マーク付き多次元 Hawjes 過程を用いた高頻度注文版データの分析

木下大輔

富山県立大学　電子情報工学科

April 23, 2021

はじめに

背景

近年、情報通信技術の発達や情報処理機器の高性能化に伴い、世界中の金融市場において取引システムの高速化や、取引データの時間解像度の向上が測られている。これらの現状を踏まえ日中データと呼ばれる日時以上の頻度で取引が記録されたデータを用いた実証分析が行われている。

目的

- 取引数と注文数をマークとしたマーク付き４次元 Hawkes 過程を提案する。
- 上記で提案したモデルを実際の複数銘柄に当てはめ、それぞれのパラメータ推定を行う。
- 推定されたパラメータを利用して時系列と銘柄間の比較を行い、各イベントの自己励起性、及びイベント間の相互励起性についての特徴を明らかにする。
- マークが Hawkes 過程の強度関数に与える影響の有無とその大きさについての分析を行う。

マーク付き多次元 Hawkes 過程とは

Hawkes 過程はイベント発生をモデルかするのに用いられてきた確率過程である。特徴としてあるイベント発生が同種類あるいは他の種類のイベントの次の発生確率を瞬間的に変化させる自己励起性あるいは相互励起性といった性質を明示的かつシンプルに記述できるということである。マーク付き多次元 Hawkes 過程はイベントの発生時点とそのイベントに紐づくマークを点過程として表現するモデルである。

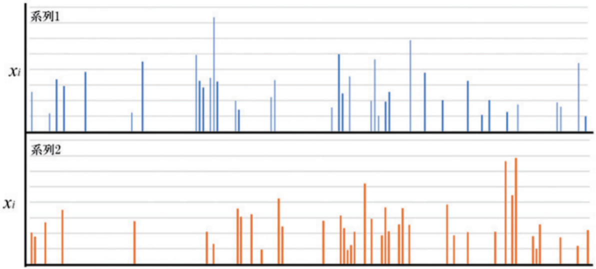


図 1: 2 次元のマーク付き Hawkes 過程の例

導入

要素

- T, X:完備過分距離空間
- (t_i, d_i, x_i):ある加算集合 J の元 $i \in J$ に対して、観測されたイベント
- $t_i \in T$:イベントが観測された時刻
- $d_i \in 1, .., d, d \in \mathbb{Z}_+$:イベントの系列番号
- $\chi_i \in X$:イベントのマークの大きさ
- (t_p, x_i):$T * X$ 状の点過程 $N = (t_i, x_i)$
- $N_{gi}(A) = N_j(A)$:点過程 N に関してマークを考えない基底過程
- (Σ, F, P, F_t):フィルター付き確率空間

$N_j(t)$ の強度関数 $\lambda_j(t)$

$$\lambda_j(t)dt := \mathbb{E}[N_j(t+dt) - N_j(t) | F_{t-}]$$

導入

要素

- $F_{t-} := \sigma(U_{s<t}, F_s)$:時点 $t > 0$ の直前までの事象に対する σ 加法族

$N_j(t)$ の強度関数 $\lambda_j(t)$

マーク付き d 次元 Hawkes 過程 N の下で第 j 番目、$j \in 1, ..., d$ の点過程 N_j の強度関数

$$\lambda_j(t) := \eta_j + \sum_{k=1}^{d} \theta_{jk} \int_{(-\infty,t)*R} \omega_j(t-s) g_k(x) N_k(ds * dx) \tag{1}$$

導入

要素

(1) 式の右辺第二項のうち $g_k(.)$ はインパクト関数と呼ばれている。本研究ではインパクト関数として以下のような線形関数を仮定する。

インパクト関数

$$g_j(x) = \frac{1 + (x_k - 1)p_j}{\mathbb{E}[1 + (x_j - 1)p_j]} \tag{2}$$

カーネル関数

(1) 式の $\omega_j()$ はカーネル関数と呼ばれあるイベントのインパクトが時間とともに減衰する速さを表す関数である。

$$\omega_j(t) = \alpha_j e^{-\alpha_j t}$$

パラメータの推定

本研究では最尤推定法を用いて多次元のマーク付き Hawkes 過程のパラメータ推定を行う。最初に推定区間 D における補正項 $A_j(t)$ を導入する。区間 $D = [T..T^*]$ において観測される多次元過程の補正項を以下のように定義する。

$$A_j(t) = \int_T^t \lambda_j(s)ds, forj \in 1, ...d \tag{3}$$

パラメータの推定

マーク付き Hawkes 過の補正項は以下のようになる

$$\Lambda_j(t) = \eta_j(t - T_*) + \sum_{k=1}^{d} \theta_{jk} \int_{(-\infty,t)\times\mathbb{R}} [\bar{\omega}_j(t-u) - \bar{\omega}_j(T_* - u)] g_k(x) N_k(\mathrm{d}u \times \mathrm{d}x), \tag{4}$$

指定区間 D を用いてマーク付き Hawkes 過程のパラメータを推定することを考えるこの時の対数最尤度関数は

$$\log L = \sum_{j=1}^{d} \int_{[T_*,T^*]\times\mathbb{R}} \log\lambda_j(t) N_j(\mathrm{d}t \times \mathrm{d}x) + \sum_{j=1}^{d} \int_{[T_*,T^*]\times\mathbb{R}} \log f_j(x) N_j(\mathrm{d}t \times \mathrm{d}x) - \sum_{j=1}^{d} \Lambda_j(T^*), \tag{5}$$

パラメータの推定

また指数カーネルの下で、強度関数 $\lambda_j(t)$ は

$$
\begin{aligned}
\lambda_j(t) &= \eta_j + \sum_{k=1}^{d} \theta_{jk} \int_{(-\infty,t)\times\mathbb{R}} \alpha_j\, e^{-\alpha_j(t-s)} g_k(x) N_k(\mathrm{d}s\times \mathrm{d}x) \\
&= \eta_j + e^{-\alpha_j(t-r)} \sum_{k=1}^{d} \theta_{jk} \int_{(-\infty,r)\times\mathbb{R}} \alpha_j\, e^{-\alpha_j(r-s)} g_k(x) N_k(\mathrm{d}s\times \mathrm{d}x) \\
&+ \sum_{k=1}^{d} \theta_{jk} \int_{[r,t)\times\mathbb{R}} \alpha_j\, e^{-\alpha_j(t-s)} g_k(x) N_k(\mathrm{d}s\times \mathrm{d}x) \\
&= \eta_j + e^{-\alpha_j(t-r)}[\lambda_j(r) - \eta_j] + \sum_{k=1}^{d} \theta_{jk} \int_{[r,t)\times\mathbb{R}} \alpha_j\, e^{-\alpha_j(t-s)} g_k(x) N_k(\mathrm{d}s\times \mathrm{d}x),
\end{aligned} \tag{6}
$$

上記による変形は過去の 1 時点における $\lambda_j(r)$ を用いて再帰的に計算可能であることを示している。推定区間 D 匂いて観測されたデータを $(t_1, d_1, x_1.., t_m, d_m, x_m, ..., t_n, d_n, x_n)$ とおく。すると定量 $logL$ は以下のように書くことができる。

$$\log\hat{L} = \sum_{m=1}^{n} \log\hat{\lambda}_{d_m}(t_m) + \sum_{m=1}^{n} \log f_{d_m}(x_m) - \sum_{j=1}^{d} \hat{\Lambda}_j(T^*). \tag{7}$$

補正項の推定量 $A_j(T^*)$ は (4) 式より

$$\hat{\Lambda}_j(T^*) = \eta_j(T^* - T_*) + \sum_{m=1}^{n} \theta_{j,d_m}\overline{w}_j(T^* - t_m)d_m(x_m),$$

使用するデータ

FLEX Full データ

本研究でパラメータ推定に使用するデータは FLEX Full データである。これは東京証券取引所が提供するサービスであり、現在値、売買高、売買代金、全気配値段・数量・注文件数、引け条件付注文数量及び件数がタイムスタンプと共に記録されている。

注文版データ

このデータは電文形式の注文フローデータであると同時に、注文版データとみなすこともできる。特定の 1 銘柄についてある時刻までの残存気配データに注目するとき注文板を再現することが可能である。株式市場における注文版とは、ある 1 銘柄の株式について市場参加者が行った売買注文の集合である。注文版に表示される注文は指値注文と呼ばれ、市場参加者がいくらで (気配価格)、どれだけの株式を (気配数量)、売り (アスク) または買い (ビッド) を取引サーバに送信したものである。

データの抽出

系列データの抽出

本研究では、系列データとして以下の 4 種類のイベントを用いる。これらは価格の弾力性や価格形成に大きな影響を及ぼすような注文版上のイベントであることが分かっている。

イベント

- 系列 1:直前の取引よりも価格上昇を示す取引:時刻 t_1, 取引量 x_i,$d_i = 1$
- 系列 1:直前の取引よりも価格下降を示す取引:時刻 t_1, 取引量 x_i,$d_i = 2$
- 系列 3:取引成立を伴わないが板中心が変化するビッド側の指値注文, キャンセル注文:時刻 t_1, 注文量 x_{di}
- 系列 4:取引成立を伴わないが板中心が変化するアスク側の指値注文, キャンセル注文:時刻 t_1, 注文量 x_{di}

データの抽出

銘柄内の全取引データから、系列 1-4 に該当するイベントを時刻 t_1, 取引量または注文量 x_i と共に抽出し、整理すると下の図のようになる。

Ask	Price	Bid
45	2,004	
54	2,003	
67	2,002	
25	2,001	
	2,000	
	1,999	
	1,998	13
	1,997	29
	1,996	32

Ask	Price	Bid
45	2,004	
54	2,003	
67	2,002	
25	2,001	
	2,000	
	1,999	16
	1,998	13
	1,997	29
	1,996	32

Ask	Price	Bid
45	2,004	
54	2,003	
67	2,002	
	2,001	
	2,000	
	1,999	16
	1,998	13
	1,997	29
	1,996	32

図 2: 注文版上の変化の再現

モデリングの有効性の検証

バーコードプロット

図 3 は 2019 年の 5 月 31 日のトヨタの 8001 番目から 9000 番目のデータに対してフィッティングを行った結果に示したバーコードプロットである. 全ての図において横軸は時間を縦軸はイベントの発生を表している。また中段は元のデータ、冗談は元データの先頭 50 サンプル分を抽出したもの、そして下段が時間変更の処理を行った後のバーコードプロットを表している。

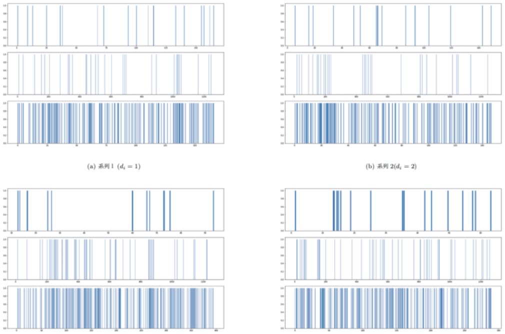
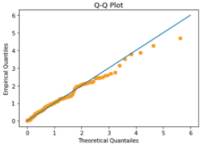


図 3: トヨタの株価に関する 4 系列のバーコードプロット

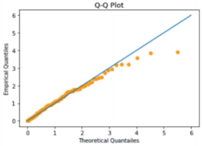
モデリングの有効性の検証

Q-Q プロット

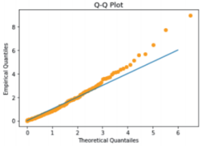
図 4 は 2019 年の 5 月 31 日のトヨタの 8001 番目から 9000 番目のデータに対して時間変更を行い、Q-Q プロットを作図したものである。横軸には想定している分布の分位点を、縦軸に比較するデータの経験分布の分位点をプロットする。



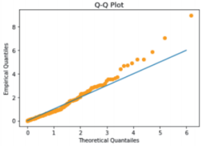
(a) 系列 1 $(d_i = 1)$



(b) 系列 $2(d_i = 2)$



(c) 系列 $3(d_i = 3)$



(d) 系列 $4(d_i = 4)$

図 4: トヨタの株価に関する 4 系列の Q-Q プロット

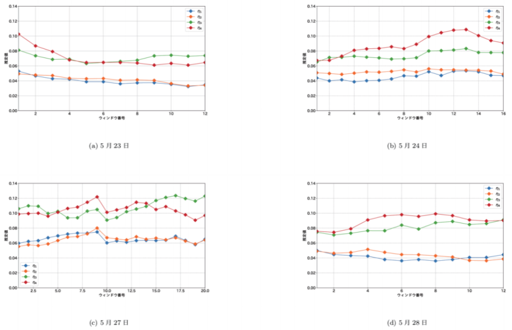
時系列方向の分析

日中、週内の分析

日中及び週内で Hawkes 過程のパラメータがどのように推移していくのかを検証する。具体的にはウィンドウ幅を１０００とし、１００データずつ開始と終了をずらしながらフィッティングを行うことでパラメータの推定値がどのように変化するのかを検証する。用いるのは 2019 年 5 月 23 日から 5 月 28 日営業日分のトヨタのデータである。

時系列方向の分析

初めにベースライン強度 $\eta = \eta_1, \eta_2, \eta_3, \eta_4$ の変化について検証する。図 5 は営業日それぞれの 1 日内でベースライン強度 η が時間と共にどのように推移するかをまとめたグラフである。比較のため縦軸のスケールを統一している。



(a) 5 月 23 日

(b) 5 月 24 日

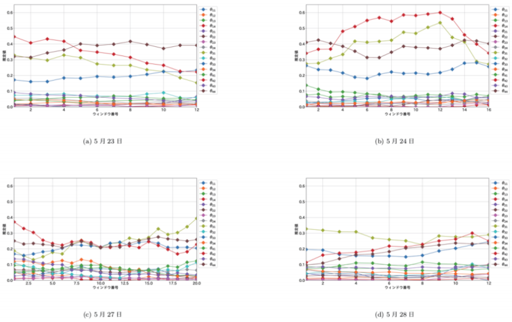
(c) 5 月 27 日

(d) 5 月 28 日

図 5: トヨタに関する η の推定値の変化

時系列方向の分析

次に分枝行列 Θ の時系列変化に着目する。Θ は強度関数の中で自己励起性及び相互励起性を表現するパラメータである。すなわち Θ の要素 θ_{jk} は系列 k で発生したイベントが系列 j の強度関数に与えるインパクトの大きさを表す。図 6 は営業日それぞれの 1 日内で、分枝行列の 16 要素それぞれの値が、時間と共にどのように推移するかをまとめたグラフである。

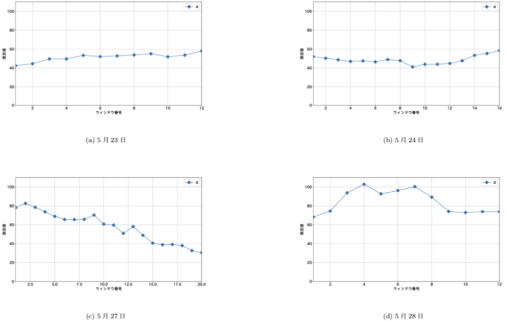


(a) 5 月 23 日　(b) 5 月 24 日　(c) 5 月 27 日　(d) 5 月 28 日

図 6: トヨタに関する Θ の推定値の変化

時系列方向の分析

次はカーネル関数のパラメータ α の変化についてである。α はイベントの発生が強度関数に与えるインパクトの減衰速度を表現したパラメータであり、その数値が大きいほどインパクトが時間経過に対して急速に減衰することを意味する。図 7 は α の時系列変化をプロットしたものである。



(a) 5 月 23 日

(b) 5 月 24 日

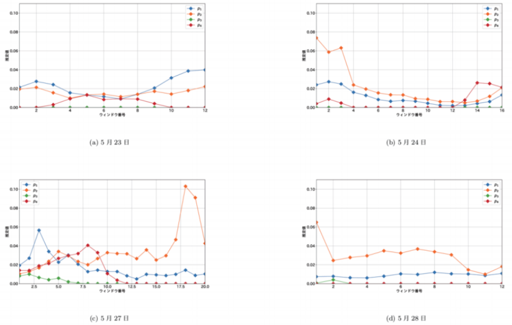
(c) 5 月 27 日

(d) 5 月 28 日

図 7: トヨタに関する α の推定値の変化

時系列方向の分析

最後はマークの大きさが系列に与える影響を分析するため (2) 式のインパクト関数 g_i のパラメータ p_j の変化を見る。p_j は大きい時、あるイベントのマークが平均を上回ったときに郷土館数へのインパクトを増加させ、下回ったときには現象させる。図 8 はそれぞれの営業日ごとに、時間軸に対して p の変化をプロットしたグラフである。

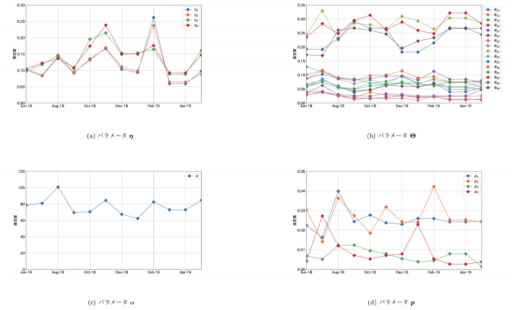


(a) 5 月 23 日 (b) 5 月 24 日 (c) 5 月 27 日 (d) 5 月 28 日

図 8: トヨタに関する p の推定値の変化

時系列方向の分析

先ほどまでのパラメータをさらに長いスパンで確認した場合それぞれのパラメータはどのような推移を見せるかを確認する。用いるデータは 2018 年 6 月 1 日から 2019 年 5 月 31 日のトヨタの注文データである。



(a) パラメータ η (b) パラメータ Θ (c) パラメータ α (d) パラメータ p

図 9: トヨタの月ごとの推定値の平均値の変化

銘柄間の分析

次に複数の銘柄に関する実証を行い比較をする。比較するのは、楽天、ソニー、トヨタ、ソフトバンク、国際帝石、丸紅、三井不動産の計 7 つの銘柄についてそれぞれ 5 月 7 日から 5 月 31 日の間の営業日分のサンプルをウィンドウサイズ 1000 としてパラメータ推定を行う。

	η_j				$\mathbb{E}[x_j]$				α
銘柄	1	2	3	4	1	2	3	4	
楽天	.100	.097	.047	.039	121.1	142.0	26.3	23.1	80.3
ソニー	.233	.235	.205	.206	48.0	48.8	8.5	8.8	86.8
トヨタ	.0976	.0881	.161	.147	42.7	44.8	8.0	7.2	84.7
ソフトバンク	.190	.190	.137	.125	45.5	43.6	8.4	7.9	96.8
国際帝石	.0932	.0874	.268	.257	43.6	41.3	5.6	5.7	112
丸紅	.102	.0884	.125	.118	53.4	52.6	7.4	6.6	94.4
三井不動産	.0503	.0520	.0737	.0627	42.2	44.9	5.6	5.6	82.7

図 10: 7 銘柄それぞれのパラメータの推定値の平均

ヒートマップ

図 11 と図 12 は分枝行列 Θ^T をヒートマップとして表現したものであり、内部にはその数値がかかれている。縦軸が発生したイベントの系列、横軸がインパクトを受ける系列が対応しており、色が濃いほどインパクトが大きいことが分かる。

銘柄間の分析

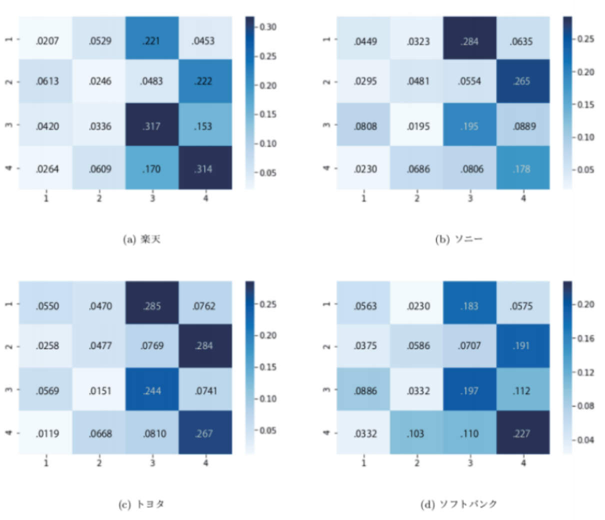


図 11: Θ^T の推定値のヒートマップ

銘柄間の分析

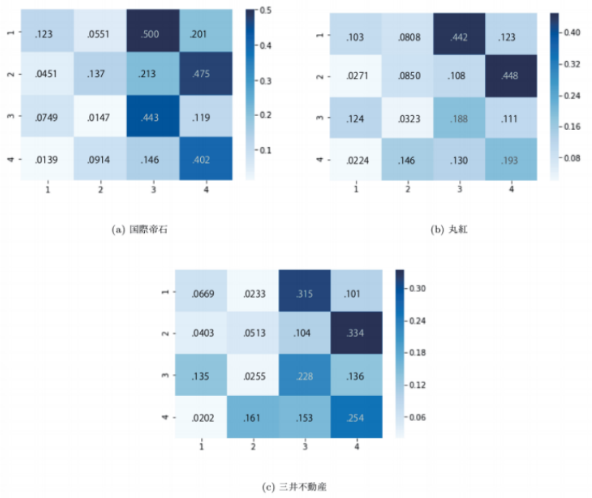


図 12: Θ^T の推定値のヒートマップ

銘柄間の分析

インパクト関数

マークと大きさが強度関数に与える影響に関して、銘柄間の比較分析を行う。本研究ではインパクト関数として (2) 式を利用している。よってインパクト関数はパラメータ p_i と共に、マーク χ の期待値 $mathbbE[\chi_j]$ にも依存する。表 4 は銘柄ごとに、系列別にパラメータ p_i の推定値をまとめた表である。

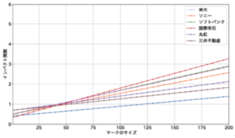
銘柄	系列ごとのp_jの平均値 1	2	3	4	系列ごとのp_jの中央値 1	2	3	4	$p_j=0$の割合(%) 1	2	3	4
楽天	1.20e-2	7.40e-3	6.85e-3	1.20e-2	1.04e-2	6.21e-3	6.59e-3	1.18e-2	0.0	0.0	0.0	0.0
ソニー	2.02e-2	2.92e-2	5.42e-3	2.81e-3	1.79e-2	2.17e-2	2.81e-10	8.95e-11	0.0	0.0	58.5	66.7
トヨタ	2.60e-2	2.41e-2	3.90e-3	2.29e-2	1.68e-2	1.62e-2	2.02e-4	2.09e-7	0.0	0.0	42.9	49.2
ソフトバンク	4.25e-2	3.60e-2	1.61e-2	2.25e-2	3.14e-2	2.25e-2	6.60e-3	1.39e-2	0.0	0.0	26.2	19.0
国際帝石	1.04e-2	1.03e-2	1.24e-2	1.17e-2	7.55e-2	7.33e-3	3.64e-3	6.71e-4	6.4	5.3	42.2	47.6
丸紅	8.12e-3	5.77e-3	6.02e-2	1.12e-1	6.34e-2	4.27e-3	2.80e-11	8.05e-3	1.4	2.7	63.0	31.5
三井不動産	2.31e-2	1.72e-2	2.29e-2	2.94e-2	1.38e-2	1.53e-2	1.53e-2	1.48e-2	0.0	0.0	29.2	29.2

図 13: インパクト関数 g_i のパラメータ p_i

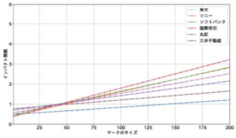
銘柄間の分析

インパクト関数

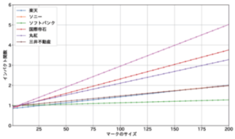
図 13 は系列 1 から系列 4 のそれぞれについて、横軸にマークの大きさ (注文サイズ)、縦軸にインパクト関数 g_i を描写したグラフである。



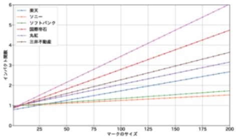
(a) 系列 1($d_i = 1$)



(b) 系列 2($d_i = 2$)



(c) 系列 3($d_i = 3$)



(d) 系列 4($d_i = 4$)

図 14: 系列 1 から系列 4 のインパクト関数

銘柄間の分析

インパクト関数

図 15 は東証一部で取引されている 7 銘柄の基本情報をまとめたものである。

銘柄名	2019 年 5 月の株価 始値	終値	リターン	出来高	呼値 (円)	呼値 1 単位のリターン(%)
楽天	1,241	1,235	−3.9	223,570,550	1	.038
ソニー	5,500	5,257	−2.0	135,357,900	1	.008
トヨタ	6,855	6,384	−3.1	98,152,400	1	.007
ソフトバンク	11,870	10,250	−6.4	132,442,000	1or5	.004 or .021
国際帝石	1,066.5	878.8	−8.4	136,364,600	0.1	.005
丸紅	794.1	683.4	−6.5	136,252,900	0.1	.006
三井不動産	2,594.5	2,631.5	0.6	65,147,200	0.5	.008

注：ソフトバンクの呼値は，期間中 1 万円を超える場合には 5 円，下回る場合には 1 円となっている．

図 15: 7 銘柄の株式の基本情報

おわりに

まとめ

- モデルの有効性を確認できた。
- 4 系列のイベントに関するマーク付き多次元 Hawkes 過程のパラメータをそれぞれ導出し、強度関数の観点から各銘柄の特性を議論した。
- 推定されたパラメータは、銘柄や推定区間によるが概ね安定的ということが確認できた。
- 約定というイベントに対する自己励起性は小さいのに対し、約定されない気配値の変化に対する自己励起性は相対的に大きい。
- 約定時のマークの大きさは強度関数に正のインパクトを与えることが明らかになった。
- 最良気配値の辺かよりは、約定イベントの方が相対的にインパクトの大きさが大きい。
- 最良気配値の変化イベントにいおいて注文量の大きさはほとんどインパクトを与えない。

おわりに

課題

- モデルの精緻化の可能性。
- 単純化をしていたので、より複雑な挙動のモデル化の可能性。
- 最適なインパクト関数の導出。
- 本研究で使用した銘柄以外にも当てはまることなのか。
- 約定時のマークの大きさは強度関数に正のインパクトを与えることが明らかになった。
- 最良気配値の辺かよりは、約定イベントの方が相対的にインパクトの大きさが大きい。
- 最良気配値の変化イベントにいおいて注文量の大きさはほとんどインパクトを与えない。