	[1.32]	(0.89)	(-1.76)	(-0.27)	(4.67)		<24.7>
9:01	-0.000	1.016	0.067	-0.042	0.003	0.203	0.376	1.70
	(-6.47)	[0.35]	(0.59)	(-0.92)	(0.03)	(5.79)		<18.2>
パネル C: 被説明変数: トータルリターン ($\Delta p^{Ti}_{c,d}$)								
時刻 t	定数項	$\Delta p^{Ti}_{t,d}$	$H^{Ti}_{t,d}\Delta p^{Ti}_{t,d}$	$V^{Ti}_d\Delta p^{Ti}_{t,d}$	$H^{Ti}_{t,d}V^{Ti}_d\Delta p^{Ti}_{t,d}$	$L^i_d\Delta p^{Ti}_{t,d}$	R^2	F 値
8:57	0.000	0.351	-0.009	0.296	-0.048	0.047	0.356	24.3
	(0.43)	[-4.65]	(-0.03)	(2.06)	(-0.13)	(0.67)		<90.6>
8:58	0.000	0.451	-0.032	0.268	-0.030	-0.008	0.382	16.3
	(0.17)	[-3.52]	(-0.09)	(1.72)	(-0.08)	(-0.11)		<86.5>
8:59	-0.000	0.693	-0.159	0.198	0.113	-0.140	0.426	3.59
	(-0.45)	[-1.79]	(-0.35)	(1.17)	(0.25)	(-2.03)		<46.5>
9:00	-0.000	1.199	-0.200	-0.137	0.175	0.277	0.452	1.38
	(-0.78)	[0.77]	(-0.33)	(-0.50)	(0.29)	(2.56)		<16.5>
9:01	-0.000	1.101	-0.144	-0.058	0.102	0.315	0.467	1.40
	(-0.55)	[0.51]	(-0.23)	(-0.28)	(0.16)	(2.83)		<17.1>

2004 年から 2005 年において，継続的に日経平均採用銘柄であり ADR が取引されていない 170 銘柄について，unbiasedness regression の推計結果を示している．( ) 内は係数の推計値がゼロと有意に異なるかを検定するための t 値，[ ] 内は係数の推計値が 1 と有意に異なるかを検定するための t 値である．F 値は，仮説 3 について F 検定を行った場合の値である．被説明変数が固有リターンおよびトータルリターンの場合は，銘柄ごとに推計し，係数の推計値の平均，R^2 の平均，残差相関修正後の t 値，F 値の平均を報告している．F 値の下の <> 内は，仮説 3 に関する F 検定において，帰無仮説を有意水準 5%で棄却した銘柄の比率である．

表 6: 固有リターンの推計における頑健性

時刻 t	定数項	$\Delta p^{Ii}_{t,d}$	$H^{Ii}_{t,d}\Delta p^{Ii}_{t,d}$	$V^{Ii}_{d}\Delta p^{Ii}_{t,d}$	$H^{Ii}_{t,d}V^{Ii}_{d}\Delta p^{Ii}_{t,d}$	$L^{i}_{d}\Delta p^{Ii}_{t,d}$	R^2	F 値
ケース 1: 固有リターンをトータルリターンと市場リターンの差とする								
8:59	-0.000	0.475	0.031	0.300	-0.005	-0.129	0.359	9.2
	(-1.79)	[-14.7]	(0.42)	(8.10)	(-0.07)	(-5.33)		<87.1>
9:00	-0.000	1.079	0.217	-0.110	-0.141	0.194	0.370	1.7
	(-0.90)	[1.40]	(1.63)	(-1.94)	(-1.02)	(5.09)		<24.1>
ケース 2: 外れ値の除外なし								
8:59	0.000	0.538	0.048	0.268	-0.038	-0.260	0.402	11.55
	(0.99)	[-14.1]	(0.78)	(7.70)	(-0.57)	(-10.2)		<80.0>
9:00	-0.000	1.091	0.186	-0.106	0.002	0.332	0.404	2.23
	(-5.11)	[1.94]	(1.49)	(-2.18)	(0.02)	(9.66)		<31.2>
ケース 3: 同行リターンの 2 パーセンタイルと 98 パーセンタイルの外側を除外								
8:59	-0.000	0.458	0.046	0.286	-0.001	-0.107	0.318	9.67
	(-4.47)	[-16.6]	(0.67)	(7.84)	(-0.02)	(-4.68)		<87.6>
9:00	-0.000	1.028	0.113	-0.094	-0.038	0.096	0.325	1.69
	(-7.98)	[0.44]	(0.74)	(-1.46)	(-0.24)	(2.76)		<24.1>
ケース 4: 9 時 0 分に特別気配が配信されていた $L^{i}_{d}=1$ の場合に除外								
8:59	-0.000	0.466	0.024	0.308	0.028		0.306	9.27
	(-4.84)	[-15.0]	(0.30)	(7.85)	(0.34)			<80.6>
9:00	-0.000	1.080	0.062	-0.092	-0.020		0.322	1.66
	(-6.15)	[1.19]	(0.37)	(-1.36)	(-0.12)			<22.4>

2004 年から 2005 年において，継続的に日経平均採用銘柄であり ADR が取引されていない 170 銘柄について，unbiasedness regression の推計結果を示している．銘柄ごと推計し，係数の推計値の平均および R^2 の平均，残差相関修正後の t 値を報告している．但し ( ) 内は係数の推計値がゼロと有意に異なるかを検定するための t 値，[ ] 内は係数の推計値が 1 と有意に異なるかを検定するための t 値である．F 値は，銘柄ごと仮説 3 について F 検定を行った場合の値の平均である．F 値の下の <> 内は，仮説 3 に関する F 検定において，帰無仮説を有意水準 5%で棄却した銘柄の比率である．

取引開始前の気配を通じて価格操作が行われているとき，気配の活発な更新が気配の過剰反応と結びつく場合には，気配更新の活発さを表すダミー変数 $H^{j}_{t,d}(j \in \{M, Ii, Ti\})$ とリターンとのクロス項の係数が負になると予想される．しかし表 5 のいずれの推計結果でも，このダミー変数を含むクロス項の係数の推計値はゼロと有意に異ならない．一方，気配更新を被説明変数とした回帰分析の表 4 の推計結果は，価格変化が大きなときに気配の更新が活発であることを示している．これは，価格操作により気配が頻繁に更新され，よって価格が大きく変化しているためである可能性がある．もし価格操作が頻繁に行われているのであれば，取引開始後に価格の情報効率性が低下すると考えられるが，表 2 および表 5 の推計結果は，取引開始直後の価格形成は情報効率的であることを示している．以上の結果は，気配更新が活発になったときに価格操作が行われているというよりも，価格変化が大きく価格発見が困難であるときに，投資家が寄付き前に活発に発注・キャンセルを繰り返しながら寄前気配を通じて学習や情報交換を行いながら価格発見をしている，という見方と整合的である．但し，本稿の目的は，気配の更新と価格発見との平均的な関係を分析することにあり，価格操作に関するより詳細な分析は，今後の課題である．

7.3 頑健性

表 5 の推計結果のうち，特に固有リターンの分析において推計上の誤差が大きくなる可能性がある．しかし表 6 に示すように，次の 4 つ方法による推計結果も表 5 とほぼ同様であった．

まず固有リターンの推計そのものについて，固有リターンをインデックス・モデルの残差としているため，インデックス・モデル推計における誤差の影響を受ける．そのためケース 1 では，固有リターンをトータルリターンと市場リターンとの差と定義した場合の推計を行う．次に，表 5 の推計では，固有リターンの 1 パーセンタイルと 99 パーセンタイルの外側にある観測値を外れ値として分析対象から除外している．この影響をみるために，ケース 2 では，すべての観測値を用いて推計

を行い，ケース3では，2パーセンタイルと98パーセンタイルの外側にある観測値を分析対象から外した推計を行う．最後にケース4では，特別気配の配信が取引開始直前の気配に影響を与えていることを考慮して，9時0分に特別気配が配信されていた観測値をすべて除外した推計を行う．この場合には，特別気配ダミーとリターンとのクロス項は回帰モデルから外される．

固有リターンについて以上の4つのケースの推計を行った結果を表6にまとめている．掲載している情報は表5と同一である．表6の推計結果は，いずれのケースも表5パネルBとほぼ同様であり，推計において固有リターンの計算方法および外れ値の影響は大きくないと考えられる．

8 おわりに

本稿では，東証における取引開始前の価格発見について，unbiasedness regressionによる分析を行った．他の取引所と同様，取引開始前に配信されている気配はノイズではなく新情報が反映している．また，価格変化が大きく価格発見が困難であると考えられる日において，取引開始前に投資家が活発に注文を出したりキャンセルしたりすることにより気配が頻繁に更新されている．さらに，価格の情報効率性は価格変化の大きさや寄前気配の更新頻度の影響を受けず，取引開始直後に価格発見はほぼ完了し，情報効率的な価格形成がなされている．

本稿の分析より，寄前気配の更新頻度は，価格発見と結びついていることがわかった．取引開始前の価格発見が困難であるとき，注文やキャンセルを活発に繰り返すことを通じて学習または情報交換を行い，本源的価値を探索していると推測される．しかし，寄前気配から具体的にどのように学習および情報交換をしているかに関しては不明である．そのため，例えば取引開始前における大口注文の発注や注文のキャンセルが価格形成にどのような影響を与えているかを分析する必要がある．また本研究では，日経平均株価採用銘柄のうちADRが取引されていない銘柄に分析対象を限定している．株価指数算出への採用やADRの取引が，取引開始時の価格発見にどのような差異をもたらしているかについての分析も，今後の課題として残されている．

参考文献

Amihud, Y., Mendelson, H. (1991). Volatility, efficiency, and trading - evidence from the Japanese stock-market. *Journal of Finance* **46**, 1765–1789

Barclay, M.J., Hendershott, T. (2003). Price discovery and trading after hours. *Review of Financial Studies* **16**, 1041–73

Barclay, M.J., Hendershott, T. (2008). A comparison of trading and non-trading mechanisms for price discovery. *Journal of Empirical Finance* **15**, 839–849

Biais, B., Bisiere, C., Pouget, S. (2009). Equilibrium discovery and preopening mechanisms in an experimental market. Mimeo.

Biais, B., Hillion, P., Spatt, C. (1999). Price discovery and learning during the preopening period in the Paris Bourse. *Journal of Political Economy* **107**, 1218–48

Cao, C., Ghysels, E., Hatheway, F.M. (2000). Price discovery without trading: evidence form the Nasdaq preopening. *Journal of Finance* **55**, 1339–65

Cao, C., Hansch, O., Wang, X.X. (2009). The information content of an open limit-order book. *Journal of Futures Markets* **29**, 16–41

Chang, R.P., Rhee, S.G., Stone, G.R., Tang, N. (2008). How does the call market method affect price efficiency? evidence from the Singapore stock market. *Journal of Banking and Finance* **32**, 2205–19

Chen, T., Cai, J., Ho, R.Y.K. (2009). Intraday information efficiency on the Chinese equity market. *China Economic Review* **20**, 527–41

Chordia, T., Roll, R., Subrahmanyam, A. (2008). Liquidity and market efficiency. *Journal of Financial Economics* **87**, 249–68

Comerton-Forde, C., Rydge, J. (2006). The influence of call auction algorithm rules on market efficiency. *Journal of Financial Markets* **9**, 199–222

Cushing, D., Madhavan, A. (2000). Stock returns and trading at the close. *Journal of Financial Markets* **3**, 45–67

Fama, E. F., French, K. R. (1993). Common risk-factors in the returns on stocks and bonds. *Journal of Financial Economics* **33**, 3–56.

George, T.J., Hwang, C.-Y. (1995). Transitory price changes and price-limit rules: evidence from the Tokyo Stock Exchange. *Journal of Financial and Quantitative Analysis* **30**, 313–27

Hillion, P., Suominen, M. (2004). The Manipulation of closing prices. *Journal of Financial Markets* **7**, 351–75

Jordan, J.S. (1982). A Dynamic-model of expectations equilibrium. *Journal of Economic Theory* **28**, 235–254

Kobayashi, T. (1977). A convergence theorem on rational expectations equilibrium with price information. Mimeo

Kyle, A.S. (1985). Continuous auctions and insider trading. *Econometrica* **53**, 1315–35

Kyle, A.S. (1989). Informed speculation with imperfect competition. *Review of Economic Studies* **56**, 317–356

Madhavan, A., Panchapagesan, V. (2000). Price discovery in auction markets: a Look inside the black box. *Review of Financial Studies* **13**, 627–58

Madhavan, A., Richardson, M., Roomans, M. (1997). Why do security prices change? a transaction-level analysis of NYSE stocks. *Review of Financial Studies* **10**, 1035–64

Subrahmanyam, A. (1997). The ex ante effects of trade halting rules on informed trading strategies and market liquidity. *Review of Financial Economics* **6**, 1–14

Tse, Y. (1999). Price discovery and volatility spillovers in the DJIA index and futures markets. *Journal of Futures Markets 19*, 911–930.

Vives, X. (1995). The speed of information revelation in a financial market mechanism. *Journal of Economic Theory* **67**, 178–204

A Robust Estimation of Realized Volatility and Covariance with Micro-market Adjustments and Round-off Errors *

Seisho Sato †
and
Naoto Kunitomo ‡

June 29, 2011

Abstract

For estimating the realized volatility and covariance by using high frequency data, Kunitomo and Sato (2008a,b) have proposed the Separating Information Maximum Likelihood (SIML) method when there are micro-market noises. The SIML estimator has reasonable asymptotic properties; it is consistent and it has the asymptotic normality (or the stable convergence in the general case) when the sample size is large under general conditions with *non-Gaussian processes* or *volatility models*. We show that the SIML estimator has the asymptotic robustness properties in the sense that it is consistent and has the asymptotic normality when there are micro-market (non-linear) adjustments and the round-off errors on the underlying stochastic processes.

Key Words

Realized Volatility with Micro-Market Noise, High-Frequency Data, Separating Information Maximum Likelihood (SIML), micro-market adjustments, Round-off errors, Robustness.

---

*KSIII-11-7-1. This is a preliminary memorandum on our research project. We thank Wataru Ohta for comments on the previous version

†Institute of Statistical Mathematics, Tachikawa-shi, Midori-cho 10-3, Tokyo 190-8562, JAPAN

‡Graduate School of Economics, University of Tokyo, Bunkyo-ku, Hongo 7-3-1, Tokyo 113-0033, JAPAN, kunitomo@e.u-tokyo.ac.jp

1. Introduction

Recently a considerable interest has been paid on the estimation problem of the realized volatility by using high-frequency data of financial price processes in financial econometrics. Since the earlier studies often had ignored the presence of micro-market noises in financial markets and there has been a consensus that the micro-market noises are important in high-frequency financial data, several new statistical estimation methods have been developed. See Bandorff-Nielsen et al. (2008) and Malliavin and Mancino (2009) for recent literatures on the related topics. In this respect Kunitomo and Sato (2008a, b) have proposed a new statistical method called the Separating Information Maximum Likelihood (SIML) estimation for estimating the realized volatility and the realized covariance by using high frequency data with the presence of possible micro-market noises. The SIML estimator has reasonable asymptotic properties; it is consistent and it has the asymptotic normality (or the stable convergence in the more general case) when the sample size is large and the data frequency interval is small under a set of regularity conditions for the *non-Gaussian* underlying processes and *volatility models*. Kunitomo and Sato (2010, 2011) have also shown that the SIML estimator has the robustness properties, that is, it is consistent and asymptotically normal even when the noise terms are autocorrelated and/or there are endogenous correlations between the market-noise terms and the (underlying) efficient market price process. There has been recent finance literature on the importance of these aspects in high frequency financial data including Engle and Sun (2007), for instance.

In this paper we shall investigate the robustness property of the SIML estimation when we have the micro-market adjustment mechanism and the round-off errors in the process of forming the observed prices. The micro-market models including the price adjustments have been discussed in the framework of *micro-market literature* in financial economics (Hansbrouck (2007), for instance). Among many micro-market models, we first take the (linear) adjustment model proposed by Amihud and Mendelson (1987) as a benchmark case. Then we shall extend it to the

nonlinear price adjustment models and we regard a continuous martingale as the hidden intrinsic value of underlying security. A new feature in this context to financial econometrics is to utilize the nonlinear (discrete) time series models and one possible non-linear model is the Simultaneous Switching Autoregressive (SSAR) model developed by Sato and Kunitomo (1996) and Kunitomo and Sato (1999). Also we shall consider the round-off error model as a non-linear transformation for financial price data. The problem of round-off error models has been recently investigated in statistics (Delattre and Jacod (1997), for instance). It reflects the common observation in actual financial markets that we have the tick-size effects (the minimum price change size and the minimum order size) and we often observe bid-ask spreads on securities in the stock markets.

In these problems there is a common econometric aspect that the observed price can be different from the underlying intrinsic value of the security and we can interpret this phenomenon as a nonlinear transformation from the intrinsic value to the observed prices. We can represent the present situation as the nonlinear statistical models of an unobservable (continuous-time) state process and the observed (discrete-time) stochastic process with measurement errors. When the effects of measurement errors are present, it seems that the existing statistical methods measuring the realized volatility and covariance have some problems to be fixed in various ways. They could handle the problem of our interest, but often they need some special consideration on the underlying mechanism of price process. On the contrary, we shall show that the SIML estimator is robust in these situations; that is, it is consistent and asymptotically normal as the sample size increases under a reasonable set of assumptions. The asymptotic robustness of the SIML method on the realized volatility and covariance has desirable properties over other estimation methods from a large number of data sets for the underlying continuous stochastic process with micro-market noise in the multivariate non-Gaussian cases. Because the SIML estimation is a simple method, it can be practically used for analyzing the multivariate (high frequency) financial time series.

In Section 2 we introduce the micro-market adjustment models and the round-off

error models. Then in Section 3 we explain the SIML estimation method and we give the asymptotic robustness properties of the SIML estimator when there are micro-market adjustments and the round-off error models. In Section 4 we shall report the finite sample properties of the SIML estimator based on a set of simulations. Finally, in Section 5 some brief remarks will be given. Some mathematical details of the proofs of theorems in Section 3 are given in Appendix A. Tables and figures based on simulations in Section 4 are given in Appendix B.

2. Micro-market adjustment Models and the Round-off error Models

2.1 A General Formulation

Let $y(t_i^n)$ be the $i-$th observation of the (log-) price at t_i^n for $j = 1, \cdots, p; 0 = t_0^n \leq t_1^n \leq \cdots \leq t_n^n = 1$. We set $\mathbf{y}_n = (y(t_i^n)$ be an $n \times 1$ vector of observations and we assume that the underlying (vector-valued) continuous process $X(t)$ $(0 \leq t \leq 1)$, which is not necessarily the same as the observed (log-)prices at t_i^n $(i = 1, \cdots, n)$ and

$$X(t) = X(0) + \int_0^t \sigma_x(s) dB_s \ \ (0 \leq t \leq 1), \tag{2.1}$$

where B_s is the standard Brownian motion, $\sigma_x(s)$ is a function adapted to the $\sigma-$field $\mathcal{F}(X_r, B_r, r \leq s)$, and the instantaneous volatility function is $\sigma_x(s)$. The main statistical objective is to estimate the quadratic variation

$$\sigma_x^2 = \int_0^1 \sigma_x(s) ds \tag{2.2}$$

of the underlying continuous process $X(t)$ $(0 \leq t \leq 1)$.

In this paper we consider the situation that the observed price $y(t_i^n)$ is different from the continuous martingale $X(t)$ and it a sequence of discrete stochastic process given by

$$y(t_i^n) = h\left(X(t), y(t_{i-1}^n), u(t_i^n)\ ,\ 0 \leq t \leq t_i^n\right)\ , \tag{2.3}$$

where the (unobservable) continuous martingale process $X(t)$ $(0 \leq t \leq 1)$ is defined by (2.1) and $u(t_i^n)$ is the micro-market noise process. For simplicity, we assume that $\mathcal{E}(u(t_i^n)) = 0$, $\mathcal{E}(u(t_i^n)^2) = \sigma_u^2$, and $h(\cdot)$ is a measurable function at $0 = t_0^n \leq t_1^n \leq \cdots \leq t_n^n = 1$ with $t_i^n - t_{i-1}^n = 1/n$ $(i = 1, \cdots, n)$.

There are several special cases of (2.1) and (2.3), which have some interesting aspects for practical applications on modeling the financial markets and the high frequency financial data. The simple (high-frequency) financial model with micro market noise can be represented by

$$y(t_i^n) = X(t_i^n) + u(t_i^n) \ , \tag{2.4}$$

where $y(t_i^n)$ $(i = 1, \cdots, n)$ is observable while the underlying continuous process $X(t)$ is given by (2.1).

The most important aspect of (2.4) is the fact that it is an additive measurement error model, which has been often assumed in the statistical literature. As we shall discuss in this section, however, there are some reasons that (2.4) is not enough for some applications. For instance, the hight-frequency financial models for micro-market price adjustments and the round-off-errors models for financial prices can be represented as some special cases of (2.1) and (2.3).

2.2 A Micro-market price adjustment model

There have been a large number of micro-market models in the area of financial economics in the past which have tried to explain the role of noise traders, insiders, bid-ask spreads, the transaction prices and the associated price adjustment processes. (See Hansbrouck (2007) for the detailed discussions on the major micro-market models in financial economics, for instance.) We illustrate the underlying arguments on the financial markets by showing Figures 2-1 and 2-2 in Appendix B. For this purpose, we denote that P and Q are the price and the quantity (demand,

supply and traded) of a security [1] . When the demand curve and supply curve for a security do not meet as Figure 2.1, there is no transaction occurred at the moment in a financial market. The minimum (desired) supply price level $\bar{P}$ is higher than the maximum (desired) demand price level $\underline{P}$, and then there is a (bid-ask) spread. When there were some information in the supply side indicating that the intrinsic value of a security at t could be less than the latest observed price at $t - \Delta t$ (i.e. $V_t - P_{t-\Delta t} < 0$, $\Delta t > 0$), however, the supply schedule would be shifted down-ward When, however, there were some information in the demand side indicating that the intrinsic value of a security at t could be higher than the latest observed price (i.e. $V_t - P_{t-\Delta t} > 0$), the demand schedule would be shifted up-ward. In these situations while the trade of a security occur at the price P^* and the quantity Q^* as in Figure 2.2, the financial market would be under pressure to further price changes.

We set $y_i = P(t_i^n)$ $(i = 1, \cdots, n)$ and $x_i = X(t_i^n)$ $(i = 1, \cdots, n)$ and we consider the (linear) micro-market price adjustment model given by

$$P(t_i^n) - P(t_{i-1}^n) = g\left[X(t_i^n) - P(t_{i-1}^n)\right] + u(t_i^n)\ , \tag{2.5}$$

where $X(t)$ (the intrinsic value of a security at t) and $P(t_i^n)$ (the observed price at t_i^n) are measured in logarithms, the adjustment (constant) coefficient g $(0 < g < 2)$, and $u(t_i^n)$ is an i.i.d. sequence of noise with $\mathcal{E}[u(t_i^n)] = 0$ and $\mathcal{E}[u(t_i^n)^2] = \sigma_u^2$.
The specific linear model of (3.2) was originally proposed by Amihud and Mendelson (1987). We take this model as the starting example because it has been one of well-known models involving transaction costs, interactions among market participants and micro-market structure. We shall depart our discussion from the Amihud-Mendelson model because we are mainly interested in the price adjustment dynamics of a security while their main purpose of study was to investigate the micro-market mechanisms by using daily (open-to-open and close-to-close) data. While Amihud and Mendelson (1987) used that $X(t_i^n)$ follows a (discrete) random walk process in the discrete time series framework, we assume that $X(t)$ is a (scalar) continuous martingale, which is given by (2.1) and $0 < \int_0^t \sigma_s^2 ds < \infty$ (a.s.).

[1] This is only an illustration for the exposition, which may be analogous to the current market practice for the periodic call option of the Tokyo Stock Exchange (TSE).

2.3 The Round-off-error model

Next, we consider the round-off-error model with the micro-market noise. One motivation has been the fact that in financial markets actual transactions occur with the minimum tick size and the observed price data do not have continuous values. The traded quantity also usually has the minimum size in actual financial markets. For instance, the Nikkei-225-futures, which have been the most important traded derivatives in Japan (as explained in Kunitomo and Sato (2008b)), has the minimum 10 yen size while the Nikkei-225-stock index is around 10,000 yen in the year of 2010. (See Hansbrouck (2007) for the details of major stock markets in the U.S., for instance.) Thus it is quite interesting and important to see the effects of round-off-errors on the estimates of the realized volatility when we have realistic round-off errors. We can illustrate the underlying typical argument on the financial markets by showing Figure 2-2 in Appendix B. When the demand curve and supply curve do meet at a point as Figure 2-2, the quantity Q^* is traded at the price P^*. Still there would be excess demand which could not be traded at the particular moment because of the positive tick-size ($\eta > 0$) and the minimum order size effects, i.e. the number of orders should be integers in actual financial markets.

We assume that

$$P(t_i^n) - P(t_{i-1}^n) = g_\eta \left[X(t_i^n) - P(t_{i-1}^n) + u(t_i^n) \right] , \tag{2.6}$$

where $u(t_i^n)$ is an i.i.d. sequence of noise with $\mathcal{E}[u(t_i^n)] = 0, \mathcal{E}[u(t_i^n)^2] = \sigma_u^2$ and the nonlinear function

$$g_\eta(x) = \eta \left[\frac{x}{\eta} \right] , \tag{2.7}$$

where $g_\eta(y)$ is the integer part of y and $[y]$ is the largest integer being less than y and η is a small positive constant.

This model corresponds to the micro-market model with the restriction of the minimum price change and η is the parameter of minimum price change. We set $y_i = P(t_i^n)$ and $x_i = X(t_i^n)$ $(i = 1, \cdots, n)$. We represent (2.7) as

$$P(t_i^n) - X(t_i^n) \tag{2.8}$$

$$= g_\eta \left[-(P(t_{i-1}^n) - X(t_{i-1}^n)) + \Delta X(t_i^n) + u(t_i^n) \right] - [P(t_{i-1}^n) - X(t_{i-1}^n) - \Delta X(t_i^n)]$$
$$= g_\eta^* \left[P(t_{i-1}^n) - X(t_{i-1}^n), \Delta X(t_i^n), u(t_i^n) \right]$$

where

$$\Delta X(t_i^n) = \int_{t_{i-1}^n}^{t_i^n} \sigma_x(s) dB_s \tag{2.9}$$

is a sequence of martingale differences.

2.4 Nonlinear Micro-market price Adjustment models

We generalize the linear price adjustment model and consider nonlinear price adjustments models. As often discussed in the cases of financial crises in the past several decades, there could be different mechanisms among the up-ward phase of financial prices and the down-ward phase of financial prices. In the context of micro-market models in financial economics, some economists have tried to find econometric models involving transaction costs and micro-market structures. In many stock markets usually there are regulations on the maximum limits of down-ward price movements within a day, for instance. One common approach in financial econometrics has been to build statistical models with asymmetrical movements of instantaneous volatility and covariance. The present approach is slightly different from the standard one because we try to consider the micro-market price adjustment processes directly. As an example of the discrete time series modeling of the nonlinear price adjustment model of the security price, we take a non-linear version of (2.5) with

$$g(x) = g_1 x \mathrm{I}(x \geq 0) + g_2 x \mathrm{I}(x < 0) \ , \tag{2.10}$$

where g_i $(i = 1, 2)$ are some constants and $\mathrm{I}(\cdot)$ is the indicator function. This has been called the SSAR (simultaneous switching autoregressive) model, whic have been investigated by Sato and Kunitomo (1996) and Kunitomo and Sato (1999). It is related to one of the threshold autoregressive models in the non-linear time series analysis. A set of sufficient conditions for the geometric ergodicity of the price process is given by

$$g_1 > 0 \ , \ g_2 > 0 \ , \ (1 - g_1)(1 - g_2) < 1 \ . \tag{2.11}$$

This condition has been discussed by Kunitomo and Sato (1999) in the context of nonlinear time series analysis. If we set $g_1 = g_2 = g$, then we have the linear adjustment case and the geometrically ergodicity condition is given by $0 < g < 2$.

More generally, we consider the model

$$P(t_i^n) - P(t_{i-1}^n) = g\left[X(t_i^n) - P(t_{i-1}^n)\right] + u(t_{i-1}^n)\ , \tag{2.12}$$

where $u(t_i^n)$ is an i.i.d. sequence of noise with $\mathcal{E}[u(t_i^n)] = 0$ and $\mathcal{E}[u(t_i^n)^2] = \sigma_u^2$. We set $y_i = P(t_i^n)$ and $x_i = X(t_i^n)$ and define a sequence of martingale differences by

$$\Delta X(t_i^n) = X(t_i^n) - X(t_{i-1}^n) = \int_{t_{i-1}^n}^{t_i^n} \sigma_x(s) dB_s\ . \tag{2.13}$$

From (2.13) and (2.14), let

$$V(t_i^n) = P(t_i^n) - X(t_i^n) - u(t_i^n) \tag{2.14}$$

and $w(t_i^n) = -\Delta X(t_i^n) + u(t_{i-1}^n)$. Then we have

$$\begin{aligned} V(t_i^n) &= V(t_{i-1}^n) + w(t_i^n) + g\left[-V(t_{i-1}^n) - w(t_i^n)\right] \\ &= g^*\left[V(t_{i-1}^n) + w(t_i^n)\right]\ , \end{aligned} \tag{2.15}$$

where $g^*(z) = z + g(-z)$, $\mathcal{E}[w(t_i^n)] = 0$, $\mathcal{E}[w(t_i^n)^2] < \infty$ and $\mathcal{E}[V(t_{i-1}^n)w(t_i^n)] = 0$.

3. The SIML Estimation and its Asymptotic Robustness Properties

3.1 The SIML Method

We summarize the derivation of the separating information maximum likelihood (SIML) estimation proposed by Kunitomo and Sato (2008a,b).. Let $y(t_i^n)$ be the $i-$th observation of the the (log-) price at t_i^n for $j = 1, \cdots, p; 0 = t_0^n \leq t_1^n \leq \cdots \leq t_n^n = 1$ and we set $\mathbf{y}_n = (y_i^n)$ be an $n \times 1$ vector of observations. The underlying (vector-valued) continuous process $X(t)$ $(0 \leq t \leq 1)$, which is not necessarily the

same as the observed (log-)prices at t_i^n $(i = 1, \cdots, n)$ and let $u_i = u(t_i^n)$ be the vector of the micro-market noise at t_i^n $(i = 1, \cdots, n)$.

We first consider the standard situation (2.4) when we have $y(t_i^n) = X(t_i^n) + u_i$, where $X(t)$ $(0 \leq t \leq 1)$ and u_i $(i = 1, \cdots, n)$ are independent with $\sigma_x^2(s) = \sigma_x^2$ (time-invariant), and u_i are independently, identically and normally distributed as $N(0, \sigma_u^2)$. Then given the initial condition y_0, we have

$$\mathbf{y}_n \sim N_n \left(y_0 \mathbf{1}_n, \sigma_v^2 \mathbf{I}_n + h_n \sigma_x^2 \mathbf{C}_n \mathbf{C}_n' \right) , \tag{3.1}$$

where

$$\mathbf{C}_n = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ 1 & \cdots & 1 & 1 & 0 \\ 1 & \cdots & 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \cdots & 0 \\ 0 & \cdots & -1 & 1 & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix}^{-1} , \tag{3.2}$$

$\mathbf{1}_n' = (1, \cdots, 1)$ and $h_n = 1/n$ $(= t_i^n - t_{i-1}^n)$.
By transforming $\mathbf{y}_n$ to $\mathbf{z}_n$ $(= (z_k))$ by

$$\mathbf{z}_n = h_n^{-1/2} \mathbf{P}_n' \mathbf{C}_n^{-1} \left(\mathbf{y}_n - \bar{\mathbf{y}}_0 \right) \tag{3.3}$$

where $\bar{\mathbf{y}}_0 = y_0 \mathbf{1}_n$, $\mathbf{P}_n = (p_{jk})$ and for $j, k = 1. \cdots, n$,

$$p_{jk} = \sqrt{\frac{2}{n + \frac{1}{2}}} \cos \left[\frac{2\pi}{2n+1} (j - \frac{1}{2})(k - \frac{1}{2}) \right] . \tag{3.4}$$

Then the transformed variables v_k $(k = 1, \cdots, n)$ are mutually independent, $v_k \sim N(0, \sigma_x^2 + a_{kn} \sigma_v^2)$ where

$$a_{kn} = 4n \sin^2 \left[\frac{\pi}{2} \left(\frac{2k-1}{2n+1} \right) \right] \ (k = 1, \cdots, n) . \tag{3.5}$$

Because the ML estimator of unknown parameters is a rather complicated function of all observations and each a_{kn} terms depend on k as well as n, one way to have a simple solution of the problem is to approximate the likelihood function. Let m and

l be dependent on n and we write m_n and l_n formally. Then we define the SIML estimator of $\hat{\sigma}_x^2$ by

$$\hat{\sigma}_x^2 = \frac{1}{m_n} \sum_{k=1}^{m_n} z_k^2 \tag{3.6}$$

and the SIML estimator of $\hat{\sigma}_v^2$ by

$$\hat{\sigma}_v^2 = \frac{1}{l_n} \sum_{k=n+1-l_n}^{n} a_{kn}^{-1} z_k^2 \ . \tag{3.7}$$

The numbers of terms m_n and l_n we use are dependent on n such that $m_n, l_n \to \infty$ as $n \to \infty$. We impose the order requirement that $m_n = O(n^\alpha)$ $(0 < \alpha < \frac{1}{2})$ and $l_n = O(n^\beta)$ $(0 < \beta < 1)$ for σ_x^2 and σ_u^2, respectively.

3.2 On Asymptotic Properties of the SIML estimator under the standard additive Model

It is important to investigate the asymptotic properties of the SIML estimator when the instantaneous volatility function $\sigma_x^2(s)$ is not constant over time. When the realized volatility is a positive (deterministic) constant a.s. (i.e. σ_x^2 is not stochastic) while the instantaneous covariance function is time varying, we have the consistency and the asymptotic normality of the SIML estimator as $n \to \infty$. For the deterministic time varying case, the asymptotic properties of the SIML estimator can be summarized as the next proposition and its proof has been given in Kunitomo and Sato (2010).

Theorem 3.1 : We assume that X_i and u_i $(i = 1, \cdots, n)$ in (2.1) and (2.4) are independent, $\sigma_x^2 = \int_0^1 \sigma_x^2(s)ds$ is a positive constant (or deterministic). positive definite matrix, $\mathcal{E}[\|\sqrt{n}(X_i - X_{i-1})\|^4] < \infty$ and $\mathcal{E}[\|u_i\|^4] < \infty$. Define the SIML estimator $\hat{\sigma}_x^2$ of σ_x^2 by (3.6) and (3.7), respectively.

(i) For $m_n = n^\alpha$ and $0 < \alpha < 0.5$, as $n \longrightarrow \infty$

$$\hat{\sigma}_x^2 - \sigma_x^2 \xrightarrow{p} 0 \ . \tag{3.8}$$

(ii) For $m_n = n^\alpha$ and $0 < \alpha < 0.4$, as $n \longrightarrow \infty$

$$\sqrt{m_n}\left[\hat{\sigma}_x^2 - \sigma_x^2\right] \xrightarrow{d} N\left[0, V\right] \ , \tag{3.9}$$

provided that

$$(3.10) V_n = 2\left[\int_0^1 \sigma_x^2(s)ds\right]^2 + 2\sum_{i,j=1}^{n}(m_n c_{ij}^2 - 1)\left[\int_{t_{i-1}}^{t_i}\sigma_x^2(s)ds \int_{t_{j-1}}^{t_j}\sigma_x^2(s)ds\right] \rightarrow V ,$$

which is positive constant and for $i, j = 1, \cdots, n$,

$$(3.11) c_{ij} = \frac{1}{m_n}\sum_{k=1}^{m_n}\{\cos[\frac{2\pi}{2n+1}(i+j-1)(k-\frac{1}{2})] + \cos[\frac{2\pi}{2n+1}(i-j)(k-\frac{1}{2})]\} .$$

There are some remarks on the limiting distribution of the SIML estimator and its asymptotic covariance formula in Theorem 2.1. The quantity $V_n^{(2)}$ defined by

$$(3.12) \qquad V_n^{(2)} = 2\sum_{i,j=1}^{n}(m_n c_{ij}^2 - 1)\left[\int_{t_{i-1}}^{t_i}\sigma_x^2(s)ds \int_{t_{j-1}}^{t_j}\sigma_x^2(s)ds\right]$$

is bounded because $\int_0^1 \sigma_x^2(s)ds$ is bounded.
Then it may be reasonable to assume the convergence of $V_n^{(2)}$ to the second part of V ($V^{(2)}$, say). When the instantaneous covariance $\sigma_x^2(s)$ is constant, then

$$(3.13) \qquad V = 2\left[\int_0^1 \sigma_x^2(s)ds\right]^2 = 2\sigma_x^4 .$$

When σ_x^2 is a random variable, we need the concept of stable convergence. The result of Theorem 3.1 can be held in the proper stochastic case with an additional assumption.

Theorem 3.2 : We assume that $X(t)$ and u_i $(i = 1, \cdots, n)$ in (2.1) and (2.4) are independent and $\sigma_x^2(s) > 0$ (positive). Additionally we assume that each elements of $\sigma_x^2(s)$ $(0 \leq s \leq 1)$ and $\sigma_x^2 = \int_0^1 \sigma_x^2(s)ds$ are *bounded* and $\mathcal{E}[\|u_i\|^4] < \infty$. Define the SIML estimator $\hat{\sigma}_x^2$ of σ_x^2 by (3.6).
(i) For $m_n = n^\alpha$ and $0 < \alpha < 0.5$, as $n \longrightarrow \infty$

$$(3.14) \qquad \hat{\sigma}_x^2 - \sigma_x^2 \xrightarrow{p} 0 .$$

(ii) For $m_n = n^\alpha$ and $0 < \alpha < 0.4$, as $n \longrightarrow \infty$ we have the weak convergence

$$Z_n = \sqrt{m_n}\left[\hat{\sigma}_x^2 - \sigma_x^2\right] \xrightarrow{w} Z^* , \tag{3.15}$$

where the characteristic function $g_n(t) = \mathcal{E}[\exp(itZ_n)]$ converges to the characteristic function of Z^*, which is written as

$$g(t) = \mathcal{E}[e^{-\frac{Vt^2}{2}}] \tag{3.16}$$

and we assume the probability convergence given by

$$V = 2\left[\int_0^1 \sigma_x^2(s)ds\right]^2 + 2\operatorname*{plim}_{n\to\infty}\sum_{i,j=1}^{n}(m_n c_{ij}^2 - 1)\left[\int_{t_{i-1}}^{t_i}\sigma_x^2(s)ds\right]^2 . \tag{3.17}$$

3.3 Robustness of the SIML estimator with micro-market adjustments and the round-off error models

We investigate the asymptotic properties of the SIML estimation under micro-market adjustment models and the round-off error models. First, we investigate the situation in Section 2.2. We have a sequence of discrete observations $P(t_i^n)$ with $0 = t_0^n < t_1^n < \cdots < t_n^n = 1$ and the main purpose is to estimate the realized volatility of the intrinsic value of the underlying security $\sigma_x^2 = \int_0^1 \sigma_x(s)^2 ds$. We re-express (3.2) as

$$\begin{aligned} P(t_i^n) &= (1-g)P(t_{i-1}^n) + gX(t_i^n) + u(t_i^n) \\ &= g\sum_{j=0}^{i-1}(1-g)^j X(t_{i-j}^n) + \sum_{j=0}^{i-1}(1-g)^j u(t_{i-j}^n) \\ &\qquad + \left[g(1-g)^i X(t_0^n) + (1-g)^i u(t_0^n)\right] \end{aligned} \tag{3.18}$$

and

$$\begin{aligned} (1-g)^j X(t_{i-j}^n) &= (1-g)^j\left[X(t_0^n) + \int_0^{t_{i-j}^n}\sigma_s dB_s\right] \\ &= (1-g)^j X(t_i^n) - (1-g)^j\left[\int_{t_{i-j}^n}^{t_i^n}\sigma_s dB_s\right] . \end{aligned} \tag{3.19}$$

Then we have the next result and the proof will be given in Appendix A, which is similar to the one given in Kunitomo and Sato (2010).

Theorem 3.3 : Assume $0 < g < 2$ in (2.5). Define the SIML estimator of the realized volatility of $X(t)$ with $m_n = n^\alpha$ $(0 < \alpha < 0.4)$ by (3.6) with $p = 1$. Then the asymptotic distribution of $\sqrt{m_n}\,[\hat{\sigma}_x^2 - \sigma_x^2]$ is asymptotically $(m_n, n \to \infty)$ equivalent to the limiting distributions given by Theorem 2.1 and Theorem 2.2 under their assumptions.

We note that the present micro-market (linear) adjustment model is quite similar to the structure of the micro-market model with autocorrelated micro-market noise discussed in Kunitomo and Sato (2010).

Second, we investigate the situation of Section 2.3 when we have a sequence of discrete observations under the round-off error models. Define

$$W(t_i^n) = P(t_i^n) - X(t_i^n) - u(t_i^n)\ . \tag{3.20}$$

If $|P(t_{i-1}^n) - X(t_i^n) - u(t_i^n)| > \eta$, then from (3.7) we have $P(t_i^n) = X(t_i^n) + u(t_i^n)$, which means $W(t_i^n) = 0$. On the other hand, if $|P(t_{i-1}^n) - X(t_i^n) - u(t_i^n)| \leq \eta$, then $P(t_i^n) = P(t_{i-1}^n)$ and $|W(t_i^n)| \leq \eta$. By defining $v_i = u(t_i^n) + W(t_i^n)$ $(i = 1, \cdots, n)$, we have the condition

$$|W(t_i^n)| \leq \eta\ . \tag{3.21}$$

By using the similar arguments to the results reported as Theorems 2.1 and 2.2 on the limiting distribution of the realized volatility estimator (Kunitomo and Sato (2010), we have the next result and the proof is given in Appendix A.

Theorem 3.4 : Assume (2.6), (2.7), and $\eta = \eta_n$ depends on n satisfying

$$\eta_n \sqrt{n} = O(1)\ . \tag{3.22}$$

Define the SIML estimator of the realized volatility of $X(t)$ with $m_n = n^\alpha$ $(0 < \alpha < 0.4)$ by (3.6). We write the limiting random variable of the normalized estimator $\sqrt{m_n}\,[\hat{\sigma}_x^2 - \sigma_x^2]$ of σ_x^2 when $n \to \infty$. Then as $\eta_n \to 0$ the limiting distributions of $\hat{\sigma}_\eta^2$

is equivalent to the distributions given by Theorem 2.1 and Theorem 2.2.

We have imposed the condition (3.21) on η, which means that itt is a parameter with small size. This condition could be relaxed because the results of simulations in Section 4 have suggested so. The SIML estimator has the asymptotic robust property against a large number of the round-off-errors models.

Third, we investigate the situation of Section 2.4 when we have a sequence of discrete observations under the non-linear adjustments models. Because the discrete time series $V(t_i^n)$ satisfies the stochastic difference equation (2.15), it is a Markovian process. In order to have the desired result, we need a set of sufficient conditions, which are some type of ergodic conditions. We summarize our results under some additional conditions with the nonlinear price adjustments and the proof will be given in Appendix A.

Theorem 3.5 : For the non-linear time series process $V(t_i^n)$ satisfying (2.14) and (2.15), we assume that there exist functions $\rho_1(\cdot)$ and $\rho_2(\cdot,\cdot)$ such that

$$\text{Cov}[V(t_i^n), V(t_j^n)] = c_1\rho_1(|i-j|)\ , \tag{3.23}$$

where c_1 is a (positive) constant and $\sum_{s=0}^{\infty}\rho_1(s) < \infty$ and

$$\text{Cov}\left[V(t_i^n)V(t_{i'}^n), V(t_j^n)V(t_{j'}^n))\right] = c_2\rho_2(|i-i^{'}|, |j-j^{'}|)\ , \tag{3.24}$$

where c_2 is a (positive) constant and $\sum_{s,s'=0}^{\infty}\rho_2(s,s^{'}) < \infty$.
Define the SIML estimator of the realized volatility of $P(t_i^n)$ with $m_n = n^{\alpha}$ $(0 < \alpha < 0.4)$ by (3.6). Then the asymptotic distribution of $\sqrt{m_n}\,[\hat{\sigma}_x^2 - \sigma_x^2]$ is asymptotically (as $m_n, n \to \infty$) equivalent to the limiting distributions given by Theorem 2.1 and Theorem 2.2.

In the above theorem we impose a set of sufficient conditions as (3.22) and (3.23), which may be relaxed. A simple example is the linear case when $g(x) = c\,x$ (c is a

constant with $0 < c < 2$ and v_i are weakly dependent process. It is straightforward to have (3.22) and (3.23) in this case. The second example is the SSAR(1) model with (3.11). It seems that we need more stringent conditions than (2.11) to have (3.22) and (3.23). There can be a large number of non-linear models for $X(t_i^n)$ and $P(t_i^n)$, and the sufficient conditions for the desired results have been under further investigation.

4. Simulations

We have investigated the robust properties of the SIML estimator for the realized variance based on a set of simulations and the number of replications is 1000. We have taken 20,000, and we have chosen $\alpha = 0.4$ and $\beta = 0.8$. The details of the simulation procedure are similar to the corresponding ones reported by Kunitomo and Sato (2008a, b).

In our simulation we consider several cases when the observations are the sum of signal and micro-market noise. The the instantaneous volatility function is given by

$$\sigma_x^2(s) = \sigma(0)^2 \left[a_0 + a_1 s + a_2 s^2\right], \tag{4.1}$$

where a_i $(i = 0, 1, 2)$ are constants and we have some restrictions such that $\sigma_x(s)^2 > 0$ for $s \in [0, 1]$. It is a typical time varying (but deterministic) case and the realized variance σ_x^2 is given by

$$\sigma_x^2 = \int_0^1 \sigma_x(s)^2 ds = \sigma_x(0)^2 \left[a_0 + \frac{a_1}{2} + \frac{a_2}{3}\right] . \tag{4.2}$$

In this example we have taken several intra-day volatility patterns including the flat (or constant) volatility, the monotone (decreasing or increasing) movements and the U-shaped movements.

Among many Monte-Carlo simulations, we summarize our main results as Tables of Appendix B. We have used several models in the form of (2.3) and each model corresponds to

Model 1 $\qquad h_1(x, y, u) = y + g(x - y) + u$ (g : a constant) ,

Model 2 $\quad h_2(x,y,u) = y + g_\eta(x-y+u)$ $(g_\eta(\cdot)$ is (3.8)) ,

Model 3 $\quad h_3(x,y,u) = y + g_\eta(x-y) + u$ $(g_\eta(\cdot)$ is (3.8)) ,

Model 4 $\quad h_4(x,y,u) = y + u + \begin{cases} g_1(x-y) & \text{if } y \geq 0 \ (g_1 : \text{a constant}) \\ g_2(x-y) & \text{if } y < 0 \ (g_2 : \text{a constant}) \end{cases}$,

Model 5 $\quad h_5(x,y,u) = y + \left[g_1 + g_2 \exp(-\gamma|x-y|^2)\right](x-y)$ $(g_1, g_2 : \text{constants})$,

Model 6 $\quad h_6(x,y,u) = y + g_1 \sin(g_2(x-y))$ $(g_1, g_2 : \text{constants})$,

Model 7 $\quad h_7(x,y,u) = y + h_2 \, o \, h_4 \, o \, h_1(x,y,u)$,

respectively.

Model 1 is the standard model when $g = 1$. When $0 < g < 2$, Model 1 corresponds to the linear model with the micro-market adjustment. Model 2 and Model 3 are the models with the round-off errors. Model 2 is the standard round-off model and Model 3 has a more complicated nonlinearity. Model 4 and Model 5 are the SSAR model and the exponential AR model, which have been known as nonlinear (discrete) time series models. Model 6 is an artificial nonlinear model with a trigonometric function. Model 7 is a combination of three nonlinear models, which corresponds to the most complicated nonlinearity in our examples.

For a comparison we have calculated the historical volatility (HI) estimates and the Realized Kernel (RK) estimates, which were developed by Bandorff-Nielsen et al. (2008). It is because there is a natural question on the comparison of the HI estimator, RK estimator and the SIML estimator, then we can compare three methods in each tables. In order to make a fair comparison we have tried to follow the recommendation by Bandorff-Nielsen et al. (2008) on the choice of kernel (Tukey-Hanning) and the band width parameter H. One important issue in the RK method has been to choose H, which depends on the noise variance and the instantaneous variance and we can interpret as $H = c\sqrt{\sigma_u^2/[\sigma_x^2/n]}$. We have found that the RK estimation gives a reasonable estimate if we had taken the reasonable value of the key parameter H. In most cases the bias and the variance of the RK estimator are larger than the corresponding values of the SIML estimator. Overall the estimates of the SIML method are quite stable and robust against the possible values of the

variance ratio even in the nonlinear situations we have considered.

For Model-1, the estimates obtained by historical-volatility (H-vol) are badly-biased, which have been known in the analysis of high frequency data. Actually, the values of H-vol are badly-biased in all cases of our simulations. Both the SIML method and the RK method give reasonable estimates and the variance of the RK estimator is sometimes smaller than the SIML estimator. (See Tables B1-B4.) For Model-1, however, the RK estimation sometimes gives biased-estimates while the SIML estimation gives reasonable estimates. (See Table B5.) For Model-2 and Model-3, the RK estimation often gives biased-estimates while the SIML estimation gives reasonable estimates. (See Tables B6-B8.) Contrary to our conjecture, for Model-4 and Model-5 both the SIML and the RK estimations often give reasonable results. Finally, for Model-6 and Model-7 the RK estimation sometimes give biased estimates while the SIML estimation gives reasonable estimates.

By examining these results of our simulations we can conclude that we can estimate both the realized volatility of the hidden martingale part. It may be surprising to find that the SIML method gives reasonable estimates even when we have nonlinear transformations of the original unobservable security (intrinsic) values. We have conducted a number of further simulations, but the results are quite similar as we have reported in this section.

5. Conclusions

In this paper, we have shown that the Separating Information Maximum Likelihood (SIML) estimator has the asymptotic robustness in the sense that it is consistent and it has the asymptotic normality under a fairly general conditions even when the standard conditions are not satisfied. They include not only the cases when the micro-market noises are possibly autocorrelated and they are endogenously correlated with the underlying continuous signal process, but also the cases when the micro-market structure has the nonlinear adjustments and the round-off errors under a set of reasonable assumptions. The micro-market factors in actual financial markets are common in the sense that we have the minimum price change and the

minimum order size rules; we often observe the bid-ask differences in stock markets, for instance. Therefore the robustness of the estimation methods of the realized volatility and covariance has been quite important. By conducting large number of simulations, we have confirmed that the SIML estimator has reasonable robust properties in finite samples even in these non-standard situations.

As a concluding remark, we should stress on the fact that the SIML estimator is very simple and it can be practically used not only for the realized volatility but also the realized covariance and the hedging coefficients from the multivariate high frequency financial series. Some applications on the analysis of stock-index futures market have been reported in Kunitomo and Sato (2008b, 2011) as illustrations.

References

[1] Amihud, Y. and H. Mendelason (1987), "Trading Mechanisms and Stock Returns : An Empirical Investigation," *Journal of Finance*, Vol.XLII-3, 533-553.

[2] Barndorff-Nielsen, O., P. Hansen, A. Lunde and N. Shephard (2008), "Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise," *Econometrica*, Vol.76-6, 1481-1536.

[3] Delattre. S. and J. Jacod (1997), "A central limit theorem for normalized functions of the increments of a diffusion process in the presense of round-off errors," *Bernoulli*, Vol 3-1, 1-28.

[4] Engle, R. and Z. Sun (2007), "When is Noise not noise : A microstructure estimate of realized volatility," Working Paper.

[5] Hansbrouck, J. (2007), *Empirical Market Microstructure*, Oxford University Press.

[6] Kunitomo, N. and S. Sato (1999), "Stationary and Non-stationary Simultaneous Switching Autoregressive Models with an Application to Financial Time Series," *Japanese Economic Review*, Blackwell, Vol. 50-2, 161-190.

[7] Kunitomo, N. and S. Sato (2008a), Separating Information Maximum Likelihood Estimation of Realized Volatility and Covariance with Micro-Market Noise, CIRJE Discussion Paper F-581, University of Tokyo, (http://www.e.u-tokyo.ac.jp/cirje/research/).

[8] Kunitomo, N. and S. Sato (2008b), "Realized Volatility, Covariance and Hedging Coefficient of the Nikkei-225 Futures with Micro-Market Noise, CIRJE Discussion Paper F-601, University of Tokyo.

[9] Kunitomo, N. and S. Sato, (2010), "On Properties of Separating Information Maximum Likelihood Estimation of Realized Volatility and Covariance with Micro-Market Noise," CIRJE Discussion Paper F-758, University of Tokyo.

[10] Kunitomo, N. and S. Sato (2011), "The SIML Estimation of the Realized Volatility and Hedging Coefficient of Nikkei-225 Futures with Micro-Market Noise," *Mathematics and Computers in Simulation*, North-Holland, in press.

[11] Malliavin, P. and M. Mancino (2009). "A Fourier Transforem Method for Non-parametric Estimation of Multivariate Volatility," *Annals of Statistics*, 37-4, 1983-2010.

[12] Sato, S. and N. Kunitomo (1996), "Some Properties of the Maximum Likelihood Estimator in Simultaneous Switching Autoregressive Model," *Journal of Time Series Analysis*, 17, 287-307.

APPENDIX A : Mathematical Derivations of Theorems

In Appendix A, we give some details of the proofs of Theorem 3.3, Theorem 3.4 and Theorem 3.5 given in Section 3. Since Theorem 3.5 essentially contains Theorem 3.3, we shall give the proof of Theorem 3.5. (The only difference is the effects of additional terms which are smaller order than $O_p(1)$.)

The proof of Theorem 3.5

(Part-I) We shall investigate the asymptotic properties of the SIML estimator in two steps. The first step is to investigate the conditions that the measurement errors are stochastically negligible.

Define $v_i = V(t_i^n)$ $(i = 1, \cdots, n)$ by (2.14) and we represent $y_i = x_i + v_i$, where $y_i = P(t_i^n), x_i = X(t_i^n)$ and $v_i = V(t_i^n)$. We set $u(t_i^n) = 0$ in (2.14) and $\sigma_x(s) = \sigma_s$ for the resulting simplicity. We write the returns in $(t_{i-1}, t_i]$ as

$$r_i = x_i - x_{i-1} = \int_{t_{i-1}}^{t_i} \sigma_s dB_s \ \ (i = 1, \cdots, n) \tag{A.1}$$

with $0 = t_0 \leq t_1 < \cdots < t_n = 1$ and $t_i - t_{i-1} = 1/n$ $(i = 1, \cdots, n)$. We note that the (instantaneous) volatility function σ_s^2 $(0 \leq s \leq 1)$ and the realized volatility $\Sigma_x = \int_0^1 \sigma_s^2 ds$ can be stochastic.

Let $z_{in}^{(1)}$ and $z_{in}^{(2)}$ $(i = 1, \cdots, n)$ be the i-th elements of

$$\mathbf{z}_n^{(1)} = h_n^{-1/2} \mathbf{P}_n' \mathbf{C}_n^{-1} (\mathbf{x}_n - \bar{\mathbf{y}}_0) \ , \ \mathbf{z}_n^{(2)} = h_n^{-1/2} \mathbf{P}_n' \mathbf{C}_n^{-1} \mathbf{v}_n, \tag{A.2}$$

respectively, where $\mathbf{x}_n = (x_i)$, $\mathbf{v}_n = (v_i)$ and $\mathbf{z}_n = (z_{in})$ are $n \times 1$ vectors with $z_{in} = z_{in}^{(1)} + z_{in}^{(2)}$.

Then by following Kunitomo and Sato (2010), we shall use the arguments developed for investigating the effects of the (possibly) autocorrelated noise term on the asymptotic distribution of $\hat{\sigma}_x^2 - \sigma_x^2$ and $\sigma_x^2 = \int_0^1 \sigma_s^2 ds$. We shall use the decomposition

$$\begin{aligned} \sqrt{m_n}\left[\hat{\sigma}_x^2 - \sigma_x^2\right] &= \sqrt{m_n}\left[\frac{1}{m_n}\sum_{k=1}^{m_n} z_{kn}^2 - \sigma_x^2\right] \\ &= \sqrt{m_n}\left[\frac{1}{m_n}\sum_{k=1}^{m_n} z_{kn}^{(1)2} - \sigma_x^2\right] + \frac{1}{\sqrt{m_n}}\sum_{k=1}^{m_n} \mathcal{E}[z_{kn}^{(2)2}] \\ &\quad + \frac{1}{\sqrt{m_n}}\sum_{k=1}^{m_n}\left[z_{kn}^{(2)2} - \mathcal{E}[z_{kn}^{(2)2}]\right] + 2\frac{1}{\sqrt{m_n}}\sum_{k=1}^{m_n}\left[z_{kn}^{(1)} z_{kn}^{(2)}\right] \ . \end{aligned} \tag{A.3}$$

Then we shall investigate the conditions that three terms except the first one of (A.3) are $o_p(1)$. It is because we could estimate the realized volatility consistently as if there were no noise terms in this situation.

Let $\mathbf{b}_k = \mathbf{e}_k^{'}\mathbf{P}_n^{'}\mathbf{C}_n^{-1} = (b_{kj})$ and $\mathbf{e}_k^{'} = (0, \cdots, 1, 0, \cdots)$ be an $n \times 1$ vector. We write $z_{kn}^{(2)} = \sum_{j=1}^{n} b_{kj} v_{fj}$ and notice that $\sum_{j=1}^{n} b_{kj} b_{k'j} = \delta(k, k^{'}) a_{kn}$. Also we shall use the notation that K_i $(i \geq 1)$ are some positive constants.

First by using the condition (3.22) and the Cauchy-Schwartz inequality, we have

$$
\begin{aligned}
\text{(A.4)} \qquad \mathcal{E}[z_{kn}^{(2)}]^2 &= \mathcal{E}[\sum_{i=1}^{n} b_{ki} v_i \sum_{j=1}^{n} b_{kj} v_j] \\
&\leq \sum_{s=0}^{n} c_1 \rho_1(s) [\sum_{i=1}^{n} b_{ki} b_{k,i-l}] \\
&\leq K_1 \times a_{kn} \ ,
\end{aligned}
$$

provided that $\mathcal{E}(v_i^2)$ are bounded and we use the notation $b_{kj} = 0$ $(j \leq 0)$. By using (3.5) and the relation $\sin x = x - (1/6)x^3 + (1/120)x^5 + o(x^7)$,

$$
\begin{aligned}
\text{(A.5)} \qquad \frac{1}{m_n} \sum_{k=1}^{m_n} a_{kn} &= \frac{1}{m_n} 2n \sum_{k=1}^{m_n} \left[1 - \cos(\pi \frac{2k-1}{2n+1}) \right] \\
&= \frac{n}{m_n} \left[2m_n - \frac{\sin \pi \frac{2m_n}{2n+1}}{\sin \pi \frac{1}{2n+1}} \right] \\
&\sim \frac{n}{m_n} \left[2m_n - \frac{(\pi \frac{2m_n}{2n+1}) - \frac{1}{6}(\pi \frac{2m_n}{2n+1})^3}{(\frac{\pi}{2n+1}) - \frac{1}{6}(\frac{\pi}{2n+1})^3} \right] \\
&= O(\frac{m_n^2}{n})
\end{aligned}
$$

Then the second term of (A.3) becomes

$$
\text{(A.6)} \qquad \frac{1}{\sqrt{m_n}} \sum_{k=1}^{m_n} \mathcal{E}[z_{kn}^{(2)}]^2 \leq K_1 \frac{1}{\sqrt{m_n}} \sum_{k=1}^{m_n} a_{kn} = O(\frac{m_n^{5/2}}{n})
$$

if $0 < \alpha < 0.4$.

For the fourth term of (A.3),

$$
\begin{aligned}
\mathcal{E} \left[\frac{1}{\sqrt{m_n}} \sum_{j=1}^{m_n} z_{kn}^{(1)} z_{kn}^{(2)} \right]^2 &= \frac{1}{m_n} \sum_{k,k'=1}^{m_n} \mathcal{E} \left[z_{kn}^{(1)} z_{k',n}^{(1)} z_{kn}^{(2)} z_{k',n}^{(2)} \right] \\
\text{(A.7)} \qquad &= \frac{1}{m} \sum_{k,k'=1}^{m} \mathcal{E} \left[2 \sum_{j,j'=1}^{n} s_{jk} s_{j'k'} \mathcal{E}(r_j r_{j'} | \mathcal{F}_{\min(j,j')}) z_{kn}^{(2)} z_{k'n}^{(2)} \right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{m_n}\sum_{k,k'=1}^{m_n}\mathcal{E}\left[2\sum_{j=1}^{n}s_{jk}s_{j,k'}\mathcal{E}(r_j^2|\mathcal{F}_{j-1})z_{kn}^{(2)}z_{k',n}^{(2)}\right] \\
&\leq K_2\mathcal{E}\left[(\sup_{0\leq s\leq 1}\sigma_s^2)\frac{2}{n}(\frac{n}{2}+\frac{1}{4})\right]\frac{1}{m_n}\sum_{k,k'=1}^{m_n}\sqrt{a_{kn}}\sqrt{a_{k',n}} \\
&\leq O_p(\sum_{k=1}^{m_n}a_{kn}) \\
&= O_p(\frac{m_n^3}{n}) \ .
\end{aligned}
$$

In the above evaluation we have used the relation

$$
\int_{t_{i-1}}^{t_i}\sigma_s^2 ds \leq \frac{1}{n}\left[\sup_{0\leq s\leq 1}\sigma_s^2\right]
$$

and

$$
|\sum_{j=1}^{n}s_{jk}s_{j,k'}| \leq [\sum_{j=1}^{n}s_{jk}^2] = n/2+1/4 \ \text{ for any } k\geq 1 \ .
$$

Hence we need the condition $0<\alpha<1/3$. When $\sigma_s=\sigma_x$, i.e., the instantaneous volatility function is constant, (A.7) becomes $O(m_n^2/n)$, which is satisfied if $0<\alpha<0.4$.

For the third term of (A.3), we need to consider the variance of

$$
z_{kn}^{(2)2}-\mathcal{E}[z_{kn}^{(2)2}] = \sum_{j,j'=1}^{n}b_{kj}b_{k,j'}\left[v_jv_{,j'}-\mathcal{E}(v_jv_{j'})\right]
$$

and we need to evaluate the expectation of $\left[z_{kn}^{(2)2}-\mathcal{E}[z_{kn}^{(2)2}]\right]\left[z_{k',n}^{(2)2}-\mathcal{E}[z_{k',n}^{(2)2}]\right]$. By using (3.23) and we utilize the fact that

$$
\text{(A.8)} \qquad \sum_{i,i'=1}^{n}\sum_{j,j'=1}^{n}b_{ki}b_{k,i'}b_{k',j}b_{k',j'}\rho_2(|i-i'|,|j-j'|)\sim K_3\times a_{kn}a_{k',n} \ .
$$

Then by collecting each terms, we obtain

$$
\begin{aligned}
\text{(A.9)} \qquad \mathcal{E}\left[\frac{1}{\sqrt{m_n}}\sum_{j=1}^{m_n}(z_{kn}^{(2)2}-\mathcal{E}[z_{kn}^{(2)2}])\right]^2 &\leq \frac{1}{m_n}\sum_{k,k'=1}^{m_n}a_{kn}a_{k'n} \\
&= O(\frac{1}{m_n}\times(\frac{m_n^3}{n})^2) \\
&= O(\frac{m_n^5}{n^2})
\end{aligned}
$$

since $\sum_{k=1}^{m} a_{kn} = O(m_n^3/n)$.

Thus the third term of (A.2) is negligible if $0 < \alpha < 0.4$.

(Part-II) The remaining task is to prove the asymptotic normality of the first term of (A.3), that is,

$$\sqrt{m_n}\left[\frac{1}{m_n}\sum_{k=1}^{m_n} z_{kn}^{(1)2} - \sigma_x^2\right] \tag{A.10}$$

because it is of the order $O_p(1)$. The proof of the asymptotic normality of (A.10) is lengthy, but quite similar to the one given in Kunitomo and Sato (2010) and thus it is omitted here. This completes the proof of Theorem 3.5. **Q.E.D.**

The proof of Theorem 3.4 : The most parts of the proof are very similar to the corresponding ones in the proof of Theorem 3.5. We write $y_i = x_i + v_i, v_i = u_i + w_i$ $(i = 1, \cdots, n)$, where $|w_i| \leq \eta_n$. Then we need to check that the effects of a sequence of random variables w_i $(i = 1, \cdots, n)$ are negligible under the additional assumption (3.21) on the threshold parameter η_n (> 0).

We shall illustrate the underlying arguments. From (A.3) and (A.4), we notice that

$$\begin{aligned} \left[z_{kn}^{(2)}\right]^2 &= \left[\sum_{i=1}^{n} b_{ki}(u_i + w_i)\right]^2 \\ &= \left[\sum_{i=1}^{n} b_{ki}u_i\right]^2 + 2\left[\sum_{i=1}^{n} b_{ki}u_i\right]\left[\sum_{i=1}^{n} b_{ki}w_i\right] + \left[\sum_{i=1}^{n} b_{ki}w_i\right]^2 . \end{aligned} \tag{A.11}$$

By using the Cauchy-Swartz inequality, under (3.21) we have

$$\left[\sum_{i=1}^{n} b_{ki}w_i\right]^2 \leq n\eta_n^2 a_{kn} . \tag{A.12}$$

Then we can find a positive constant such that

$$\mathcal{E}\left[z_{kn}^{(2)}\right]^2 = \left[\sum_{i=1}^{n} b_{ki}(u_i + w_i)\right]^2 \leq K_4 a_{kn}\left[1 + \eta_n\sqrt{n}\right]^2 . \tag{A.13}$$

By using the similar arguments to other terms in the decomposition of (A.3) as (A.11), we can apply the same arguments as the proof of Theorem 3.5. Then we have the desired result in Theorem 3.4. **Q.E.D.**

APPENDIX B : TABLES and FIGURES

In Tables the variances (σ_x^2) are calculated by the SIML estimation method while H-vol and RK are calculated by the historical volatility estimation and the realized kernel estimation methods, respectively. The true-val means the true parameter value in simulations and mean, SD and MSE correspond to the sample mean, the sample standard deviation and the sample mean squared error of each estimator, respectively.

B-1 : Estimation of Realized Volatility (Model-1)
$(a_0 = 1, a_1 = 0, a_2 = 0; \sigma_u^2 = 1.00\text{E} - 04, g = 0.2)$

n=20000	σ_x^2	H-vol	RK
true-val	1.00E+00	1.00E+00	1.00E+00
mean	1.01E+00	2.33E+00	1.04E+00
SD	1.97E-01	2.32E-02	6.58E-02
MSE	3.89E-02	1.78E+00	6.00E-03

B-2 : Estimation of Realized Volatility (Model-1)
$(a_0 = 1, a_1 = 0, a_2 = 0; \sigma_u^2 = 1.00\text{E} + 00, g = 0.2)$

n=20000	σ_x^2	H-vol	RK
true-val	1.00E+00	1.00E+00	1.00E+00
mean	9.96E-01	1.11E-01	9.71E-01
SD	1.93E-01	2.35E-03	6.30E-02
MSE	3.74E-02	7.90E-01	4.80E-03

B-3 : Estimation of Realized Volatility (Model-1)
$(a_0 = 1, a_1 = 0, a_2 = 0; \sigma_u^2 = 1.00\text{E} + 00, g = 1.5)$

n=20000	σ_x^2	H-vol	RK
true-val	1.00E+00	1.00E+00	1.00E+00
mean	1.00E+00	3.00E+00	1.01E+00
SD	1.94E-01	4.03E-02	6.55E-02
MSE	3.78E-02	4.00E+00	4.34E-03

B-4 : Estimation of Realized Volatility (Model-1)

$(a_0 = 1, a_1 = 0, a_2 = 0; \sigma_u^2 = 1.00\text{E}-05, g = 1.0)$

n=20000	σ_x^2	H-vol	RK
true-val	1.00E+00	1.00E+00	1.00E+00
mean	9.88E-01	1.40E+00	9.97E-01
SD	1.99E-01	1.40E-02	6.53E-02
MSE	3.97E-02	1.60E-01	4.27E-03

B-5 : Estimation of Realized Volatility (Model-1)

$(a_0 = 1, a_1 = 0, a_2 = 0; \sigma_u^2 = 1.00\text{E}-06, g = 0.01)$

n=20000	σ_x^2	H-vol	RK
true-val	1.00E+00	1.00E+00	1.00E+00
mean	8.40E-01	2.51E-02	2.48E-01
SD	1.66E-01	5.41E-04	2.76E-02
MSE	5.31E-02	9.50E-01	5.66E-01

B-6 : Estimation of Realized Volatility (Model-2)

$(a_0 = 7, a_1 = -12, a_2 = 6; \sigma_u^2 = 2.00\text{E}-02, \eta = 0.5)$

n=20000	σ_x^2	H-vol	RK
true-val	4.50E+01	4.50E+01	4.50E+01
mean	4.60E+01	1.37E+02	5.36E+01
SD	1.05E+01	6.19E+00	3.65E+00
MSE	1.11E+02	8.46E+03	8.68E+01

B-7 : Estimation of Realized Volatility (Model-3)

$(a_0 = 7, a_1 = -12, a_2 = 6; \sigma_u^2 = 1.00\text{E}-02, \eta = 0.5)$

n=20000	σ_x^2	H-vol	RK
true-val	4.50E+01	4.50E+01	4.50E+01
mean	4.54E+01	3.95E+02	6.19E+01
SD	1.05E+01	6.69E+00	4.07E+00
MSE	1.10E+02	1.22E+05	3.02E+02

B-8 : Estimation of Realized Volatility (Model-3)

$(a_0 = 1, a_1 = 0, a_2 = 0; \sigma_u^2 = 1.00\text{E}+00, \eta = 0.005)$

n=20000	σ_x^2	H-vol	RK
true-val	1.00E+00	1.00E+00	1.00E+00
mean	1.00E+00	6.85E-01	9.97E-01
SD	1.94E-01	8.66E-03	6.21E-02
MSE	3.77E-02	9.92E-02	3.87E-03

B-9 : Estimation of Realized Volatility (Model-4)

$(a_0 = 1, a_1 = 0, a_2 = 0; \sigma_u^2 = 1.00\text{E}+00, g_1 = 0.2, g_2 = 5)$

n=20000	σ_x^2	H-vol	RK
true-val	1.00E+00	1.00E+00	1.00E+00
mean	1.01E+00	2.22E+00	1.01E+00
SD	1.93E-01	6.46E-02	6.25E-02
MSE	3.71E-02	1.49E+00	3.93E-03

B-10 : Estimation of Realized Volatility (Model-4)

$(a_0 = 1, a_1 = 0, a_2 = 0; \sigma_u^2 = 1.00\text{E}-03, g_1 = 0.2, g_2 = 5)$

n=20000	σ_x^2	H-vol	RK
true-val	1.00E+00	1.00E+00	1.00E+00
mean	1.02E+00	6.65E+01	1.11E+00
SD	1.94E-01	1.66E+00	7.46E-02
MSE	3.79E-02	4.30E+03	1.85E-02

B-11 : Estimation of Realized Volatility (Model-5)

$(a_0 = 1, a_1 = 0, a_2 = 0; \sigma_u^2 = 1.00\text{E}+00, g_1 = 1.9, g_2 = -1.7, \gamma = 10000)$

n=20000	σ_x^2	H-vol	RK
true-val	1.00E+00	1.00E+00	1.00E+00
mean	9.99E-01	6.39E+00	1.00E+00
SD	1.92E-01	3.66E-01	6.53E-02
MSE	3.68E-02	2.91E+01	4.26E-03

B-12 : Estimation of Realized Volatility (Model-6)

$(a_0 = 1, a_1 = 0, a_2 = 0; \sigma_u^2 = 1.00\text{E} + 00, \sin(z * 0.1))$

n=20000	σ_x^2	H-vol	RK
true-val	1.00E+00	1.00E+00	1.00E+00
mean	1.00E+00	5.26E-02	8.32E-01
SD	2.14E-01	2.23E-03	6.79E-02
MSE	4.59E-02	8.97E-01	3.27E-02

B-13 : Estimation of Realized Volatility (Model-6)

$(a_0 = 1, a_1 = 0, a_2 = 0; \sigma_u^2 = 1.00\text{E} + 00, 0.01 * \sin(z * 100))$

n=20000	σ_x^2	H-vol	RK
true-val	1.00E+00	1.00E+00	1.00E+00
mean	7.67E-01	4.49E-01	7.75E-01
SD	1.79E-01	3.78E-03	6.05E-02
MSE	8.64E-02	3.03E-01	5.41E-02

B-14 : Estimation of Realized Volatility (Model-7)

$(a_0 = 1, a_1 = 0, a_2 = 0; \sigma_u^2 = 1.00\text{E} - 04, g_1 = 0.2, g_2 = 5; g = 0.01; \eta = 0.01)$

n=20000	σ_x^2	H-vol	RK
true-val	1.00E+00	1.00E+00	1.00E+00
mean	1.18E+00	3.62E+00	1.81E+00
SD	2.30E-01	1.04E-01	1.16E-01
MSE	8.36E-02	6.85E+00	6.69E-01

In Figures 3.1 and 3.2 P and Q stand for the price and the quantity, respectively. D and S are the demand curve and supply curve, respectively. η in Table 3.2 denotes the minimum tick size and Q^* is the quantity traded in Figure 3.2.

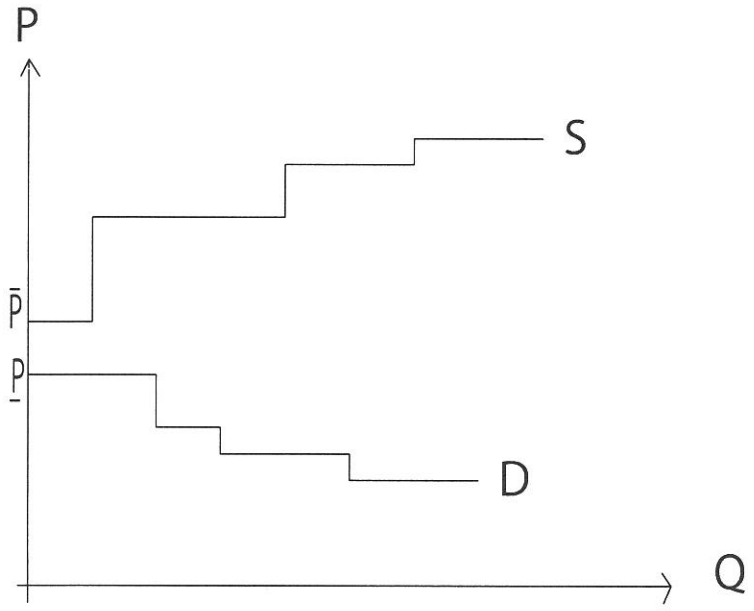


Fig. 2-1

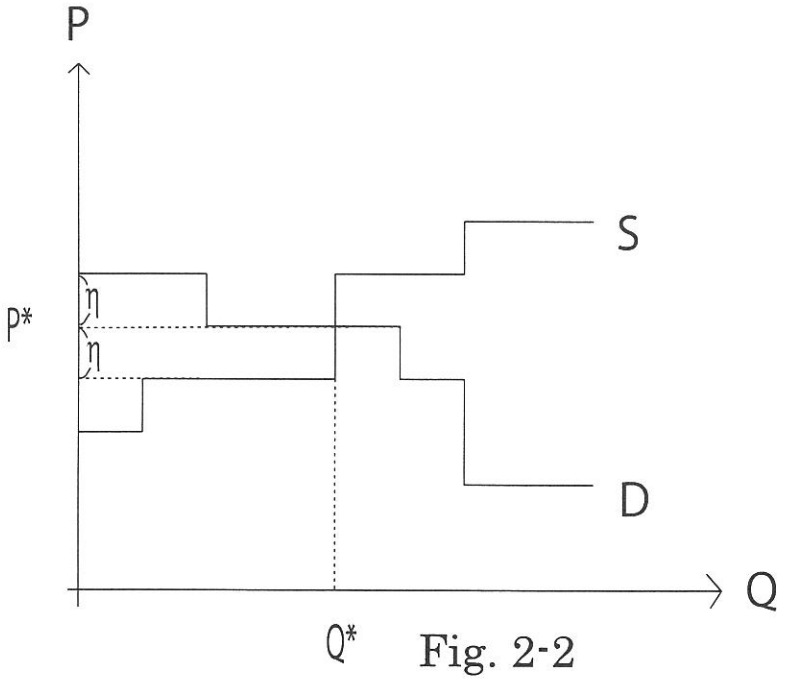


Fig. 2-2

Bayesian Estimation of Probability of Information based trading

Kosuke Oya[1]

Graduate School of Economics, Osaka University, Japan

Key words: PIN, adjusted PIN, PSOS, MCMC

1. Introduction

The analysis of microstructure of financial market has received a lot of attention both from academic researchers and practitioners in recent years. In particular, the key issue for the analysis is the asymmetric information among the traders. Although typical analysis at an earlier stage such as Roll (1984) accommodates the assumption of homogeneous information, that is, there is no information asymmetry, it is become more important to figure out the degree of information trading in the market. Easley and O'Hara (2004) investigate the roles of public and private information in affecting a firm's cost of capital. They argue that stocks with more information asymmetry have higher expected returns through their rational expectations equilibrium model with asymmetric information. This argument is confirmed in Easley, Hvidkjaer and O'Hara (2002) empirically using a structural microstructure model to provide estimates of information-based trading for a large cross section of stocks. The estimation for the probability of informed trading (PIN) in the market on a particular day is proposed by Easley, Kiefer, O'Hara and Paperman (1996). The model is

Email addresses: oya@econ.osaka-u.ac.jp (Kosuke Oya)

[1]This work is supported by the grant-in-aid for scientific research (A) 22243021 and 10153313 from the Japanese Society for the Promotion of Science. The author would like to thank W. Ohta for data processing, T. Watanabe, Y. Omori and T. Nakatsuma for several suggestions about MCMC method. Special thanks to H. Takehara who reminds me of the problem of excess zero counts.

based on the sequential trading model by Glosten and Milgrom (1985). Easley and O'Hara (2004) provide the theoretical framework about PIN based on microeconomic theory and show that the information shift from public to private increases the equilibrium required return. Easley, Hvidkjaer and O'Hara (2002) conduct the empirical study and show that a proxy for information asymmetry is positively and significantly related to average stock returns in U.S. market. On the other hand, Duarte and Young (2009) argue that the PIN model can not capture positive correlation between the numbers of buyer and seller-initiated transactions even though we observe positive correlation for many stocks. They propose the extended PIN model and two measures for information asymmetry and illiquidity to remedy the correlation problem. The adjusted PIN is the PIN component related to asymmetric information and PSOS (probability of symmetric order flow shock) is the PIN component related to illiquidity. They conclude that adjusted PIN is not priced and PSOS is priced. Although both the original PIN and the extended PIN models provide the useful measures for the information asymmetry and illiquidity, these two models can not handle excess zero counts which are often observed in actual market. This excess zero counts make estimation of the model more difficult. In this paper, we propose a zero inflated Poisson mixture model to deal with the excess zero counts problem.

The remainder of the paper is as follows. In next section, we give an overview of the extended PIN model by Duarte and Young (2009). Section 3 provides an introduction of zero inflated Poisson mixture model and Bayesian inference for the model is shown in Section 4. Section 5 presents an empirical illustration that shows the importance of dealing with the excess zero counts. Section 6 concludes.

2. Probability of Information Trading

The original PIN model suffers from the inability to matching the sample moments for the actual market data. Duarte and Young (2009) propose the extended PIN model to remedy the deficit of the original PIN model through

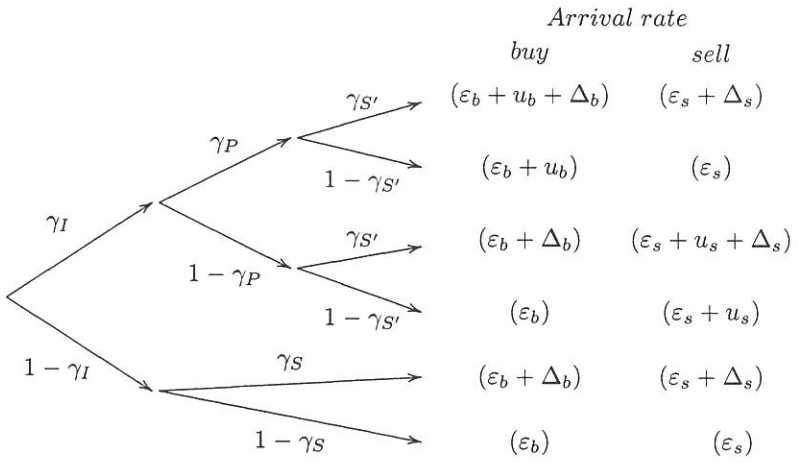


Figure 1: Trading process for Extended PIN model

the incorporation of symmetric order flow shock.

2.1. Extended PIN model

The PIN is the probability that there was informed trading in the market on a particular day. At beginning of the day, a private information event will occur with probability γ_I. The probability that the informed trader receives the positive private signal is γ_P when the private information event occurs on the day. The probabilities of symmetric order flow shock conditional on the absence and the arrival of private information are defined as γ_S and $\gamma_{S'}$, respectively. Suppose that the traders arrive according to Poisson processes where the arrival rates of uninformed buy and sell orders are ε_b and ε_s, and those of informed buy and sell orders are u_b and u_s. The informed buy and sell orders are triggered by the positive and negative private signals, respectively. Further we assume that there are the additional arrival rate Δ_b of buy order and that of Δ_s sell order caused by the symmetric order flow shocks. Figure 1 depicts the trading process. There are six states in the market. For example, the first state exhibits that there is positive information with the symmetric order flow shock. Then the expected order flows for buy and sell are $(\varepsilon_b + u_b + \Delta_b)$ and $(\varepsilon_s + \Delta_s)$,

respectively. In contrast, the case where there is only public information in the market is defined by the last state.

We suppose that the order is executed immediately. So we refer the number of order flows as the number of transactions hereafter if there is no confusion.

Let the numbers of buy and sell transactions are $\boldsymbol{X}_t = (X_{b,t}, X_{s,t})'$. We assume that $\{ \boldsymbol{X}_t \}_{t=1}^{n}$ is the independent random sequence. The joint distribution of the numbers of buy and sell transactions on the day t is given by a Poisson mixture distribution

$$
\begin{aligned}
\Pr(\boldsymbol{X}_t \mid \boldsymbol{\theta}, \boldsymbol{\gamma}) \;=\; & \gamma_I \gamma_P \gamma_{S'}\, P_o(x_{b,t}, \varepsilon_b + u_b + \Delta_b)\, P_o(x_{s,t}, \varepsilon_s + \Delta_s) \qquad (1) \\
& + \gamma_I \gamma_P (1 - \gamma_{S'})\, P_o(x_{b,t}, \varepsilon_b + u_b)\, P_o(x_{s,t}, \varepsilon_s) \\
& + \gamma_I (1 - \gamma_P) \gamma_{S'}\, P_o(x_{b,t}, \varepsilon_b + \Delta_b)\, P_o(x_{s,t}, \varepsilon_s + u_s + \Delta_s) \\
& + \gamma_I (1 - \gamma_P)(1 - \gamma_{S'})\, P_o(x_{b,t}, \varepsilon_b)\, P_o(x_{s,t}, \varepsilon_s + u_s) \\
& + (1 - \gamma_I) \gamma_S\, P_o(x_{b,t}, \varepsilon_b + \Delta_b)\, P_o(x_{s,t}, \varepsilon_s + \Delta_s) \\
& + (1 - \gamma_I)(1 - \gamma_S)\, P_o(x_{b,t}, \varepsilon_b)\, P_o(x_{s,t}, \varepsilon_s)
\end{aligned}
$$

where $\boldsymbol{\theta} = (\varepsilon_b, \varepsilon_s, u_b, u_s, \Delta_b, \Delta_s)'$, $\boldsymbol{\gamma} = (\gamma_I, \gamma_P, \gamma_S, \gamma_{S'})'$ and $P_o(x, c) = e^{-c} c^x / x!$ is a probability mass function of Poisson random variable.

The covariance of $X_{b,t}$ and $X_{s,t}$ implied by the Poisson mixture distribution is not necessarily be always negative unlike the original PIN model. Duarte and Young (2009) apply Maximum likelihood method to estimate unknown $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ and propose *adjusted PIN* as the probability of informed trade that is the ratio of the expected informed order flow to the total expected order flow and $PSOS$ as the unconditional probability that a given trade will come from a shock to both buy and sell order flows as

$$
adjPIN = \frac{IN}{IN + SOS + \varepsilon_b + \varepsilon_s}, \; and \; PSOS = \frac{SOS}{IN + SOS + \varepsilon_b + \varepsilon_s}
$$

where $IN = \gamma_I(\gamma_P u_b + (1 - \gamma_P) u_s)$ and $SOS = (\Delta_b + \Delta_s)(\gamma_I \gamma_{S'} + (1 - \gamma_I)\gamma_S)$.

Although they propose a likelihood ratio test statistic for testing the parameter restriction of the model, a model selection procedure based on the LR

Table 1: Buy & sell order flows

Z_t	buy order flow	sell order flow	$\Pr(Z_t \mid \gamma)$
1	$\varepsilon_b + u_b + \Delta_b$	$\varepsilon_s + \Delta_s$	$\gamma_I \gamma_P \gamma_{S'}$
2	$\varepsilon_b + u_b$	ε_s	$\gamma_I \gamma_P (1 - \gamma_{S'})$
3	$\varepsilon_b + \Delta_b$	$\varepsilon_s + u_s + \Delta_s$	$\gamma_I (1 - \gamma_P) \gamma_{S'}$
4	ε_b	$\varepsilon_s + u_s$	$\gamma_I (1 - \gamma_P)(1 - \gamma_{S'})$
5	$\varepsilon_b + \Delta_b$	$\varepsilon_s + \Delta_s$	$(1 - \gamma_I)\gamma_S$
6	ε_b	ε_s	$(1 - \gamma_I)(1 - \gamma_S)$

statistic is not straightforward since the asymptotic distribution of test statistic is not distributed as χ^2 under some null hypothesis and the pairwise model selection by testing hypotheses often leads ambiguous results.

2.2. State variable

We introduce a state variable Z_t which enables to see whether the private information arrives or not on the specific day, the information is positive or negative and there is the symmetric order-flow shock or not. The state variable Z_t takes value 1 when there is positive private information with symmetric order flow shock, 2 when there is positive private information without symmetric order flow shock, 3 when there is negative private information with symmetric order flow shock, 4 when there is negative private information without symmetric order flow shock, 5 and 6 when there is no private information with and without symmetric order flow shock, respectively. The buy and sell order flows arrive according to Poisson distributions with intensities and their probabilities given in Table 1.

We assume that Z_t's are independent random sequence with probabilities

$\Pr(Z_t = k)$ where $k = 1, \ldots, 6$. For each element of $\boldsymbol{\gamma}$, we have

$$\gamma_I = \sum_{k=1}^{4} \Pr(Z_t = k \mid \boldsymbol{\gamma}), \qquad \gamma_P = \frac{\sum_{k=1}^{2} \Pr(Z_t = k \mid \boldsymbol{\gamma})}{\sum_{k=1}^{4} \Pr(Z_t = k \mid \boldsymbol{\gamma})},$$

$$\gamma_{S'} = \frac{\Pr(Z_t = 1 \mid \boldsymbol{\gamma})}{\sum_{k=1}^{2} \Pr(Z_t = k \mid \boldsymbol{\gamma})}, \qquad \gamma_S = \frac{\Pr(Z_t = 5 \mid \boldsymbol{\gamma})}{\sum_{k=5}^{6} \Pr(Z_t = k \mid \boldsymbol{\gamma})}.$$

2.3. Likelihood conditional on state variable

Before defining the zero inflated PIN model, we show the distribution of $\boldsymbol{X}_t$ of the extended PIN model conditional on the state variable Z_t which is represented as the product of the two Poisson distributions with the intensities $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$

$$\Pr(\boldsymbol{X}_t \mid Z_t = k, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \begin{cases} P_o(x_{b,t}, \varepsilon_b + u_b + \Delta_b)\, P_o(x_{s,t}, \varepsilon_s + \Delta_s) & \text{for } k = 1, \\ P_o(x_{b,t}, \varepsilon_b + u_b)\, P_o(x_{s,t}, \varepsilon_s) & \text{for } k = 2, \\ P_o(x_{b,t}, \varepsilon_b + \Delta_b)\, P_o(x_{s,t}, \varepsilon_s + u_s + \Delta_s) & \text{for } k = 3, \\ P_o(x_{b,t}, \varepsilon_b)\, P_o(x_{s,t}, \varepsilon_s + u_s) & \text{for } k = 4, \\ P_o(x_{b,t}, \varepsilon_b + \Delta_b)\, P_o(x_{s,t}, \varepsilon_s + \Delta_s) & \text{for } k = 5, \\ P_o(x_{b,t}, \varepsilon_b)\, P_o(x_{s,t}, \varepsilon_s) & \text{for } k = 6. \end{cases}$$

The joint probability function of $\boldsymbol{X}_t$ and Z_t for PIN model is given as

$$\Pr(\boldsymbol{X}_t, Z_t \mid \boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{k=1}^{6} \Big\{ \Pr(X_t \mid \boldsymbol{\theta}, Z_t = k) \Pr(Z_t = k \mid \boldsymbol{\gamma}) \Big\}^{1(Z_t = k)}$$

where $1(Z_t = k)$ is the indicator function that takes unity when $Z_t = k$ and zero otherwise. The likelihood function can be readily available from the definition of the joint probability function of $\boldsymbol{X}_t$ and Z_t.

3. Zero inflated Poisson mixture model

We often face excess zero counts for the number of security trade in actual financial market as compared to that through the Poisson distribution. There are several ways to handle such excess zero counts, such as Zero Inflated Poisson (ZIP) model and Hurdle Poisson model. The negative binomial distribution can be adopted instead of Poisson distribution. See Mullahy (1986) and Winkelmann (2008) for details.

3.1. Model

We consider Zero Inflated PIN (ZI-PIN) model which is the extension of the PIN model. Suppose that the excess zero counts occur in a state where there is no private information, that is $Z_t = 6$. Although we have assumed that there exists some public information for the extended PIN model, we now suppose that the excess zero counts arise when there is neither public nor private information and when the severe order imbalance exists caused by public information. To deal with such situation, we define a following joint distribution function of the numbers of buy and sell transactions as

$$\begin{aligned}
\Pr(\boldsymbol{X}_t \mid w, \boldsymbol{\theta}, \boldsymbol{\gamma}) \;=\; & \gamma_I \gamma_P \gamma_{S'}\, P_o(x_{b,t}, \varepsilon_b + u_b + \Delta_b)\, P_o(x_{s,t}, \varepsilon_s + \Delta_s) \qquad (2)\\
& + \gamma_I \gamma_P (1-\gamma_{S'})\, P_o(x_{b,t}, \varepsilon_b + u_b)\, P_o(x_{s,t}, \varepsilon_s)\\
& + \gamma_I (1-\gamma_P) \gamma_{S'}\, P_o(x_{b,t}, \varepsilon_b + \Delta_b)\, P_o(x_{s,t}, \varepsilon_s + u_s + \Delta_s)\\
& + \gamma_I (1-\gamma_P)(1-\gamma_{S'})\, P_o(x_{b,t}, \varepsilon_b)\, P_o(x_{s,t}, \varepsilon_s + u_s)\\
& + (1-\gamma_I)\gamma_S\, P_o(x_{b,t}, \varepsilon_b + \Delta_b)\, P_o(x_{s,t}, \varepsilon_s + \Delta_s)\\
& + (1-\gamma_I)(1-\gamma_S)\, Q(\boldsymbol{x}_t \mid w, \varepsilon_b, \varepsilon_s)
\end{aligned}$$

where

$$Q(\boldsymbol{x}_t \mid w, \varepsilon_b, \varepsilon_s) \;=\; \begin{cases} w + (1-w) P_o(0, \varepsilon_b)\, P_o(0, \varepsilon_s) & x_{b,t} = x_{s,t} = 0 \\ (1-w) P_o(x_{b,t}, \varepsilon_b)\, P_o(x_{s,t}, \varepsilon_s), & \textit{otherwise} \end{cases}$$

where $0 < w < 1$.

3.2. adjPIN and PSOS for ZI-PIN model

The expected order flow in state where there is no information is smaller than that for the extended PIN model. The latter is $(\varepsilon_b + \varepsilon_s)$ and the former is $\{1 - (1-\gamma_I)(1-\gamma_S)w\}(\varepsilon_b + \varepsilon_s)$. Thus the *adjPIN* and the *PSOS* for $ZI\text{-}PIN$ model defined below.

$$\begin{aligned}
adjPIN \;&=\; \frac{IN}{IN + SOS + \{1 - (1-\gamma_I)(1-\gamma_S)w\}(\varepsilon_B + \varepsilon_S)},\\
PSOS \;&=\; \frac{SOS}{IN + SO + \{1 - (1-\gamma_I)(1-\gamma_S)w\}(\varepsilon_B + \varepsilon_S)}.
\end{aligned}$$

4. Bayesian Inference

In this section, we utilize a model selection procedure which is not rely on the asymptotic theory for non-regular case by use of marginal likelihood obtained through Markov chain Monte Carlo method.

4.1. Likelihood Function

The distribution of $\boldsymbol{X}_t$ of the *ZI-PIN* model which is represented as the product of the two Poisson distributions conditional on the state variable Z_t, w and $\boldsymbol{\theta}$.

$$\Pr(\boldsymbol{X}_t \mid Z_t = k, w, \boldsymbol{\theta}) = \begin{cases} P_o(x_{b,t}, \varepsilon_b + u_b + \Delta_b)\, P_o(x_{s,t}, \varepsilon_s + \Delta_s) & \text{for } k = 1, \\ P_o(x_{b,t}, \varepsilon_b + u_b)\, P_o(x_{s,t}, \varepsilon_s) & \text{for } k = 2, \\ P_o(x_{b,t}, \varepsilon_b + \Delta_b)\, P_o(x_{s,t}, \varepsilon_s + u_s + \Delta_s) & \text{for } k = 3, \\ P_o(x_{b,t}, \varepsilon_b)\, P_o(x_{s,t}, \varepsilon_s + u_s) & \text{for } k = 4, \\ P_o(x_{b,t}, \varepsilon_b + \Delta_b)\, P_o(x_{s,t}, \varepsilon_s + \Delta_s) & \text{for } k = 5, \\ Q(\boldsymbol{x}_t \mid w, \varepsilon_b, \varepsilon_s) & \text{for } k = 6. \end{cases}$$

The joint probability function of $\boldsymbol{X}_t$ and Z_t for *ZI-PIN* model is given as

$$\Pr(X_t, Z_t \mid w, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{k=1}^{6} \Big\{ \Pr(X_t \mid Z_t = k, w, \boldsymbol{\theta}) \Pr(Z_t = k \mid \boldsymbol{\gamma}) \Big\}^{1(Z_t = k)}$$

where $1(Z_t = k)$ is the indicator function that takes unity when $Z_t = k$ and zero otherwise.

Denote the number of $Z_t = k$ among n trading days denoted as $n_k = \sum_{t=1}^{n} 1(Z_t = k)$ and the numbers of buy and sell transactions for state k as $m_{bk} = \sum_{t=1}^{n} x_{b,t} 1(z_t = k)$ and $m_{sk} = \sum_{t=1}^{n} x_{s,t} 1(z_t = k)$ where $k = 1, \ldots, 6$. Further we denote $d_t = 1(x_{b,t} = x_{s,t} = 0)$ and $D = \sum_{t=1}^{n} d_t 1(Z_t = 6)$. Then $Q(\boldsymbol{x}_t \mid w, \varepsilon_b, \varepsilon_s)$ can be represented as

$$Q(\boldsymbol{x}_t \mid w, \varepsilon_b, \varepsilon_s) = \Big\{ w + (1-w) e^{-(\varepsilon_b + \varepsilon_s)} \Big\}^{d_t} \left\{ (1-w) \prod_{i=b,s} P_o(x_{i,t}, \varepsilon_i) \right\}^{1-d_t}.$$

It is noted that the parameter w can take negative value with some lower bound. In such case, we obtain a model with zero deflation.

Define $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$ and $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)$. Then the likelihood function of *ZI-PIN* model that is the joint probability function of $(\boldsymbol{X}, \boldsymbol{Z})$ conditional on w, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ is given as follows.

$$\Pr(\boldsymbol{X}, \boldsymbol{Z} \mid w, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{t=1}^{n} \prod_{k=1}^{6} \Big\{ \Pr(\boldsymbol{X}_t \mid Z_t = k, w, \boldsymbol{\theta}) \Pr(Z_t = k \mid \boldsymbol{\gamma}) \Big\}^{1(Z_t = k)}. \quad (3)$$

4.2. Prior distribution for Poisson intensity $\boldsymbol{\theta}$

We consider the prior distributions for the vector of the intensities $\boldsymbol{\theta}$ of Poisson distributions whose element should be positive. Thus we adopt Gamma distributions $\mathcal{G}(\alpha_k, \beta_k)$, $\mathcal{G}(\dot{\alpha}_k, \dot{\beta}_k)$ and $\mathcal{G}(\ddot{\alpha}_k, \ddot{\beta}_k)$ for the prior distributions of ε_k, u_k and Δ_k for $k = b, s$

$$\pi(\boldsymbol{\theta}) \quad \propto \quad \prod_{k=b,s} \varepsilon_k^{\alpha_k - 1} u_k^{\dot{\alpha}_k - 1} \Delta_k^{\ddot{\alpha}_k - 1} e^{-\beta_k \varepsilon_k} e^{-\dot{\beta}_k u_k} e^{-\ddot{\beta}_k \Delta_k} \quad (4)$$

4.3. Prior distribution for w *and* $\boldsymbol{\gamma}$

We suppose that the prior distribution of w and $\boldsymbol{\gamma} = (\gamma_I, \gamma_P, \gamma_S, \gamma_{S'})'$ are Beta distributions $\mathcal{BE}(\delta_i, \tau_i)$ for $i = w, I, P, S, S'$

$$\pi(w, \boldsymbol{\gamma}) \propto w^{\delta_w - 1}(1 - w)^{\tau_w - 1} \prod_{i=I,P,S,S'} \gamma_i^{\delta_i - 1}(1 - \gamma_i)^{\tau_i - 1}. \quad (5)$$

4.4. Posterior distribution of w, $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$ *and* $\boldsymbol{Z}$

Posterior distribution of w, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ conditional on $\boldsymbol{X}$ and $\boldsymbol{Z}$ is given using the likelihood function (3) and the prior distributions of w, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ as

$$\begin{aligned} \pi(w, \boldsymbol{\theta}, \boldsymbol{\gamma} \mid \boldsymbol{X}, \boldsymbol{Z}) &= \frac{\Pr(\boldsymbol{X}, \boldsymbol{Z} \mid w, \boldsymbol{\theta}, \boldsymbol{\gamma})\, \pi(\boldsymbol{\theta})\, \pi(w, \boldsymbol{\gamma})}{\int\int\int \Pr(\boldsymbol{X}, \boldsymbol{Z} \mid w, \boldsymbol{\theta}, \boldsymbol{\gamma})\, \pi(\boldsymbol{\theta})\, \pi(w, \boldsymbol{\gamma})\, \mathrm{d}w\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{\gamma}} \\ &\propto \Pr(\boldsymbol{X}, \boldsymbol{Z} \mid w, \boldsymbol{\theta}, \boldsymbol{\gamma})\, \pi(\boldsymbol{\theta})\, \pi(w, \boldsymbol{\gamma}). \end{aligned} \quad (6)$$

4.5. Conditional posterior distribution of $\boldsymbol{\gamma}$

Though the joint posterior distribution is complicated as see in (6), the posterior distribution of the element of $\boldsymbol{\gamma}$ conditional on w, $\boldsymbol{\theta}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ is given

as following Gamma distributions

$$\gamma_I \mid \boldsymbol{\gamma}_{-I}, w, \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z} \sim \mathcal{BE}\Big(n_1 + n_2 + n_3 + n_4 + \delta_I, n_5 + n_6 + \tau_I \Big), \quad (7)$$

$$\gamma_P \mid \boldsymbol{\gamma}_{-P}, w, \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z} \sim \mathcal{BE}\Big(n_1 + n_2 + \delta_P, n_3 + n_4 + \tau_P \Big), \quad (8)$$

$$\gamma_{S'} \mid \boldsymbol{\gamma}_{-S'}, w, \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z} \sim \mathcal{BE}\Big(n_1 + n_3 + \delta_{S'}, n_2 + n_4 + \tau_{S'} \Big), \quad (9)$$

$$\gamma_S \mid \boldsymbol{\gamma}_{-S}, w, \boldsymbol{X}, \boldsymbol{Z} \sim \mathcal{BE}\Big(n_5 + \delta_S, n_6 + \tau_S \Big). \quad (10)$$

The sampling from the conditional distributions (7), (8), (9) and (10) is straightforward when the state variable $\boldsymbol{Z}$ is given.

4.6. Sampling for w and $\boldsymbol{\theta}$ with M-H algorithm

Since the conditional posterior distribution of w and $\boldsymbol{\theta}$ on $\boldsymbol{\gamma}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ is still complicated to generate random sample, we adopt the Metropolis-Hastings (M-H) algorithm as follows. Let $g(w, \boldsymbol{\theta}) = \log\{\, \pi(w, \boldsymbol{\theta} \mid \boldsymbol{\gamma}, \boldsymbol{X}, \boldsymbol{Z}) \,\}$ that is

$$\begin{aligned}
g(w, \boldsymbol{\theta}) \;=\; & m_{b1} \log(\varepsilon_b + u_b + \Delta_b) - n_1(\varepsilon_b + u_b + \Delta_b) \\
& + m_{s3} \log(\varepsilon_s + u_s + \Delta_s) - n_3(\varepsilon_s + u_s + \Delta_s) \\
& + (m_{b3} + m_{b5}) \log(\varepsilon_b + \Delta_b) - (n_3 + n_5)(\varepsilon_b + \Delta_b) \\
& + (m_{s1} + m_{s5}) \log(\varepsilon_s + \Delta_s) - (n_1 + n_5)(\varepsilon_s + \Delta_s) \\
& + m_{b2} \log(\varepsilon_b + u_b) - n_2(\varepsilon_b + u_b) \\
& + m_{s4} \log(\varepsilon_s + u_s) - n_4(\varepsilon_s + u_s) \\
& + (m_{b4} + m_{b6} + \alpha_b - 1) \log \varepsilon_b - (n_4 + n_6 - D + \beta_b)\varepsilon_b \\
& + (m_{s2} + m_{s6} + \alpha_s - 1) \log \varepsilon_s - (n_2 + n_6 - D + \beta_s)\varepsilon_s \\
& + (\dot{\alpha}_b - 1) \log u_b - \dot{\beta}_b u_b + (\dot{\alpha}_s - 1) \log u_s - \dot{\beta}_s u_s \\
& + (\ddot{\alpha}_b - 1) \log \Delta_b - \ddot{\beta}_b \Delta_b + (\ddot{\alpha}_s - 1) \log \Delta_s - \ddot{\beta}_s \Delta_s \\
& + D \log\Big\{ w + (1 - w) e^{-(\varepsilon_b + \varepsilon_s)} \Big\} + (n_6 - D) \log(1 - w) \\
& + (\delta_w - 1) \log w + (\tau_w - 1) \log(1 - w). \qquad (11)
\end{aligned}$$

For notational simplicity, we denote a parameter vector $\boldsymbol{\vartheta} = (w, \boldsymbol{\theta}')'$ and define the approximation of $g(\boldsymbol{\vartheta})$ using the Taylor expansion around the mode $\hat{\boldsymbol{\vartheta}}$ as $h(\boldsymbol{\vartheta})$

$$g(\boldsymbol{\vartheta}) \approx h(\boldsymbol{\vartheta}) \equiv g(\hat{\boldsymbol{\vartheta}}) + g'_{\hat{\vartheta}}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\vartheta}}) + \frac{1}{2}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\vartheta}})' g_{\hat{\vartheta}\hat{\vartheta}'}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\vartheta}})$$

where

$$g_{\hat{\vartheta}} = \frac{\partial}{\partial \boldsymbol{\vartheta}} g(\boldsymbol{\vartheta})\Big|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}} \quad \text{and} \quad g_{\hat{\vartheta}\hat{\vartheta}'} = \frac{\partial^2}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}'} g(\boldsymbol{\vartheta})\Big|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}}.$$

Since

$$h(\boldsymbol{\vartheta}) = const. - \frac{1}{2}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\vartheta}})' \Theta^{-1} (\boldsymbol{\vartheta} - \hat{\boldsymbol{\vartheta}}), \quad \Theta^{-1} = -g_{\hat{\vartheta}\hat{\vartheta}'},$$

then we provide the proposal distribution for the conditional posterior distribution of $\boldsymbol{\vartheta}$ as $\mathrm{N}(\hat{\boldsymbol{\vartheta}}, \Theta)$. It is noted that the first and rest components of $\boldsymbol{\vartheta}$ should be in $(0, 1)$ and be non-negative value, respectively. Then we define the proposal density for the candidate $\boldsymbol{\vartheta}^*$ as

$$q(\boldsymbol{\vartheta}^{(i-1)}, \boldsymbol{\vartheta}^*) \propto \exp\left\{ -\frac{1}{2}(\boldsymbol{\vartheta}^* - \hat{\boldsymbol{\vartheta}})' \Theta^{-1} (\boldsymbol{\vartheta}^* - \hat{\boldsymbol{\vartheta}}) \right\} 1(\boldsymbol{\theta}^* \geq 0) \cdot 1(0 < w^* < 1).$$

The acceptance probability of the candidate $\boldsymbol{\vartheta}^*$ is

$$\eta(\boldsymbol{\vartheta}^{(i-1)}, \boldsymbol{\vartheta}^*) = \min\left\{ 1, \frac{\pi(\boldsymbol{\vartheta}^* \mid \boldsymbol{\gamma}, \boldsymbol{X}, \boldsymbol{Z})\, q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}^{(i-1)})}{\pi(\boldsymbol{\vartheta}^{(i-1)} \mid \boldsymbol{\gamma}, \boldsymbol{X}, \boldsymbol{Z})\, q(\boldsymbol{\vartheta}^{(i-1)}, \boldsymbol{\vartheta}^*)} \right\}.$$

M-H algorithm

1. Set the initial value $\boldsymbol{\vartheta}^{(0)}$ and other conditional variables.
2. For $i = 1, \ldots, N$, we conduct following step.
   (a) generate $\boldsymbol{\vartheta}^*$ through the proposal density defined above and calculate the acceptance probability of $\boldsymbol{\vartheta}^*$.
   (b) Generate an uniform random number ν and set $\boldsymbol{\vartheta}^{(i)}$ as follow

$$\boldsymbol{\vartheta}^{(i)} = \begin{cases} \boldsymbol{\vartheta}^*, & \nu \leq \eta(\boldsymbol{\vartheta}^{(i-1)}, \boldsymbol{\vartheta}^*) \\ \boldsymbol{\vartheta}^{(i-1)}, & \nu > \eta(\boldsymbol{\vartheta}^{(i-1)}, \boldsymbol{\vartheta}^*). \end{cases}$$

3. Go to the other sampling block.

4.7. Conditional probability of Z_t

The conditional probability of the state variable $\boldsymbol{Z}_t$ is given as $\Pr(Z_t^{(i)} = k \mid \boldsymbol{X}_t, w^{(i)}, \boldsymbol{\theta}^{(i)}, \boldsymbol{\gamma}^{(i-1)})$

$$\Pr(Z_t = k \mid \boldsymbol{X}_t, w, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \frac{\Pr(\boldsymbol{X}_t \mid Z_t = k, w, \boldsymbol{\theta}, \boldsymbol{\gamma}) \Pr(Z_t = k \mid \boldsymbol{\gamma})}{\sum_{\ell=1}^{6} \Pr(\boldsymbol{X}_t \mid Z_t = \ell, w, \boldsymbol{\theta}, \boldsymbol{\gamma}) \Pr(Z_t = \ell \mid \boldsymbol{\gamma})}. \quad (12)$$

4.8. Sampling

The parameters and the state variables being sampled are w, $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$ and $Z = \{Z_t\}_{t=1}^n$. The hyper-parameters that should be determined prior to the random number generation are α_k, $\dot{\alpha}_k$, $\ddot{\alpha}_k$, β_k, $\dot{\beta}_k$, $\ddot{\beta}_k$ for $k = b, s$, and (δ_ℓ, τ_ℓ) for $\ell = w, I, P, S'$.

(I) Set the initial values $w^{(0)}$, $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$.

- $\alpha_k = \tilde{\varepsilon}_k \times \beta_k$ and $\beta_k = \tilde{\varepsilon}_k / varx$ where $\tilde{\varepsilon}_k$ is a maximum likelihood estimate of ε_k and $varx = \tilde{\varepsilon}_k$ for $k = b, s$. Similarly, we set $(\dot{\alpha}_k, \dot{\beta}_k)$ and $(\ddot{\alpha}_k, \ddot{\beta}_k)$.
- $(\delta_\ell, \tau_\ell) = (1, 1)$ for $\ell = w, I, P, S'$.
- For $(\varepsilon_b^{(0)}, \varepsilon_s^{(0)})$, $(u_b^{(0)}, u_s^{(0)})$, $(\Delta_b^{(0)}, \Delta_s^{(0)})$, $w^{(0)}$ and $\gamma_k^{(0)}$ for $k = I, P, S'$ are set to be the maximum likelihood estimates of them.

(II) Generate $\boldsymbol{Z}^{(0)} = \{Z_t^{(0)}\}_{t=1}^n$ through the multinomial distribution with the conditional probability $\Pr(Z_t^{(0)} = k \mid \boldsymbol{X}_t, w^{(0)}, \boldsymbol{\theta}^{(0)}, \boldsymbol{\gamma}^{(0)})$ given in (12) and set $D^{(0)}$, $n_j^{(0)}$, $m_{bj}^{(0)}$ and $m_{sj}^{(0)}$ using $Z^{(0)}$ for $j = 1, \ldots, 6$.

(III) Generation of $w^{(i)}$, $\boldsymbol{\theta}^{(i)}$ and $\boldsymbol{Z}^{(i)}$ for $i = 1, 2, \ldots, N + M$

(a) Generate $w^{(i)}$ and $\boldsymbol{\theta}^{(i)}$ using M-H algorithm which is described above.

(b) Generate $\boldsymbol{Z}^{(i)} = \{Z_t^{(i)}\}_{t=1}^n$ through the multinomial distribution with the conditional probability $\Pr(Z_t^{(i)} = k \mid \boldsymbol{X}_t, w^{(i)}, \boldsymbol{\theta}^{(i)}, \boldsymbol{\gamma}^{(i-1)})$ given in (12)

(c) Update $D^{(i)}$, $n_k^{(i)}$, $m_{bk}^{(i)}$ and $m_{sk}^{(i)}$ using $\boldsymbol{Z}^{(i)}$ for $k = 1, \ldots, 6$.

(d) Update $\boldsymbol{\gamma}^{(i)}$.

$$\begin{aligned}
\gamma_I^{(i)} \mid \boldsymbol{X}, \boldsymbol{Z}^{(i)} &\sim \mathcal{BE}\Big(\sum_{k=1}^{4} n_k^{(i)} + \delta_I - 1, \sum_{k=5}^{6} n_k^{(i)} + \tau_I - 1 \Big) \\
\gamma_P^{(i)} \mid \boldsymbol{X}, \boldsymbol{Z}^{(i)} &\sim \mathcal{BE}\Big(\sum_{k=1}^{2} n_k^{(i)} + \delta_P - 1, \sum_{k=3}^{4} n_k^{(i)} + \tau_P - 1 \Big) \\
\gamma_{S'}^{(i)} \mid \boldsymbol{X}, \boldsymbol{Z}^{(i)} &\sim \mathcal{BE}\Big(n_1^{(i)} + n_3^{(i)} + \delta_{S'} - 1, n_2^{(i)} + n_4^{(i)} + \tau_{S'} - 1 \Big) \\
\gamma_S^{(i)} \mid \boldsymbol{X}, \boldsymbol{Z}^{(i)} &\sim \mathcal{BE}\Big(n_5^{(i)} + \delta_S - 1, n_6^{(i)} + \tau_S - 1 \Big).
\end{aligned}$$

(e) Set $i = i + 1$ and go to (a), until $i > N + M$ where N is the length of burn in period and M is the number of samples to store.

(IV) Calculate $adjPIN$ and $PSOS$ for $i = N + 1, \ldots, N + M$,

$$\begin{aligned}
adjPIN^{(i)} &= \frac{IN^{(i)}}{IN^{(i)} + SO^{(i)} + \{1 - (1 - \gamma_I^{(i)})(1 - \gamma_S^{(i)})w^{(i)}\}(\varepsilon_B^{(i)} + \varepsilon_S^{(i)})} \\
PSOS^{(i)} &= \frac{SO^{(i)}}{IN^{(i)} + SO^{(i)} + \{1 - (1 - \gamma_I^{(i)})(1 - \gamma_S^{(i)})w^{(i)}\}(\varepsilon_B^{(i)} + \varepsilon_S^{(i)})}
\end{aligned}$$

where

$$\begin{aligned}
IN^{(i)} &= \gamma_I^{(i)} \left\{ \gamma_P^{(i)} u_B^{(i)} + (1 - \gamma_P^{(i)}) u_S \right\}, \\
SO^{(i)} &= (\Delta_B^{(i)} + \Delta_S^{(i)}) \left\{ \gamma_I^{(i)} \gamma_{S'}^{(i)} + (1 - \gamma_I^{(i)}) \gamma_S^{(i)} \right\}.
\end{aligned}$$

4.9. Model comparison

The model we have considered includes more parameters than the original PIN model. The ZI-PIN model implies the several model specification with a variety of parameter restrictions. We consider the following six models. $\mathcal{M}_0$ is the original and $\mathcal{M}_5$ is a full specification of ZI-PIN model.

$$\begin{aligned}
\mathcal{M}_0 &: \ \gamma_{S'} = \gamma_S = 0, \ u_b = u_s, \\
\mathcal{M}_1 &: \ \gamma_{S'} = 0, \ u_b = u_s, \ \Delta_b = \Delta_s, \\
\mathcal{M}_2 &: \ \gamma_{S'} = 0, \ u_b = u_s, \\
\mathcal{M}_3 &: \ \gamma_{S'} = 0, \\
\mathcal{M}_4 &: \ \gamma_S = \gamma_{S'}, \\
\mathcal{M}_5 &: \ \textit{Unrestricted model.}
\end{aligned}$$

The model selection is made on the marginal likelihood using the modified harmonic mean estimator proposed by Geweke (1999).

5. Empirical illustration

In this section, we provide some empirical illustration to make the point considered in this paper clear. We use the intra day data of several stocks listed on the first section of Tokyo Stock Exchange (TSE). The numbers of buyer and seller initiated trades for the five minutes after market open from July 1, 2009 to December 30, 2009 which consists of 123 days are used to estimate the model proposed in this paper.

Figure 2 is indicates the plot and histograms of the number of trade at bid and ask for the security code 1332 (Nippon Suisan: fishery, agriculture & forestry). We observe slight more large numbers at zero for both number of trade at bid and ask as compare to a Poisson distribution. The percentages of the former and the latter are 6.50% and 7.32%, respectively. The correlation coefficient between the numbers of trade at bid and ask is 0.262.

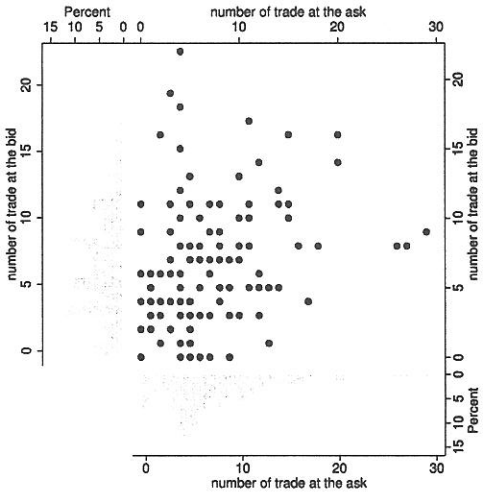


Figure 2: Number of trade at bid and ask for code 1332

To see the effect of the excess zero counts, we show the maximum likelihood estimates for ZI-PIN ($\mathcal{M}_5$) and PIN ($\mathcal{M}_5$) models in Table 2. The estimates of the parameter w for excess zero is small, but it is significance at 1% level. So

Table 2: Maximum likelihood estimates for ZI-PIN and PIN models

	ZI-PIN			PIN		
	Estimate	s.e.	p-value	Estimate	s.e.	p-value
γ_I	0.353	0.116	0.001	0.321	0.078	0.000
γ_P	0.483	0.218	0.014	0.743	0.160	0.000
$\gamma_{S'}$	0.229	0.142	0.055	0.552	0.153	0.000
γ_S	0.317	0.108	0.002	0.505	0.093	0.000
ε_b	3.851	0.447	0.000	2.913	0.416	0.000
ε_s	4.138	0.538	0.000	3.250	0.433	0.000
u_b	7.338	2.426	0.002	8.641	1.173	0.000
u_s	7.311	1.613	0.000	8.317	3.124	0.004
Δ_b	7.256	2.250	0.001	4.058	0.891	0.000
Δ_s	5.135	0.915	0.000	5.587	0.602	0.000
w	0.055	0.032	0.044			
$adjPIN$	0.186	0.036	0.000	0.197	0.034	0.000
$PSOS$	0.254	0.057	0.000	0.360	0.052	0.000

the estimated $adjPIN$ and $PSOS$ for PIN model take similar values as those for ZI-PIN model.

For MCMC estimation, we draw 10,000 sample after the initial 10,000 sample are discarded through the sampling algorithm described in the previous section. Table 3 reports the model selection result. The maximum log of marginal likelihood is attained for $\mathcal{M}_5$ that is a model without parameter restriction. The summary statistics of the posterior distributions for $(\boldsymbol{\theta}, w, \boldsymbol{\gamma})$ by ZI-PIN $(\mathcal{M}_5)$ model is given in Table 4. Figures 3 and 4 are the sample auto correlation function, the posterior density for the parameters of ZI-PIN $(\mathcal{M}_5)$ model, respectively. We have introduced the state variable Z_t which indicates that whether the information arrives or not, the information is positive and there is the symmetric order flow shock.

The probability $\Pr(Z_t = k \mid \boldsymbol{X}_t, w, \boldsymbol{\theta}, \boldsymbol{\gamma})$ is estimated as the sample mean of $\{Z_t^{(i)} = k\}_{t=1}^n$, $i = 1, \ldots, 10,000$ which is the sample generated by MCMC. We denote this estimated probability as $\widehat{\Pr}(Z_t = k)$. The probability $\boldsymbol{\gamma} = (\gamma_I, \gamma_P, \gamma_{S'}, \gamma_S)'$ for each t can be obtained from $\widehat{\Pr}(Z_t = k)$. We classify a day t as the day with information if the estimated γ_I for day t is greater than 0.5. The other categories are classified according to the similar way. The classified

days are shown in Figure 5. We confirm that there is more buy transactions when there is positive information and vice versa. Figure 6 shows the changes of category depending the day of the week.

Table 3: Log likelihood and Log of marginal likelihood for code 1332

	Log likelihood	M-H	Log of marginal L.	s.e.
$\mathcal{M}_0$	-748.258	0.851	-117.871	0.012
$\mathcal{M}_1$	-723.528	0.849	-117.517	0.014
$\mathcal{M}_2$	-717.267	0.841	-118.278	0.016
$\mathcal{M}_3$	-716.341	0.832	-118.125	0.017
$\mathcal{M}_4$	-718.298	0.825	-117.853	0.060
$\mathcal{M}_5$	-712.837	0.834	-117.130	0.016

M-H is the acceptance ratio of M-H algorithm.
s.e. is the standard error of log of marginal likelihood.

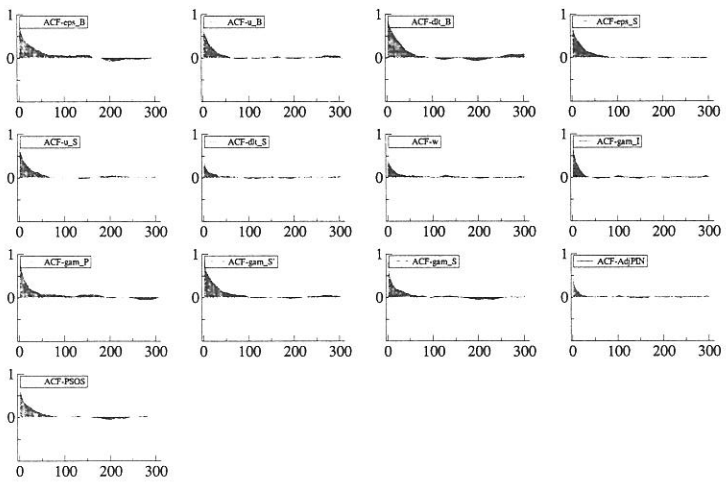


Figure 3: Sample auto correlation function for $\mathcal{M}_5$

Next example illustrates the severe excess zero counts case. The security code is 5201 (Asahi Glass Company Limited: Glass & ceramics products). The sample correlation between the numbers of trade at bid and ask is 0.752. The percentages of zero counts at bid and ask are 27.64%.

Table 4: Summary statistic of posterior distribution of $\mathcal{M}_5$ for code 1332

	mean	s.d.	95%L	median	95%U	Geweke	Inef
γ_I	0.463	0.112	0.261	0.459	0.696	0.036	18.028
γ_P	0.512	0.166	0.199	0.512	0.826	0.069	32.029
$\gamma_{S'}$	0.329	0.151	0.094	0.309	0.661	0.125	36.287
γ_S	0.345	0.123	0.108	0.344	0.593	0.152	24.202
ε_b	3.579	0.522	2.451	3.609	4.483	0.293	37.491
ε_s	3.836	0.536	2.794	3.835	4.865	0.117	30.110
u_b	6.559	1.523	3.782	6.565	9.541	0.280	24.336
u_s	5.969	1.492	2.811	6.041	8.822	0.007	27.838
Δ_b	6.658	1.856	3.356	6.770	10.054	0.151	42.604
Δ_s	5.340	0.834	3.590	5.389	6.852	0.486	12.708
w	0.082	0.046	0.024	0.073	0.186	0.446	15.366
$adjPIN$	0.199	0.033	0.134	0.199	0.264	0.031	9.864
$PSOS$	0.282	0.058	0.175	0.281	0.402	0.953	27.145

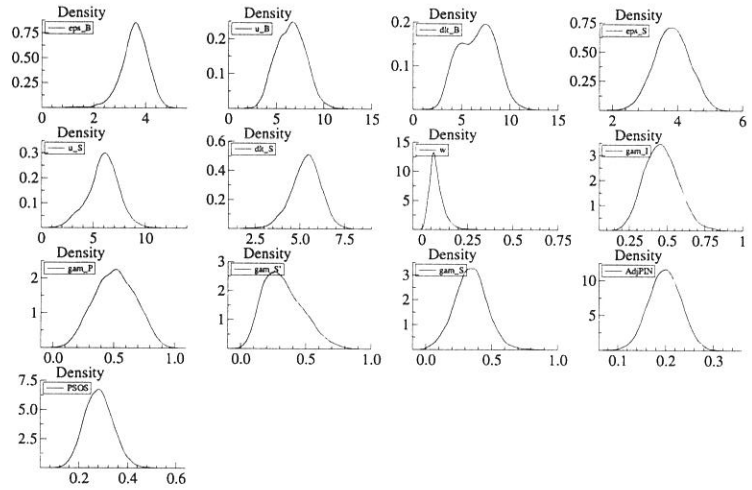


Figure 4: Posterior density for $\mathcal{M}_5$

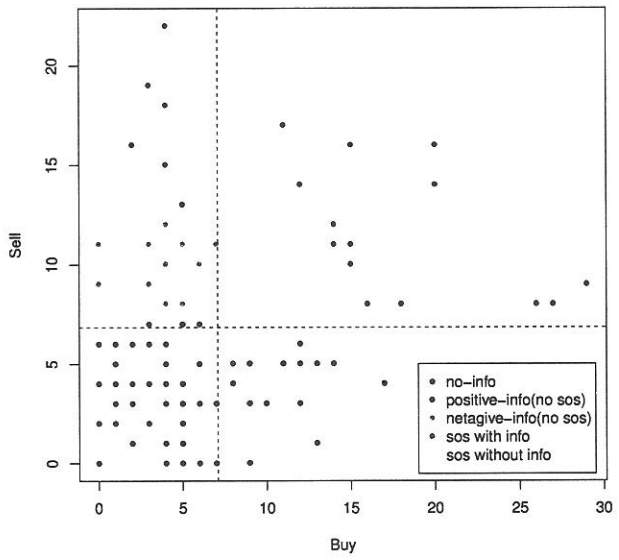


Figure 5: Transaction classify by estimated state for code 1332

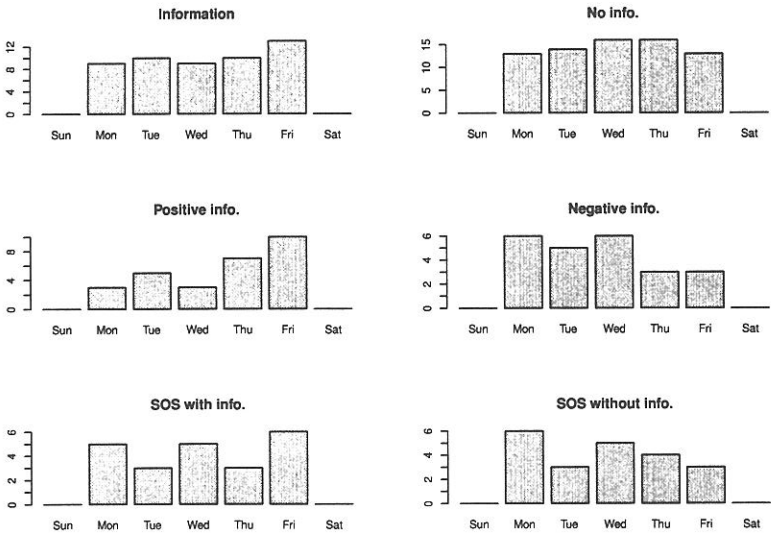


Figure 6: Transaction classify by estimated state for code 1332

Table 5: Log likelihood and Log of marginal likelihood for code 5201

	Log likelihood	M-H	Log of marginal L.	s.e.
$\mathcal{M}_0$	-678.913	0.965	-115.341	0.012
$\mathcal{M}_1$	-656.142	0.960	-115.084	0.014
$\mathcal{M}_2$	-654.596	0.952	-115.734	0.015
$\mathcal{M}_3$	-652.819	0.933	-116.073	0.017
$\mathcal{M}_4$	-646.523	0.820	-122.943	0.905
$\mathcal{M}_5$	-646.476	0.855	-120.401	0.323

Table 6: Summary statistic of posterior distribution of $\mathcal{M}_1$ for code 5201

	mean	s.d.	95%L	median	95%U	Geweke	Inef
γ_I	0.193	0.072	0.069	0.187	0.346	0.338	13.264
γ_P	0.523	0.203	0.147	0.516	0.929	0.947	14.198
γ_S	0.249	0.068	0.121	0.246	0.388	0.312	11.329
ε_b	10.820	0.708	9.353	10.844	12.141	0.624	10.881
ε_s	11.671	0.768	10.049	11.713	13.093	0.692	10.546
u	10.406	1.736	7.289	10.311	14.137	0.462	9.448
Δ	10.055	1.073	8.147	9.983	12.331	0.467	9.820
w	0.454	0.069	0.329	0.450	0.602	0.850	5.640
$adjPIN$	0.087	0.028	0.036	0.086	0.144	0.138	10.923
$PSOS$	0.179	0.046	0.092	0.178	0.271	0.207	10.792

The ML and MCMC estimations for the model without zero inflation do not work due to the excess zero counts. The ZI-PIN model works well for both ML and MCMC estimations. The result of ML estimation is omitted to save space. For MCMC estimation, the log marginal likelihood for each model is reported in Table 5. The selected ZI-PIN model is $\mathcal{M}_1$. The summary of the result is given in Table 6. It is remarkable that the mean of the posterior distribution of w is 0.454. This suggest that we would suffer severe bias if we ignore the data with zero counts of the sample.

6. Conclusion

The excess zero counts we often face in empirical study make an inference by the extended PIN model difficult. Although the one possible way to avoid such difficulty is to discard the data with zero counts, it would incur some bias

for the $adjPIN$ and $PSOS$ introduced in Duarte and Young (2009). On the other hand, the Zero inflated PIN model proposed in this paper enables us to deal with such extra zero counts case. In this paper, we also propose the model selection procedure which does not rely on the asymptotic theory using the marginal likelihood and it makes the model selection easier than the method proposed in the previous study.

References

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

Easley, D. and M. O'Hara (2004), "Information and the cost of capital," *Journal of Finance*, 59, 1553–1583.

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

Easley, D., Hvidkjaer, S. and M. O'Hara (2002), "Is information risk a determinant of asset return?," *Journal of Finance*, 57, 2185–2221.

Geweke, J. (1999), "Using simulation methods for Bayesian econometric models: inference, development, and communication," *Econometric Reviews*, 18, 1–73.

Mullahy, J. (1986), "Specification and testing of some modified count data models," *Journal of Econometrics*, 33, 341-365.

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

Winkelmann, R. (2008), *Econometric Analysis of Count Data*, Springer-Verlag, Berlin.

高頻度データ分析: 不確実性指標と予測可能性 *

慶應義塾大学　林　高樹

2011 年 9 月 15 日

概要

アルゴリズム取引や高頻度トレードが普及してゆく中で, 高頻度データの有効利用が金融機関, 投資家など市場参加者に求められている. 取引所間の国際的な競争, IT 技術の進展を背景に, 各取引市場のマッチングエンジンはますます高速化し, リアルタイムで配信される市場データやそれらをヒストリカルに蓄積して提供される高頻度データも年を追う毎に "高頻度" 化されている. 本研究では, そのような状況の中, 高頻度の価格データが持つ "情報量" を "文脈木" によって実際に計量し, 高頻度 (短時間) における市場の "効率性・非効率性" について考察を加える.

Keywords: 高頻度データ, 可変長マルコフモデル, エントロピー, 市場の効率性, 高頻度トレード, アルゴリズム取引

1　はじめに

本研究では, 金融証券市場の価格やボラティリティの予測性という観点から, 高頻度の価格データが持つ "情報量' の大きさの計測を試みる. 金融証券価格の予測性は, 市場の効率性・非効率性と密接に関連している. 言うまでもなく, 市場の効率性はファイナンス研究分野において実務上も理論上最も重要なテーマの一つであり, その中において, 市場データのみを用いたシグナル生成による売買, 実務における "テクニカル分析" による正のリターンを得る可能性, いわゆる "ウィーク・フォーム" の市場効率性に関する実証研究も多数行われてきている.

90 年代半ば以降, 高頻度データを用いることによって, ごく短い将来の価格変動に対する予測性や (非) 効率性に関する実証研究も行われるようになってきている. 著者の理解によれば, それらはおおむね短い時間におけるある程度の予測性を見出しつつも, 取引コストを勘案すると超過利益をあげるのは難しいという結論のようである. 本研究に関連する先行研究として, まず Papageorgiou (1997), Tanaka-Yamazaki (2003), Ohira et al. (2002) を挙げる. これらの文献では, 外国為替レートのティックデータを用い, 高頻度領域での上下動の時系列に関する条件付き確率の評価を行った. 実証分析の結果, 前二者は, 2 次のマルコフ性, 後者は 4 次のマルコフ性を見出し, (次の値動きに関する) 将来予測の可能性を指摘した.

本稿は, 文脈木 (context tree) を用いて, 高頻度データに含まれる情報量を, データの圧縮性という観点から計量し, さらにそれをベースにした予測可能性という観点から, 市場の効率性について考える. 具体的には, 高頻度時系列データの変化を置き換えた "アルファベット" 列を, 簡潔に符号化 (圧縮) し, 得られた文脈木を使って次に発生するであろう文字 (上下動) を予測することを試みる. 仮に, ある時系列の中に, 繰り返しパターンが存在するとすれば, そのようなパターンの発生頻度が高いほど短い長さの符号で表現した方が効率的である.

---

* Very preliminary and incomplete. Comments and suggestions are welcome.

換言すれば, もとの時系列データに含まれるパターンの有無は, 元のデータがどれだけ “圧縮’ されるかを計量することによってその程度を調べることができる. 本稿で取り上げる, データの効率的な表現 (圧縮) と, 予測を行う研究は, 情報理論, 計算機科学を中心に研究が進められてきた (例, Cover and Thomas (2006)). 今日まで, 文脈木を高頻度データ分析に応用した例はあるものの数は極めて少ない. 文脈木においては, トレーニング列に新しく現れた文字列のパターンは “辞書” に “文脈” として登録され, 一方既に辞書に登録されている文脈は, それがトレーニング列に現れるごとに発生頻度の情報が更新されてゆくことになる. このようにして得られた辞書を用いれば, 登録された文脈の発生頻度情報を使うことによって, 今手元にあるテスト列を辞書に照らし合わせることによって, 次に現れるであろう文字に関する予測をすることができる.

文脈木は, 情報理論分野において Rissanen (1983) によって最初に提案されて以来, バイオ・インフォマティックス, 計量言語学等, 幅広い分野で応用されてきた. 文脈木は, “可変長マルコフ (連鎖) モデル”(Variable Length/Order Markov, 以下 “VOM モデル”) とも呼ばれ, 一言で表せば, マルコフの依存性を表すメモリの大きさが過去の文字列に依存するようなマルコフ連鎖である. メモリを必要に応じてのみ使用することで, データの表現の自由度を維持しつつ, モデルの単純性・倹約性を実現し, これらを両立させていることがこのモデルの成功の要因とされている. 学術研究にとどまらず, “ビッグデータ” を戦略的に活用することが様々なビジネス分野において求められている今日, 文脈木/VOM に代表される大規模データを効率的に圧縮可能な技術への社会的要請は高い.

VOM モデルを, 高頻度データの予測に応用した先攻研究は幾つか存在するが, その利便性・潜在的有用性にかかわらずそれを用いた実証研究が十分に行われているとは言い難い. 本研究は, 国内株式のティックデータに対して同方法論を適用することで, 高頻度領域における株価形成に関する新たな実証的知見を得るのが最終的な目標である.

文脈木によって価格系列の乱雑性 (complexity) を計測し, それを介して市場の効率性を検証しようする試みとしては, 2 値 (上昇, 下落) の日次データについて調べた研究例に, Shmilovici et al. (2003), Giglio et al. (2008) 等が, 3 値 ({ 上昇, 下落, 不変 }) の高頻度データでは Shmilovici et al. (2009) 等の例がある. 本稿は, 主に Shmilovici et al. (2009) の方法論を採用する.

高頻度データの高頻度化が年々進展している今日, より高頻度領域における価格変動の時系列特性を調べる必要性も増している. 例えば, 2010 年 1 月より, 東証において arrowhead が導入されているが, その帰結としての高頻度領域におけるデータ特性について実証的に研究調査することは意義が高いと考える.

2 方法論

準備

以下では, 主に Begleiter et al. (2004) を参考にしながら, 記号を導入し, 問題設定を行う. Σ を有限アルファベット集合とする. 例えば, $\Sigma = \{a, b, c\}$(上昇, 下落, 不変) などである. 学習者 (learner) はトレーニング列 $q_1^n = q_1 q_2 \cdots q_n$ が与えられる, 但し, $q_i \in \Sigma$, $q_i q_j$ は 2 つの文字 (アルファベット) の結合を表す. 学習者の目的は, 学習データ q_1^n をもとに, 過去のある履歴が与えられた時に任意の将来の実現に対する確率を与えるようなモデル $\hat{P}$ を得ることである. 具体的には, 任意の文脈 $s \in \Sigma^*$ と文字 $\sigma \in \Sigma$ に対して, 学習者は条件付き確率評価 (推定値)$\hat{P}(\sigma|s)$ を生成せねばならない. ここで, $\Sigma^* = \bigcup_{k \geq 0} \underbrace{\Sigma \times \Sigma \times \cdots \Sigma}_{k \text{ 個}}$ は, 全ての有限長のアルファベット列から構成される集合である.

文脈木 (可変長マルコフモデル)

固定長の n 次マルコフモデルは, 任意の長さ N の文脈 $s \in \Sigma^N$ と文字 $\sigma \in \Sigma$ に対して, 確率分布 $P(\sigma|s)$ を推定しようとするが, 可変長マルコフモデル (Variabile Order Markov Model, VOM) においては, トレーニング列内に含まれる当該文脈の発生頻度に応じて, 直前の文脈の長さ $|s|$ が変化する. 高次元のマルコフモデルに比して, データの記述性を若干犠牲にすることにより, 相対的に短いトレーニングデータで容易に解析を行うことができることに特長がある.

本稿では, Begleiter et al. (2004) の報告に基づき, VOM アルゴリズムの中でも, 特に, Prediction by Partial Match (PPM) と呼ばれるアルゴリズムの一種を用いることにする. PPM は, スピードに劣るものの, 高い圧縮性能を持つアルゴリズムであることが知られている. 以下では, 特に, Shmilovici et al. (2009) で用いられているバージョンを採用することとする. 以下では, 同論文に沿ってアルゴリズムの概要を紹介する.

通常, VOM の予測アルゴリズムは, カウンティング (counting), スムージング (smoothing), 文脈選択 (context selection) の 3 つのフェーズから構成される. カウンティング・フェーズにおいて, 最大深さ D の初期の文脈木が作られる. 木の根元 (root) から, 節 (ノード) まで伸びる枝 (パス) が一つの文脈 (context) を表す. ここで, 枝を伸ばしてノードを一つ追加することは一つ前の文字を加えて文脈を伸ばすことに対応する (文脈はトレーニング列の中で時間的に遡る方向に現れる). 各ノードは最大 $|\Sigma|$ 個の子ノードを持つが, 木はバランスしている (balanced) とは限らない, すなわち, 全ての枝が同じ長さであるとは限らないし, 全てのノードが同じ個数の子ノードを持つとも限らない.

第 1 フェーズでは, まず文字列を, 一回に一文字づつ構文解析 (parse) する. パースされた文字 σ_i と深さ D の文脈 σ_{i-D}^{i-1} が T 内の仮のパスを決める–もしそれが存在していなければあらたに構築する. 各ノードには, (そのノードに至るパス＝文脈を所与として) 各シンボルの出現回数を数える $|\Sigma|$ 個のカウンターが備えられている. アルゴリズムは, 文脈を次のように更新する. 文脈 σ_{i-D}^{i-1} によって定義されるパスに沿って木を縦断し, 最も深いノード (葉) に到達するまで, 全てのノード内にある文字 σ_i の個数を加えていく. 以下, トレーニング列内の文脈 s の後に文字 $\sigma \in \Sigma$ が現れる回数を $N_\sigma(s)$ と書くことにする.

第 2 フェーズでは, 計測されたカウントを用いて, 予測 $\hat{P}(\sigma|s)$ を生成する. PPM においては, トレーニング列に一度も現れないパターンに発生確率をどう付与するかによって, バリエーションが存在するが, ここでは, 上述の Shmilovici et al. (2009) にの方法をそのまま用いることにする. すなわち,

$$\hat{P}(\sigma|s) = \frac{\frac{1}{2} + N_\sigma(s)}{\frac{|\Sigma|}{2} + \sum_{\sigma' \in \Sigma} N_{\sigma'}(s)}$$

によって計算する.

第 3 フェーズでは, トレーニング列へのオーバー・フィッティングを避け, かつメモリ領域の節約と計算効率向上のために, フェーズ 2 で得られた暫定モデルの大きさの縮小を図る. いま, $s = \sigma_k \sigma_{k-1} \cdots \sigma_1$ が葉ノードに到達している時, その親ノードは, s の最長の "接尾語"(suffix), $s' = \sigma_{k-1} \cdots \sigma_1$, である. そこで, 親ノードと比較して, 次の文字 σ の予測において, 追加情報量の大きさの点で貢献しない葉を刈り込む (prune). 具体的には, 親子の符号長 (理論値) の差が $C(|\Sigma|+1)\log(N+1)$ 以上である時のみ, 子ノード s を保持する (Shmilovici et al. (2003)). 但し, C は定数である.[*1] 枝の切り落としは, 文脈木の中で, 再帰的に実行されて

[*1] 同論文では, Rissanen (1983) を参考に $C = 2$ とおいている. 一方, 続編の Shmilovici et al. (2009) においては, $C = 0.50$ と設定されている. なお, これらの文献には誤植と思われる箇所がある.

ゆく.

以上が情報圧縮のアルゴリズムであるが, これにより得られた VOM モデル (文脈木) を用いて, 次の文字に対する予測を行うことができる. すなわち, (トレーニング列とは別の) 文字列 s が与えられた時, 次の文字の予測値 $\hat{\sigma}$ は, VOM モデルにおける発生確率が最大となるような文字を選択すれば良い, すなわち, $\hat{\sigma} = \arg\max_{\sigma'} \hat{P}(\sigma'|s)$ のように選べばよい.

圧縮率、予測精度

トレーニング列をどれだけ良く学習できたかを評価する量として, 圧縮率を用いる. アルファベット集合 Σ を持つ長さ w の文字列 $x_1^w = x_1 x_2 \cdots x_w$ が, 完全に "予測不能"(一様確率でかつ独立) に並んでいるのであれば, これを 0,1 で表現するのに必要なビット数は, $w \log_2 |\Sigma|$ である (もちろん, 非対称なケースはその分布のエントロピーを計算すれば良い). 一方, VOM によって符号化 (圧縮) された符号の長さは, $-\log_2 \hat{P}(x_1^w) = -\sum_{t=1}^{w} \log_2 \hat{P}(x_t|x_1 \cdots x_{t-1})$ にて与えられる. したがって, 両者の比である圧縮率

$$r = \frac{-\log_2 \hat{P}(x_1^w)}{w \log_2 |\Sigma|}$$

を見れば, トレーニング列のランダム性・不規則性 (この場合, 一様かつ独立) を評価することができるはずである. 従って, 対称な単純ランダムウオークをモンテカルロシミュレーションで多数発生させそれを VOM モデルに学習させることで, "効率的市場" の下での圧縮率の従う参照分布が得られるから, 実証データより計測される圧縮率に対しているしては, その (経験)p 値が計算される.

次に, 通常の判別分析の評価と同様, モデル予測値と実績値の正答率を評価することもできる. すなわち, 文脈 s に対して確率評価 $\hat{P}(\sigma|s)$ を最大にするような文字の予測値 $\hat{\sigma}$ を, 実現値 $\tilde{\sigma}$ と比較し, 正答率, 誤答率をデータ期間にわたり評価すれば良い. これはトレーニング列 (内挿データ), テスト列 (外挿データ) のどちらにおいても計測可能である. たとえば, アルファベット集合が 3 文字 $|\Sigma| = 3$ であれば, 9×9 の行列 ("confusion matrix") で表現することができる. 情報分野, 機械学習分野で用いられる評価尺度である, "F 尺度", "再現率 (recall)" などを計算しても良い.

3 実証分析

データ

以下では, 日経 NEEDS 提供の個別株式ティック・データを用いた分析結果を紹介する. 使用したティック・データは, 東証一部銘柄の中で時価総額や流動性の特に高いものから構成される TOPIX コア 30 構成銘柄 (2010 年 10 月末時点) の約定データである. データ期間は, 東証に arrowhead の導入された 2010 年一年間とし, 市場営業日別に分析を行う. 各日内において, トレーニング列を "動的" に生成 (rolling window) し, 圧縮および 1 時点先予測を行い, これを一時点ずつずらしながら終値まで繰り返し, その日のパフォーマンスを集計する. 分析に用いた時間軸は, 通常の物理時間 (暦時間) ではなく, 取引ごとに時計が進むと考える "トレード時間" である (cf. Griffin and Oomen (2008)). 分析を行うために, 生のティック・データを次のように加工した.

- データをティック時間刻みで間引き (飛ばし幅 $m = 1, 5, 10, 20, 50$) する ($m = 1$ は全データ使用).[*2]

[*2] 約定ティック・データは "ビッド・アスク・バウンス" と呼ばれる "マイクロストラクチャ・ノイズ" の影響が強いことが知られて

- 価格変化を, 文字数 2 のアルファベット, $\Sigma = \{a, b\}$ に変換する (a は上昇, b は下落). なお, ティック・データには, 価格変動ゼロの取引の割合が高いが, このような "ゼロ・リターン" があった場合には, 一つ前の値動きの方向 (上昇, 下落) を継続させることにする.
- トレーニング列のサイズ (ウィンドウ幅) として, 上記文献を参考に, また, 上記 m の大きさ, 各銘柄の 1 日当たりのティックデータの長さを考慮に入れ, $w = 50$(区間) と設定する.[*3]
- 文脈木の深さの最大値としては, やはり同文献を参考に $D = 6$ と設定する.[*4]

分析結果

ページ数の都合上, Topix コア 30 銘柄のうちの一つ, 小松製作所 (6301) の結果のみを図 1–3 に示す.

各図の左が, 一日内における平均正答率 (外挿予測) 対平均圧縮率の散布図である. 各点が 1 日に相当する. 一方, 右図は, 平均正答率とその日における (外挿期間内の) 上昇回数の相対頻度の時系列プロットである. 図 1 は, $k = 1, 5$ のケース, 図 2 は $k = 10, 20$, 図は $k = 50$ のケースである. 上昇回数の相対頻度とは, 仮にモデルなしにつねに上昇を予測した場合の勝率である. 図より, 平均圧縮率が高い日は 平均正答率も高い (左肩上がり) ことが明らかである. また, 平均正答率は, 対応する平均上昇率よりも明らかに大きい. これらの傾向は k が大きくなってゆくにつれ, 軽減されてゆき, $k = 50$ では消滅することが分かる.

図 4–6 に, セブン&アイ・ホールディングス (3382), 武田薬品工業 (4502), 新日本製鉄 (5401) の結果 (いずれも, $k = 1, 5$ のケースのみ) を示す. 小松製作所 (6301) と同様な結果が得られた.

課題

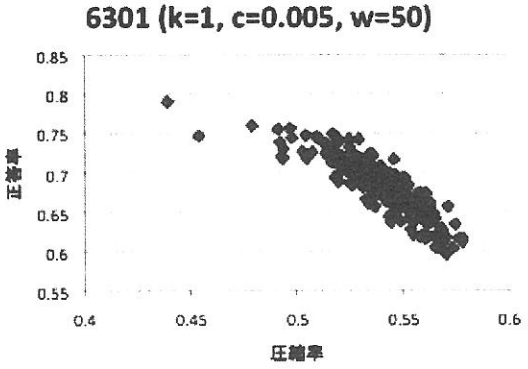
実のところ, 今回使用のウィンドウ幅 $w = 50$ は, (フェーズ 2 においた使用した枝伐採のための係数 $C = 0.005$ に対して) 十分長いとは言えない. 一方, 事前の予備的分析において, C の値をいろいろと変えて結果を眺めてみたところ, 先行研究 (Shmilovici et al. (2009)) にで設定された C の値が今回のデータに対しては大きすぎることが判明したため, 暫定的ではあるが $C = 0.005$ を選んだ ("最適化" された数値ではない). 実証分析の前に, 2 値の乱数列 ("公正な" コイン) を同じ長さ ($w = 50$) で 1 万回発生させ, 作成した PPM アルゴリズムを適用して圧縮率の分布を調べたところ, $C = 0.005$ の時の 10 %点が 0.5253, 90% 点が 0.6312 であった. 一方, $C = 0.05$ の時には, 0.6723, 0.866, $C = 0.01$ の時には, 0.8262, 0.9896 であった. C の値が小さい場合にはオーバーフィッティングする可能性がある半面, 大きな C の値では, 木を伐採しすぎ, うまく学習できなくなる恐れがある. モンテカルロ実験では $C = 0.25$ 程度の大きさの時, 文脈木のノード数は平均して約 1 個 (つまり, メモリをもたない) にまでなることが確認された. 従って, 上記の実際の高頻度データを用いた分析において, 得られた圧縮率がおおむね 0.4–0.6 の範囲に収まっているが, これをもって直ちにデータが (対称な) ランダムデータではないと断じることができない (市場の "効率性" を棄却できない).

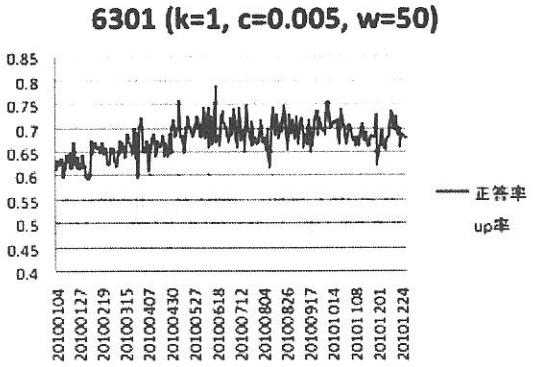
そもそも, (相対的に短いデータに対して) 圧縮率が高いことが, 学習がうまくいっているためなのか, オーバーフィッティングしているだけなのかこれだけの材料では判断ができず, w と C の望ましい値について今後の検討を要する.

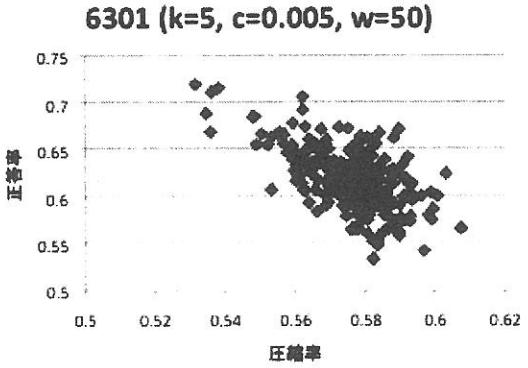
いるが, 本研究では, (超) 高頻度領域での価格変動の記述が目的であることから, 敢えてこれを除くような処理は施さないことにした.

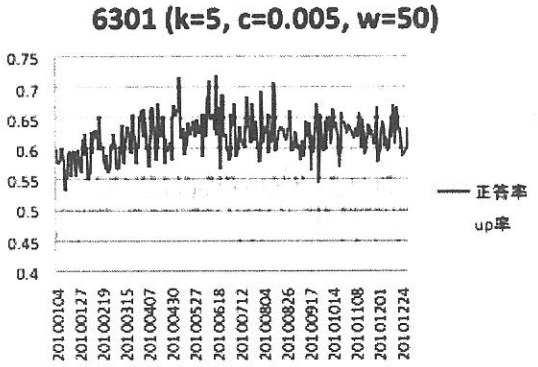
[*3] Shmilovici et al. (2009) は, $w = 50, 75, 100$ を使用している. フェーズ 2 における C の選択とともに, 望ましい w の大きさは今後の検討課題である.

[*4] 文献では, $D \leq \log(N + 1)/\log|\Sigma|$ が推奨されているようである.

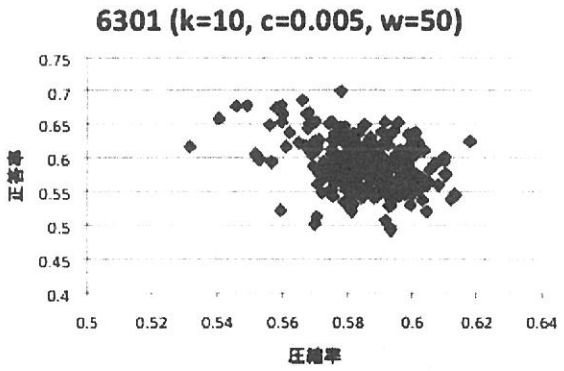


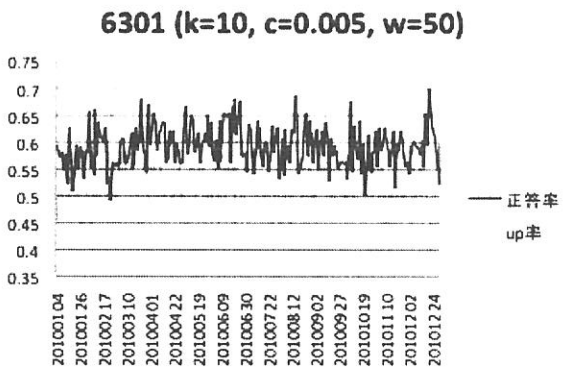


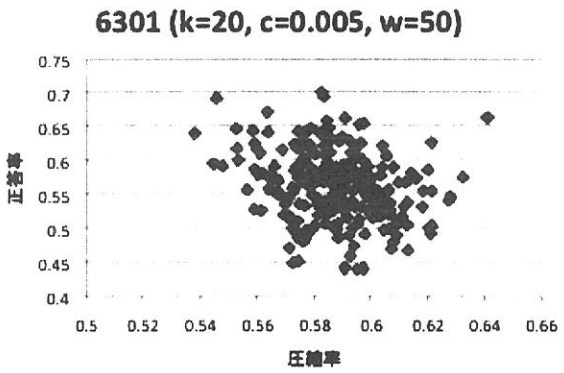


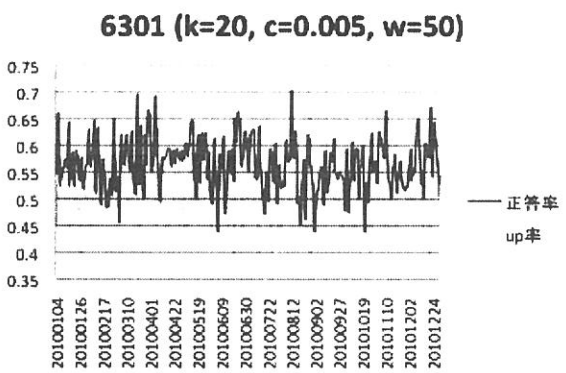


一方, 上図より, 圧縮がうまく行われる時, 正答率が上昇する関係が明確に見られる. 圧縮率は日によって異なるが, 少なくとも今回構築した PPM が直前 50 個のデータ列のパターンをうまく学習し, 一つ先をうまく予測していることを示している. 使用データの長さが銘柄によって異なるが, 例えば, 小松製作所 (6301) の場合, 2010 年一日あたり平均 5249 件のレコードがあり, これより学習・予測を一日当たり (rolling window を 1 ステップずつオーバーラップさせて)$k = 1$ のケースでは 5200 回程度行っている. 従って, 仮に価格変動の上下が等確率なランダム系列であれば, この "外挿" 予測の平均正答率に対する標準偏差 (約 0.007) の大きさに対し, 上図の平均正答率 (モデル予測対実際の上昇比率) の差が殆どの日で有意となることは明らかである (等確率でない場合には 2 項分布の標準偏差はさらに小さくなるので市場の "効率性" は棄却される!?). しかしながら, 時間間隔が長くなる (k の値が大きくなる) に従いそのような収益機会は消滅してゆく. 恐らく (やはり) ビット・アスク・バウンスが主要因と考えられるが, それ以外の要因の可能性も含め今後の詳細な分析が必要である. なお, 見かけ上の "非効率性" が存在したとしてもそれを実際に利用して超過利益を挙げられなければ市場が "効率的" であるとは言えないのは当然であるが (例えばビッド・アスク・バウンスは指値の待ち行列のためそこから収益を上げるのは単純ではない), IT 技術がすさまじい勢いで発展し, コロケーション・サービスなどが提供されるようになっている今日, 今回確認された瞬間的に消えてしまう "非効率性" は, 他者より早







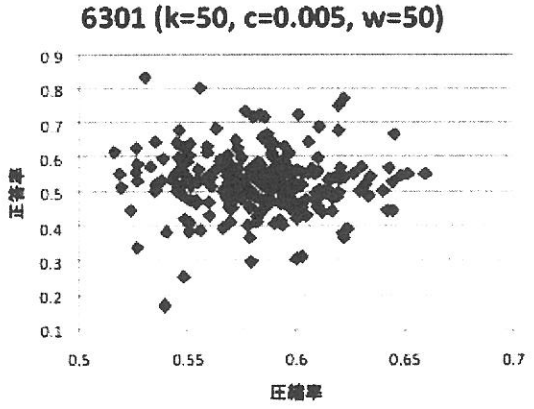


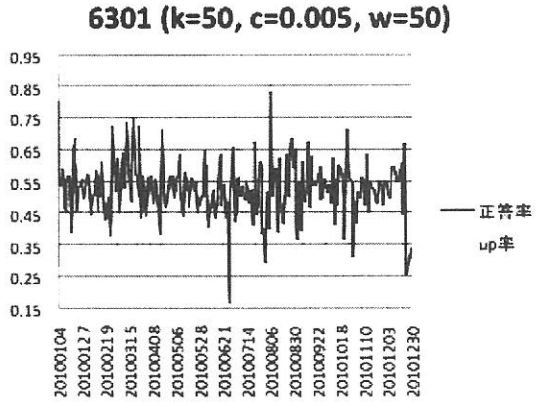
く市場情報を入手し分析し, 他者より早く発注する高速売買実践者にとってのまさに収益源なのかもしれない.

4 まとめ

本研究は, 高頻度領域における価格変動の記述に焦点を当てたものである. 方法論として, 文脈木を採用し, 東証第一部の代表的銘柄数銘柄について, それらの価格変動の情報量, 予測性について調査した. その記述性と簡便性を特徴とする文脈木の, 金融データ, なかんずく高頻度データへの応用研究はこれまで多くはなく, 同方法論を用いた実証研究の余地は今後大きいと考えられる

今回は, 圧縮率の高いことで知られている PPM アルゴリズムの一種を用いたが, 生成される文脈木はアルゴリズムのチューニング・パラメータ, 特にフェーズ 2 における係数 C の大きさに大きく依存する. トレーニング列の適正なサイズ w の値と共に, どのようにこれらを選択すれば良いかという問題は今後の調査が必要である. また, VOM アルゴリズムとしては, PPM 以外にも, CTW(Willems et al. (1995, 1996, 1998)) などの有力なものが知られている. これを含めた他のアルゴリズム間で 予測性能に関する比較検討も課題である (例, Giglio et al. (2008)).





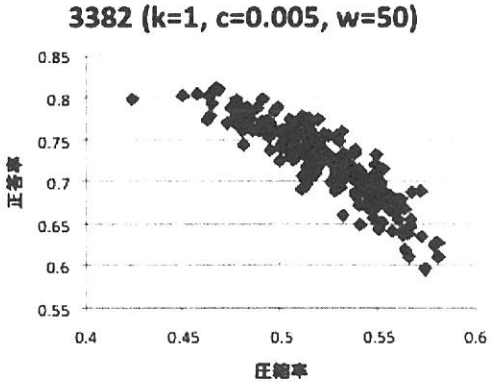
当然ながら, 市場の "効率性" の程度を検証するためには, 超過収益の獲得可能性について, 予測シグナルに基づく模擬売買などによって調べることが必要である. 既存研究 (例えば, 上記の Shmilovici et al. (2009)) においては, 仮に将来に関するある程度の予測力を持つモデルが得られても, 取引コストを考慮すると有意な収益が得られないというのが, 通常なされる報告である. このような取引コストを勘案した上での収益性の評価は市場の (弱い意味での) 効率性を評価する上で重要である.

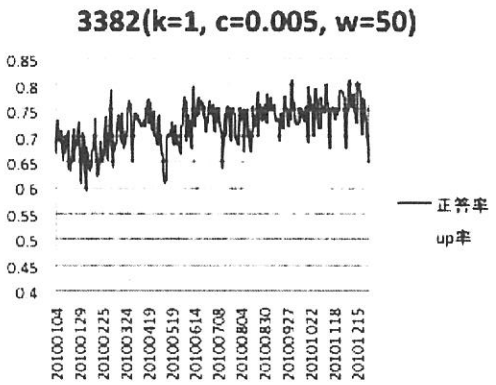
高頻度データは, 実のところ, そのサンプリング頻度によって, 時系列特性が変化することが指摘されている (例えば, Mandelbrot らのグループの提唱する "マルチフラクタル" 性など; Calvet and Fisher (2008)). VOM はデータの記述性が高く応用範囲が広いとされるモデルではあるが, そもそも (それよりも記述性の高い) 高次数のマルコフ性すら, 金融市場における高頻度データを記述するには適切でない可能性がある. マルチフラクタル性などの実証的性質を考慮に入れながら, 例えば異なる頻度によって得られた分析結果間の比較など, 今後より詳細な調査, 検証を通じて, VOM モデルの適用妥当性について考察を進める必要があろう.

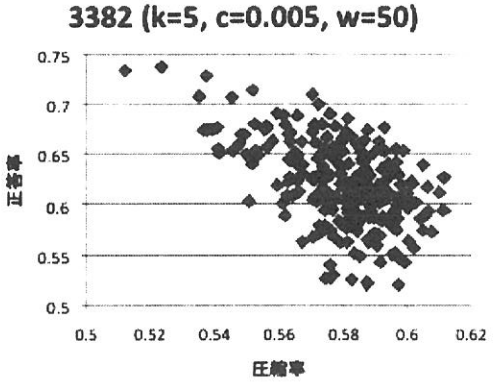
複数ある文脈木の候補群から適当なものを一つ選ぶという文脈木選択問題も, 理論上, 応用上ともに重要なテーマである (Garivier and Leonardi (2011)). "予測力" の高いアルゴリズムの開発とともに, 長期的かつ大きな研究課題である.

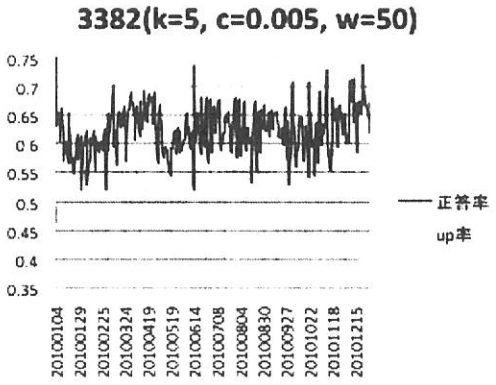
謝辞

本研究は, 文部科学賞科学研究費研究 (基盤研究 (A)「ファイナンス計量分析の新展開と金融市場」(課題番号 21243019), 研究代表者: 東京大学国友直人教授) および, 日本証券奨学財団研究調査助成金からの資金援助









により行われている．研究遂行にあたり，みずほ第一フィナンシャルテクノロジー社藤野直樹氏より，実務面からのコメントを頂いた．ここに謝意を表する．

参考文献

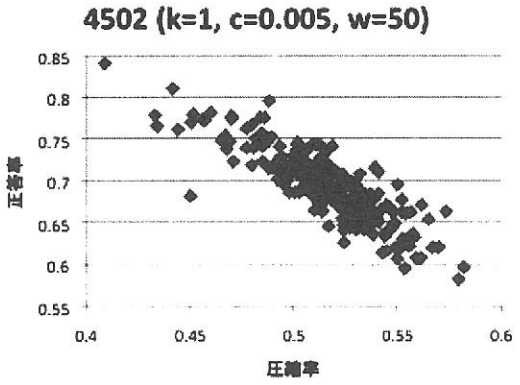
Begleiter, Ron, Ran El-Yaniv, and Golan Yona (2004) "On Prediction Using Variable Order Markov Models," *J. Artificial Intelligence Res.*, Vol. 22, pp. 385–421.

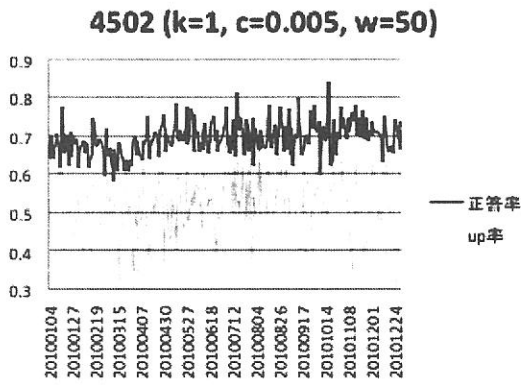
Calvet, Laurent E. and Adlai J. Fisher (2008) *Multifractal Volatility: Theory, Forecasting, and Pricing*, Burlington: Academic Press.

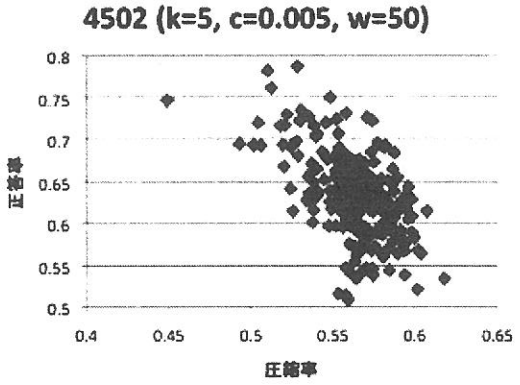
Cover, Thomas M. and Joy A. Thomas (2006) *Elements of Information Theory*, New York: Wiley.

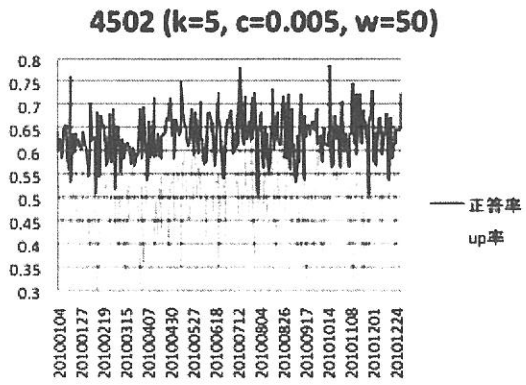
Garivier, Aurélien and Florencia Leonardi (2011) "Context Tree Selection: A Unifying View," *Stoc. Proc. Appl.*, Vol. (in press).

Giglio, Ricardo, Raul Matsushita, Figueiredo Annibal, Gleria Iram, and Da Silva Sergio (2008) "Algorithmic Complexity Theory and the Relative Efficiency of Financial Markets," *Europhysics Let.*, Vol.









84, No. 4, p. 48005.

Griffin, Jim E. and Roel C. Oomen (2008) "Sampling Returns for Realized Variance Calculations: Tick Time or Transaction Time," *Economet. Rev.*, Vol. 27, pp. 230–253.

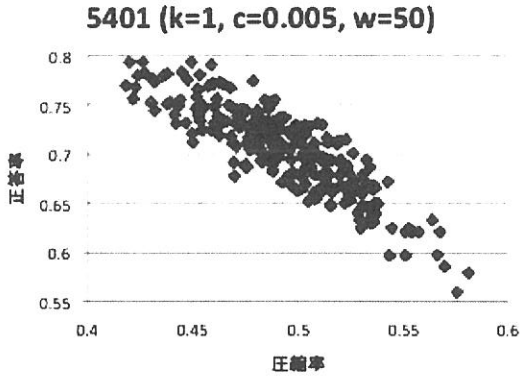
Ohira, Toru, Naoya Sazuka, Kouhei Marumo, Tokiko Shimizu, Misako Takayasu, and Hideki Takayasu (2002) "Probability of Currency Market Exchange," *Physica A*, Vol. 308, pp. 368–374.

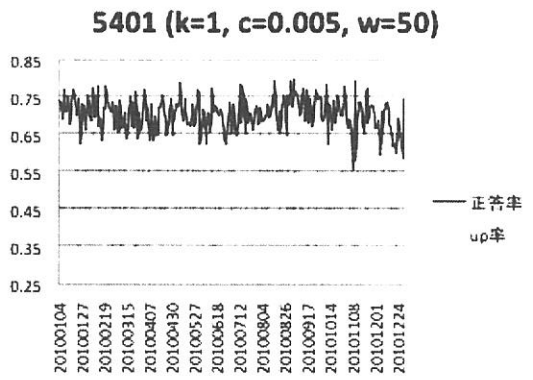
Papageorgiou, Constantine P. (1997) "High Frequency Time Series Analysis and Prediction Using Markov Models," *Computational Intelligence for Financial Engineering (CIFEr), Proceedings of the IEEE/IAFE 1997*, pp. 182–188.

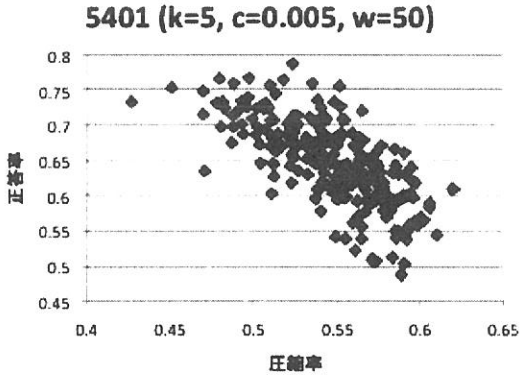
Rissanen, Jorma (1983) "A Universal Data Compression System," *IEEE Trans. Inform. Theory*, Vol. 29, No. 5, pp. 656–664.

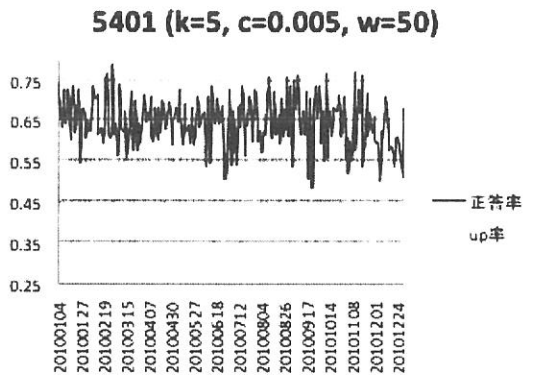
Shmilovici, Armin, Yael Alon-Brimer, and Shmuel Hauser (2003) "Using a Stochastic Complexity Measure to Check the Efficient Market Hypothesis," *Comput. Econ.*, Vol. 22, pp. 273–284.

Shmilovici, Armin, Yoav Kahiri, Irad Ben-Gal, and Shmuel Hauser (2009) "Measuring the Efficiency of the Intrady Forex Market with a Universal Data Compression Algorithm," *Comput. Econ.*, Vol. 33,









pp. 131–154.

Tanaka-Yamazaki, Mieko (2003) "Stability of Markovian Structure Observed in High Frequency Foreign Exchange Data," *Ann. Inst. Statist. Math.*, Vol. 55, No. 2, pp. 437–446.

Willems, Frans, Yuri Shtarkov, and Tjalling Tjalkens (1995) "Context Tree Weighting: Basic Properties," *IEEE Trans. Inform. Theory*, Vol. 41, pp. 653–664.

Willems, Frans, Yuri Shtarkov, and Tjalling Tjalkens (1996) "Context Tree Weighting for General Finite-Context Sources," *IEEE Trans. Inform. Theory*, Vol. 42, pp. 1514–1520.

Willems, Frans, Yuri Shtarkov, and Tjalling Tjalkens (1998) "The Context-Tree Weighting Method: Extensions," *IEEE Trans. Inform. Theory*, Vol. 44, pp. 792–798.