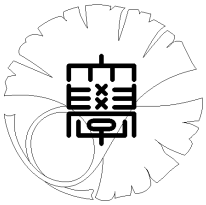
数理科学実践研究レター　2021–11 September 17, 2021

多次元 **Hawkes** 過程による性別に関連する不満投稿の解析

by

郷田 昌稔、矢野良輔、水野貴之



UNIVERSITY OF TOKYO

GRADUATE SCHOOL OF MATHEMATICAL SCIENCES

KOMABA, TOKYO, JAPAN

多次元 Hawkes 過程による性別に関連する不満投稿の解析

郷田昌稔[1]（東京大学大学院数理科学研究科）
Masatoshi Goda (Graduate School of Mathematical Sciences, The University of Tokyo)

矢野良輔（東京海上日動リスクコンサルティング）
Ryosuke Yano (Tokio Marine and Nichido Risk Consulting Co. Ltd.)

水野貴之（国立情報学研究所）
Takayuki Mizuno (National Institute of Informatics)

概 要

本研究はスパースな構造を持つ多変量 Hawkes 過程を用いて，幾つかのグループ間の Web サービスにおける投稿の伝搬をモデル化する手法を提案する．推定には疑似最尤推定量と L^1 罰則項付疑似最尤推定量を組み合わせた手法を用いる．実例として，日本の性別に関連する不満投稿に関してユーザーの年齢・性別で 12 のグループに分類し、Hawkes 過程を用いてモデリングを行なった．

1 はじめに

インターネット上には，SNS などの Web サービスを介して様々なテキストデータが投稿されている．これらの投稿は相互の閲覧などを通じて様々なカテゴリ間でトレンドを形成していると考えられる．本論文では、この伝播をモデル化する手法を提案する．

Hawkes 過程は強度過程が自身の確率積分で表される点過程であり，自己励起性を持つことから地震 [7] や板情報 [10, 1] のモデルなど様々な分野で利用されている．自己励起性はトレンドの指標と見なすことが出来る為，Web 投稿の解析にも当てはまりが良い．多変量 Hawkes 過程を用いて各グループの相互関係を調べる場合，重要でないパラメータを 0 と定めた疎なモデルを考慮することで，重要な相関関係のみを抽出することが出来る．このようなモデルを推定するために，疑似最尤推定量と L^1 罰則項付疑似最尤推定量を組み合わせた手法を提案する．

実例として，日本の性別に関連する不満投稿に関してユーザーの年齢・性別で 12 のグループに分類し分析を行なった．日本ではジェンダー格差を指摘する研究が数多くなされており，世代間での認識の差が指摘されている [9, 8]．また，2021 年 2 月には，東京 2020 オリンピック大会組織委員会の会長が女性に対する性差別的な発言をして辞任した．この辞任は国民の不満が広がった事によるものとされている[2]．このような背景から，女性の性別に関連する不満が多い事が予想され，世代間における不満の伝搬構造は一つの興味である．本研究の解析により，この各世代・性別間における不満の伝搬構造を定量化する事が出来る．さらに，各グループ内の投稿の伝播構造に関する推定結果を Hawkes グラフとカーネル関数の時間積分のヒートマップの形で可視化した．

2 データと手法

2.1 データ

本研究では株式会社 Insight Tech が提供する不満買取センター[3]と呼ばれる Web サービスによって収集された不満調査データセットを使用する [6]．このサービスでは，ユーザーはアプリケーションやウェブページを通じて日常の様々な体験に関する不満を投稿する事ができ、他のユーザーの投稿も閲覧可能である．各投稿には投稿時刻のラベルが付与されている為，データセットは連続時間観測の時系列データと見なす事が出来る．また，各投稿時刻におけるユーザーの年齢と性別も参照可能であり，これらを用いて表 1 に示すように 12 のカテゴリに分類した．

[1] goda@ms.u-tokyo.ac.jp
[2] https://edition.cnn.com/2021/02/11/sport/yoshiro-mori-resignation-intl-hnk/index.html
[3] https://fumankaitori.com

表 1: **カテゴリ**

性別 \ 年齢	0-19	20-29	30-39	40-49	50-59	60-
女性	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
男性	Group 7	Group 8	Group 9	Group 10	Group 11	Group 12

さらに，投稿内容を性別に関連する不満に限定する為に，文章の中に「男」と「女」が含まれている投稿を抽出した．表 2 にデータの例を示す.

表 2: **ランダム抽出したデータ例**

No.	不満内容	投稿時刻	性別	誕生年	Group
663	自分には甘いが相手のことになると相手を貶めるためだけに男尊女卑とか差別とか軽々しく使うのやめてほしい。コメンテーターレベルの議員は特に税金の無駄なんでやめてほしい。	2016-05-16 16:18:35	男	1977	9
4242	テレビでのオネェキャラへのいじりや流行り物のように男装女装を取り上げる限り日本のセクシャルマイノリティに対する理解は広まるどころか間違った方向に行くのではないかなと不安。服装、体、心、性対象を男性か女性かの二種類のどちらかに統一するものだという考えは間違えている。まずはメディアがきちんとした知識をつけて欲しい	2016-09-28 12:09:26	女	1996	2
4969	外国人選手へのインタビューを吹き替えする時に、女性選手の言葉の語尾を「〜わよ」「〜のよ」「〜わ」、男性選手の言葉の語尾を「〜なんだ」「〜だよ」などと訳すのをやめて欲しい。日本人が話しているような言葉遣いで表現してほしい。	2016-10-24 02:40:23	女	1988	2
6311	他の男に取られるくらいなら殺そうと思った 打倒だと思う 身勝手な女が多すぎる	2016-12-13 21:14:41	男	1973	10
6634	美容院の店員がチャライ。男性店員は女性客には本当にチャラく見える。	2016-12-27 11:56:22	男	1980	9
6890	女性ばっかりファッションのことをやるが... 男性もやって欲しい。	2017-01-08 06:40:16	男	1976	10
7808	性別の選択が男女しかなかった。LGBT からの不満は受け付けてくれないの?	2017-02-05 03:49:12	女	1992	2

2.2 モデル

カーネルが指数型の d 次元 Hawkes 過程とは強度過程が以下のように表される点過程 $N = (N^1, \ldots, N^d)$ である.

$$\lambda_t^i = \mu_i + \sum_{j=1}^{d} \int_{(0,t)} \alpha_{ij} e^{-\beta_{ij}(t-s)} dN_s^j,\ i = 1, \ldots, d.$$

ここで，μ_i, α_{ij} は非負値，β_{ij} は正値のパラメータである．確率積分の項により自己励起性が表現される．特にカーネルの時間に対する積分値 $\rho_{ij} = \frac{\alpha_{ij}}{\beta_{ij}}$ はイベント j が起きた際にイベント i がどの程度起こりやすくなるかを表す指標となる．この値をもとに有向グラフを描いたものは Hawkes グラフと呼ばれ [4]，カテゴリの相互関係が可視化できる.

2.3 推定方法

$(\Omega, \mathscr{F}, P)$ を確率空間，$\Theta \subset [0,\infty)^{d+d^2} \times (0,\infty)^{d^2}$ を相対コンパクトなパラメータ空間，$\theta^* = (\mu_1^*, \ldots, \mu_d^*, \alpha_{11}^*, \ldots, \alpha_{1d}^*, \alpha_{21}^*, \ldots, \alpha_{dd}^*, \beta_{11}^*, \ldots, \beta_{1d}^*, \beta_{21}^*, \ldots, \beta_{dd}^*) \in \Theta$ をパラメータの真値，N_t を強度過程が $\lambda_t^i = \mu_i^* + \sum_{j=1}^{d} \int_{(0,t)} \alpha_{ij}^* e^{-\beta_{ij}^*(t-s)} dN_s^j,\ i = 1, \ldots, d$ で表される d 次元 Hawkes 過程とする．この章における我々の目的は $\omega \in \Omega$ と $T > 0$ が与えられた際にデータ $\{N_t(\omega)\}_{t\in[0,T]}$ から θ^*

を推定する事である．強度過程をパラメータの関数として見たものを $\lambda_t^i(\theta)$ と書く．擬似対数尤度

$$l_T(\theta) = \sum_{i=1}^{d} \int_0^T \log\left(\epsilon_T^i + \lambda_s^i(\theta)\right) dN_s^i - \sum_{i=1}^{d} \int_0^T \left(\epsilon_T^i + \lambda_s^i(\theta)\right) ds$$

を最大にする θ を擬似最尤推定量（QMLE）と呼ぶ．また，目的関数

$$-l_T(\theta) + \gamma\sqrt{T}\left(\sum_{i=1}^{d} \mu_i + \sum_{i,j=1}^{d} \alpha_{ij}\right)$$

を最小にする θ をチューニングパラメータ γ の LASSO 推定量（L^1 罰則項付疑似最尤推定量）と呼ぶ．ここで，$\mu_i = 0$ のケースで擬似対数尤度を well-def にする為，$\epsilon_T^i > 0$ を加えている事に注意する．この時，以下の性質が数値実験により確認される．

- QMLE : $\mu_i^* = 0$ や $\alpha_{ij}^* = 0$ ならば，適当な確率で $\hat{\mu}_i = 0$ や $\hat{\alpha}_{ij} = 0$ と推定．非零のパラメータには漸近正規性が成立.
- LASSO: $\mu_i^* = 0$ や $\alpha_{ij}^* = 0$ ならば，QMLE より正確に $\hat{\mu}_i = 0$ や $\hat{\alpha}_{ij} = 0$ と推定．推定量にバイアスが生じる.

故に，LASSO で 0 のパラメータを推定し，正のパラメータを QMLE で再推定する 2 段階推定が有効であると分かる．

3 結果

$\epsilon_T^i = \frac{1}{T^2}, \gamma = 0.5$ として，QMLE と LASSO を組み合わせた手法による不満データの解析結果は以下の通りである．プログラムは Python3 で実装し[4]，最適化には L-BFGS-B 法を用いた．

$$\hat{\mu} = \big(0.0, 0.7442, 1.9413, 0.0403, 0.6293, 0.2421, 0.0, 0.0, 0.0, 0.2627, 0.0, 0.0\big),$$

$$\hat{\alpha} = \begin{pmatrix}
0.3889 & 0.0063 & 0.0469 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0097 & 0.0 & 0.0433 & 0.0 \\
0.0 & 0.3192 & 0.3922 & 0.1114 & 0.5059 & 0.0 & 0.0 & 0.0 & 0.1068 & 0.6006 & 0.0 & 0.0 \\
0.0 & 0.3046 & 1.5796 & 0.1686 & 0.5681 & 0.0 & 0.0 & 0.0 & 0.0 & 1.6673 & 0.0 & 0.0 \\
0.0 & 0.8914 & 1.0734 & 0.4289 & 0.3516 & 0.0 & 0.0 & 1.0213 & 0.4986 & 0.7692 & 0.0 & 0.0 \\
0.0 & 0.0343 & 0.1854 & 0.0 & 0.2593 & 0.0 & 0.0 & 0.199 & 0.0874 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0324 & 0.0013 & 0.0 & 0.0 & 0.258 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1346 & 0.0 \\
0.1334 & 0.0203 & 0.0024 & 0.0 & 0.1025 & 0.1335 & 0.0 & 0.0 & 0.0355 & 0.0 & 0.0153 & 0.0 \\
0.0 & 0.1078 & 0.0363 & 0.1403 & 0.1351 & 0.0 & 0.0 & 0.0 & 0.1118 & 0.2838 & 0.0 & 0.0 \\
0.0 & 0.2465 & 0.3657 & 0.3353 & 0.4448 & 0.0 & 0.0 & 0.0 & 0.2929 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0107 & 0.0512 & 0.0045 & 0.184 & 0.0 & 0.0 & 0.0 & 0.0812 & 0.2867 & 0.4019 & 0.0 \\
0.0 & 0.0712 & 0.0888 & 0.0032 & 0.1338 & 0.0 & 0.0 & 0.0 & 0.0559 & 0.0 & 0.5785 & 0.0 \\
0.0 & 0.0456 & 0.0065 & 0.0 & 0.0965 & 0.0979 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0269 & 0.0
\end{pmatrix},$$

$$\hat{\beta} = \begin{pmatrix}
1.8846 & 2.044 & 2.4282 & * & * & * & * & * & 2.0119 & * & 1.8354 & * \\
* & 1.5205 & 2.0965 & 1.8373 & 2.7367 & * & * & * & 1.9767 & 2.5105 & * & * \\
* & 1.5203 & 7.0284 & 0.3979 & 2.5592 & * & * & * & * & 2.2953 & * & * \\
* & 3.3362 & 5.3614 & 2.5917 & 2.328 & * & * & 2.096 & 2.2077 & 2.5117 & * & * \\
* & 1.7559 & 2.4978 & * & 2.6568 & * & * & 1.8125 & 1.5966 & * & * & * \\
* & 2.1235 & 2.0871 & * & * & 2.4063 & * & * & * & * & 2.0047 & * \\
2.2488 & 2.0805 & 2.1363 & * & 2.3354 & 2.4174 & * & * & 2.3423 & * & 2.2609 & * \\
* & 2.2988 & 2.4466 & 2.1271 & 1.8205 & * & * & * & 2.1045 & 1.929 & * & * \\
* & 2.6917 & 3.9023 & 2.7543 & 2.2145 & * & * & * & 2.4289 & * & * & * \\
* & 2.026 & 1.4952 & 1.9325 & 2.2706 & * & * & * & 2.0143 & 1.8652 & 2.7102 & * \\
* & 1.3024 & 1.9844 & 2.0162 & 2.327 & * & * & * & 2.1765 & * & 2.8969 & * \\
* & 2.0452 & 2.2732 & * & 2.3534 & 2.4012 & * & * & * & * & 2.1106 & *
\end{pmatrix},$$

[4] https://github.com/goda235/MHP_QMLE_LASSO

$$\hat{\rho} = \begin{pmatrix}
0.2064 & 0.0031 & 0.0193 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0048 & 0.0 & 0.0236 & 0.0 \\
0.0 & 0.2099 & 0.1871 & 0.0606 & 0.1848 & 0.0 & 0.0 & 0.0 & 0.054 & 0.2392 & 0.0 & 0.0 \\
0.0 & 0.2004 & 0.2248 & 0.4237 & 0.222 & 0.0 & 0.0 & 0.0 & 0.0 & 0.7264 & 0.0 & 0.0 \\
0.0 & 0.2672 & 0.2002 & 0.1655 & 0.151 & 0.0 & 0.0 & 0.4873 & 0.2259 & 0.3062 & 0.0 & 0.0 \\
0.0 & 0.0195 & 0.0742 & 0.0 & 0.0976 & 0.0 & 0.0 & 0.1098 & 0.0548 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0152 & 0.0006 & 0.0 & 0.0 & 0.1072 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0672 & 0.0 \\
0.0593 & 0.0098 & 0.0011 & 0.0 & 0.0439 & 0.0552 & 0.0 & 0.0 & 0.0151 & 0.0 & 0.0068 & 0.0 \\
0.0 & 0.0469 & 0.0149 & 0.066 & 0.0742 & 0.0 & 0.0 & 0.0 & 0.0531 & 0.1471 & 0.0 & 0.0 \\
0.0 & 0.0916 & 0.0937 & 0.1217 & 0.2009 & 0.0 & 0.0 & 0.0 & 0.1206 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0053 & 0.0343 & 0.0023 & 0.081 & 0.0 & 0.0 & 0.0 & 0.0403 & 0.1537 & 0.1483 & 0.0 \\
0.0 & 0.0547 & 0.0447 & 0.0016 & 0.0575 & 0.0 & 0.0 & 0.0 & 0.0257 & 0.0 & 0.1997 & 0.0 \\
0.0 & 0.0223 & 0.0029 & 0.0 & 0.041 & 0.0408 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0127 & 0.0
\end{pmatrix}.$$

図 1 はノードを Poisson パラメータ μ_i で，エッジを相関パラメータ ρ_{ij} で重みづけた Hawkes グラフである．さらに，図 2 は $\hat{\rho}$ のヒートマップである．

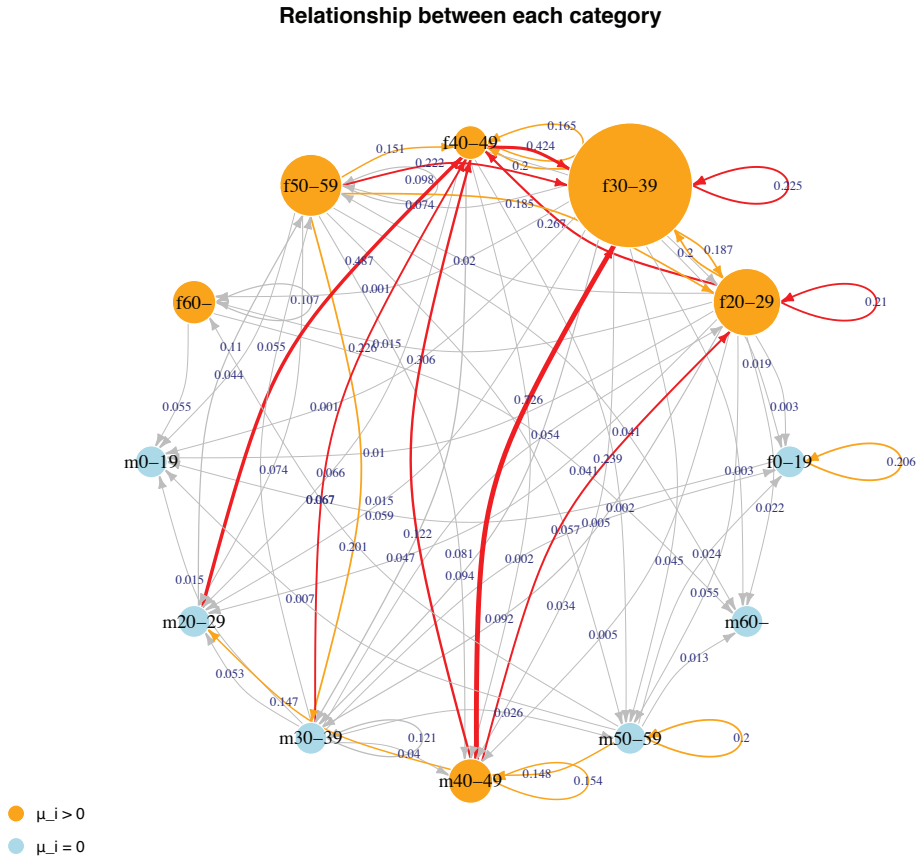


図 1: **Hawkes グラフ．** $\hat{\rho}$ の値が大きな上位 10 のエッジを赤色に，次点で値が大きな上位 10 のエッジを橙色で表示している．

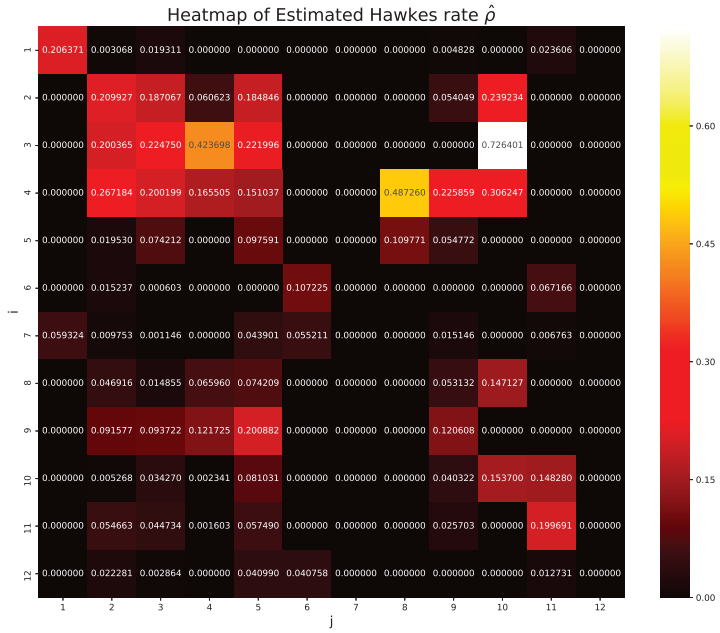


図 2: $\hat{\rho}$ **のヒートマップ．** $\hat{\rho}_{ij}$ はグループ j の投稿がグループ i の投稿に与える影響の大きさを表す．

4 結論

上述の結果から以下の構造が読み取れる．先ず，Poisson 的な不満の発生は 20〜30 代女性が多い．また，30〜40 代女性が 20〜40 代（特に 40 代）男性から受ける影響が特筆して大きい．一方で 40〜50 代男性は自身以外からの影響が小さい．最後に男女ともに 10 代の投稿数は少ない事が分かる．

モデルの妥当性を自然言語処理の観点から確認する．相互作用があるグループ間では投稿内容に類似性が見られると考えられる．Word2Vec[5] 及び BERT[3] を用いて投稿内容をベクトル化し，グループ間でコサイン類似度を計算したものと多次元 Hawkes 過程により推定された $\hat{\rho}$ の間でスピアマンの順位相関係数を計算する．結果は表 3 のようになり，どちらの手法でもスピアマン順位相関係数は 0.4 前後であるが，それは十分に大きい値である．実際に相関が 0 を帰無仮説とする検定を行なった場合の p 値は極めて小さく，順位相関が有意である事が確認できる．これらの結果は Hawkes 過程がトレンドの伝搬を記述できていることを裏付けている．

表 3: **グループ間のコサイン類似度と** $\hat{\rho}$ **のスピアマンの順位相関係数．**

	スピアマン順位相関係数	p 値
Word2Vec	0.3750	3.632×10^{-6}
BERT	0.4277	8.903×10^{-8}

5 終わりに

日本の Web サービスにおける性別に関連する不満の世代・性別間の伝搬構造を多変量 Hawkes 過程を用いて定量化した．推定方法としては，LASSO と QMLE を組み合わせた手法を提案した．この手法に関しては，数値実験によりスパースな構造を持つ多変量 Hawkes 過程のパラメータを精度よく推定する事が示唆された．最後に、Hawkes グラフとカーネル関数の時間積分値のヒートマップにより、各グループ間の（自己）相関を可視化した。

推定結果から，20〜39 歳の女性の投稿は Poisson 的な投稿の確率が高く、自身や 20〜49 歳の男性の投稿の影響を強く受けていることが分かった．対照的に，40〜59 歳の男性グループは他のグループからの影響が小さく，自己励起の度合いが高いことが分かった．

6 謝辞

本研究では，国立情報学研究所の IDR データセット提供サービスにより株式会社 Insight Tech から提供を受けた「不満調査データセット」を利用した．また，本研究の会合の調整や議論の総括をして頂いた柏原崇人先生に深く感謝申し上げます．

A 付録

節 2.3 で紹介した QMLE は全てのパラメータが正であり $\epsilon_T^i = 0$ とした場合にその漸近正規性が証明されている [2]．本論文における QMLE も非零の真値を持つパラメータに関する成分に関しては漸近正規性が成り立つことが確認できる．すなわち，QMLE を $\hat{\theta}_T$ とした時，その誤差 $\sqrt{T}(\hat{\theta}_T - \theta^*)$ の非零の真値を持つパラメータに関する成分が，$T \to \infty$ とした時に正規分布に分布収束する．従って，LASSO でパラメータ選択を行なった後に QMLE で再チューニングを行うと，最終的な非零のパラメータの推定値には漸近正規性が従うことが予想される．例として，以下のようにパラメータを設定したカーネルが指数型の 3 次元 Hawkes 過程に関する数値実験を見る．

$$\mu = (0.1, 0.0, 0.1), \quad \alpha = \begin{pmatrix} 0.0 & 0.2 & 0.2 \\ 0.4 & 0.2 & 0.0 \\ 0.0 & 0.0 & 0.3 \end{pmatrix}, \quad \beta = \begin{pmatrix} * & 0.4 & 0.4 \\ 0.7 & 0.4 & * \\ * & * & 0.5 \end{pmatrix}.$$

ここで，$*$ は真値を持たない未定義のパラメータである．観測時間を $T = 1000$，モンテカルロシミュレーションの回数を 300 回と設定する．図 3 は上記のモデルに対して計算された LASSO と QMLE を組み合わせた手法による推定値 $\check{\theta}_T$ の誤差 $\sqrt{T}(\check{\theta}_T - \theta^*)$ のヒストグラムである．

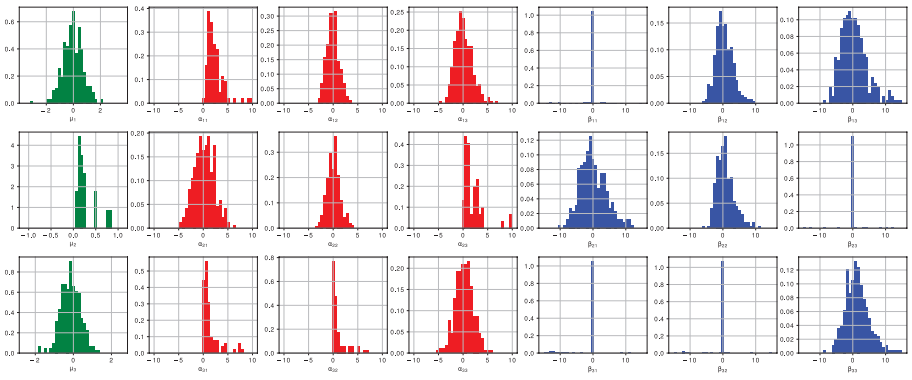


図 3: $\sqrt{T}(\check{\theta}_T - \theta^*)$ **のヒストグラム．**

非零のパラメータに関する推定値のヒストグラムは正規分布に近い形となっている事が確認できる．また，QMLE と LASSO のそれぞれで各パラメータを何回 0 と推定したかを表 4 に示す．

表 4: **各推定量が** 0 **と推定した回数．**試行回数は 300 回.

QMLE						**LASSO**					
μ_1	0	μ_2	**270**	μ_3	0	μ_1	0	μ_2	**282**	μ_3	1
α_{11}	**176**	α_{21}	0	α_{31}	**189**	α_{11}	**219**	α_{21}	0	α_{31}	**204**
α_{12}	0	α_{22}	0	α_{32}	**189**	α_{12}	0	α_{22}	0	α_{32}	**207**
α_{13}	0	α_{23}	**227**	α_{33}	0	α_{13}	0	α_{23}	**244**	α_{33}	0

全体として LASSO の方がより正確に 0 のパラメータを正しく推定していることが分かる．これら性質の数学的に厳密な証明は本論文の主旨から逸脱するため，今後の課題とする．

参考文献

[1] Frédéric Abergel, Marouane Anane, Anirban Chakraborti, Aymen Jedidi, and Ioane Muni Toke. *Limit Order Books*. Cambridge University Press, 1st edition, 2016. ISBN: 978-1-107-16398-0.

[2] Simon Clinet and Nakahiro Yoshida. Statistical inference for ergodic point processes and application to limit order book. *Stochastic Processes and their Applications*, 127(6):1800–1839, 2017.

[3] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186, 2019.

[4] Paul Embrechts and Matthias Kirchner. Hawkes graphs. *Theory of Probability and Its Applications*, 62(1):163–193, 2018.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2, pages 3111–3119, 2013.

[6] Kensuke Mitsuzawa, Maito Tauchi, Mathieu Domoulin, Masanori Nakashima, and Tomoya Mizumoto. Fkc corpus: a japanese corpus from new opinion survey service. In *Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results*, pages 11–18, 2016.

[7] Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

[8] Misaki Takeya Toshiaki Tanaka, Toshihiro Sadasue. Intergenerational disparity on knowledge and understanding of lgbt. *Papers of Kyushu Women's University*, 54(2):115–127, 2017.

[9] Seisuke Tsuda. Factors that affect gender role attitudes: Focus on effects of interaction between sex, academic background and generation. *Journal of educational research for human coexistence*, 6:87–100, 2019.

[10] Hai Chuan Xu and Wei Xing Zhou. Modeling aggressive market order placements with hawkes factor models. *PLOS ONE*, 15(1), 2020.