Applications of Data Science in Social Science
in honor of Prof. Nobuhiko Terui
supported by JSPS KAKENHI

Program and Abstracts

Feb. 17-18th, 2022

Graduate School of Economics and Management

Tohoku University

financially supported by

Grant-in-Aid for Scientific Research

(A) 20H00576, (B) 21H03400

Feb 17, 2022

9:30-11:00 【Session 1】Chair: **Masataka Ban** (University of Tsukuba)

・**Sotaro Katsumata** (Osaka University)

Online Review Text Analysis Incorporating Valence and Vote: Extending the Topic Model to Include Performance Measures

・**Mirai Igarashi** (University of Tsukuba)

Finding Influential Users by Topic in Unstructured User-generated Content

・**Hiroshi Onishi** (Tokyo University of Science)

Does Deregulation of Health Claims on Functional Food Package Design Change Consumers' Purchase Behaviors?

11:10-12:10 【Session 2】Chair: **Shohei Hasegawa** (Hosei University)

・**Junji Nakano** (Chuo University)

Statistical model for article citation network in Web of Science

・**Manabu Asai** (Soka University)

High-Dimensional VAR Models via Sparse Precision Matrix

Lunch 12:10-13:10

13:10-14:10 【Session 3】Chair: **Wirawan Dony Dahana**（Osaka University）

・**Naoto Kunitomo** (The Institute of Statistical Mathematics)

Local SIML Estimation of Some Brownian and Jump Functionals under Market Microstructure Noise

・**Nobuaki Hoshino** (Kanazawa University)

ビッグミクロデータのプライバシー保護

14:20-15:20【Session 4】Chair: **Tsukasa Ishigaki** (Tohoku University)

・**Takemi Yanagimoto** (The Institute of Statistical Mathematics)

無情報事前分布の理解：公平な分配と未知母数から

・**Hiroshi Itosu** (Tokyo University of Science)

パラメータ識別可能なグラフィカルモデルによる欠測を含んだ2×2分割表の対称性の検定

15:30-16:30 【Session 5】Chair: **Yasumasa Matsuda** (Tohoku University)

・**Makoto Abe** (The University of Tokyo)

Deriving Lifetime Value of a Customer Who Exhibits Non-Poisson Purchase Behavior:

Application to Marketing Intervention for Customer Retention
・**Thomas Otter** (Goethe University)
Discrete Choice in Marketing through the Lens of Rational Inattention

16:40-17:40 【Session 6】Chair: **Wirawan Dony Dahana**（Osaka University）
・**Nobuhiko Terui** (Tohoku University)
Service Economy and Data Science

Feb 18, 2022

9:30-11:00【Session 7】Chair: **Shohei Hasegawa** (Hosei University)
・**Greg Allenby** (Ohio State University)
A Choice Model of Utility Maximization and Regret Minimization
・**Jaehwan Kim** (Korea University)
The Impact of Gig Economy on the Product Quality through the Labor Market: Evidence from Ride-sharing and Restaurant Quality
・**Takuya Satomura** (Keio University)
A Multiple Duration Choice Model for Service Data

11:10-12:10 【Session 8】Chair: **Masataka Ban** (University of Tsukuba)
・**Yoshimasa Uematsu** (Tohoku University)
Robust False Discovery Rate Control via Debiased Rank Lasso
・**Junichi Hirukawa** (Niigata University)
Innovation algorithm of fractionally integrated ($I(d)$) process and applications on the estimation of parameters

Lunch 12:10-13:10

13:10-14:10 【Session 9】Chair: **Tsukasa Ishigaki** (Tohoku University)
・**Qingfeng Liu** (Otaru University of Commerce)
Machine Collaboration
・**Takaki Sato** (Tohoku University)
コロナ禍における感染拡大防止対策の人流抑制効果に関する実証分析

14:20-15:20 【Session 10】Chair: **Yasumasa Matsuda** (Tohoku University)

・**Masataka Ban** (University of Tsukuba)
購入型クラウドファンディングの資金調達パターン分析
・**Shohei Hasegawa** (Hosei University)
BTYD モデルによる Q&A サイトの投稿行動の分析

15:30-16:30 【Session 11】Chair: **Wirawan Dony Dahana**（Osaka University）
・**Li Yinxing** (Tohoku University)
Product Embedding For The Large-scale Disaggregated Sales Data
・**Shogo Takedomi** (Tohoku University, BRIDGESTONE)
待ち時間予測モデルを利用した pMP モデリングによる緊急ロードサービス施設配置最適化

大阪大学 OSAKA UNIVERSITY ROBERT H. SMITH SCHOOL OF BUSINESS

Online Review Text Analysis Incorporating Valence and Vote: Extending the Topic Model to Include Performance Measures

Applications of Data Science in Social Science
Symposium in honor of Prof. Nobuhiko Terui, supported by JSPS KAKENHI
@Tohoku University
Feb. 17, 2022

Sotaro Katsumata
Graduate School of Economics, Osaka University
katsumata@econ.osaka-u.ac.jp

P. K. Kannan
Robert H. Smith School of Business, University of Maryland
pkannan@umd.edu

Introduction

- CGM (consumer generated media) influence firms' financial performance
  - e.g., Chevalier and Mayzlin (2006), Dellarocas, Zhang, Awad (2007), Liu (2006)
- GGMs for consumers and firms
  - Spread their product information without additional financial costs (Tirunillai and Tellis, 2014)
  - Effective source of information for firms to obtain external evaluations of their products (Dellarocas 2003).
- Consumer reviews are written in natural language
  - Some analytical frameworks are presented recently (Berger et al. 2020; Humphreys and Wang 2017; Balducci and Marinova 2018)
- It is necessary to understand some other characteristics of consumer reviews
  - Polarity and Importance

Polarity (Valence) and Importance (Vote)

- **"Polarity" of online review**
  - Some reviews have a negative impact on firm performances
    - Volume, Variance, and Valence need to be considered (Dellarocas et al. 2003)
  - Positive reviews have a positive impact on market outcomes and *vice versa*
    - Chevalier and Mayzlin (2006), Liu (2006), Moe and Trusov (2011), Moe and Schweidel (2012)
  - Review polarity depends on the products
    - Burns and Hou, (2017), Schoenmueller, Netzer, and Stahl (2020)
- **"Importance" of online review**
  - Viewers can vote and see how many votes have been cast
    - Mudambi and Schuff (2010), Lu, Wu, and Tseng (2018)
  - Viewers prefer posts attract many votes
    - Egebark and Ekstrom (2018)
  - Associated with actual corporate financial performance
    - Ding, et al. (2017)
- Necessary to consider both valence and vote together
  - Interactivity (sociality) of UGC as one of the open issues for digital marketing (Kannan and Li, 2017)
  - Several studies have presented models to analyze valence and text simultaneously (e.g., Büschken and Allenby 2016, 2020)

Collective Evaluation for CGM

- **Valence** reflects the sentiment of reviews
  - (e.g., Blei and McAuliffe 2010; Tirunillai and Tellis 2014)
  - Difficult to accurately determine the impact on financial outcomes (e.g., Chevalier & Mayzlin 2006; 2006; Moe and Trusov 2011).
- **Vote** is an indicator to examine the quality of reviews
  - (Mudambi and Schuff 2010).
  - This "double voting" mechanism: a fair indicator
  - (Mayzlin, Dover, and Chevalier 2014).
  - Collect vote from others also increases trust
  - (Egebark and Ekstrom 2018; Ding, et al. 2017).
- Impact of review on the market is different
  - Necessary to separate important topics from less important ones (Mimno et al. 2011)
  - Incorporating this valence/vote in the topic model, we can select important topics

Model

- Three Components of the model

1. **Text**
2. **Valence**, the rating score
   - used as useful information to examine market trends even before the development of text analysis methods (e.g., Ansari, Essagaier and Kohli 2000).
3. **Vote**, the number of "helpful" votes
   - The reviews by viewers (Mudambi and Schuff 2010)

- Index
  - N: The total number of tokens (individual words) observed in all documents
  - D: the total number of documents (reviews)
  - K: the number of potential topics needs to be determined by the analyst

Review Text

- w_i: i-th observed token (word)
  - $w_{iv} = \begin{cases} 1, & \text{if the word } v \text{ is observed in token } i \\ 0, & \text{else} \end{cases}$

$$w_i \sim Categorical_V(\tilde{\psi}_i)$$
$$\tilde{\psi}_{iv} = \prod_{k=1}^{K} \psi_{kv}^{z_{ik}},$$

- z_i: K-dimensional parameter
  - $z_{ik} = \begin{cases} 1, & \text{if token } i \text{ belongs to topic } k \\ 0, & \text{else} \end{cases}$
- ψ_{kv}: a parameter which obtained for each topic.

Writers' Valence

- y_{dj}: Rating of perspective j of review d
  - if the data is collected on a 5-point scale in item j, define $Q_j = 4$ and $y_{dj} \in \{0,1,2,3,4\}$

$$y_{dj} \sim Binomial(Q_j, \tilde{\phi}_{dj}),$$
$$\tilde{\phi}_{dj} = \prod_{k=1}^{K} \phi_{kj}^{c_{dk}},$$

- c_d: the K-dimensional parameter
  - $c_{dk} = \begin{cases} 1, & \text{if document } d \text{ belongs to topic } k \\ 0, & \text{else} \end{cases}$
- ϕ_{kj}: parameter that influences the height of item j's score obtained for each topic.
- Expected value of the score of item j of the review belonging to topic k is $Q_j \phi_{kj}$.

Viewers' Vote

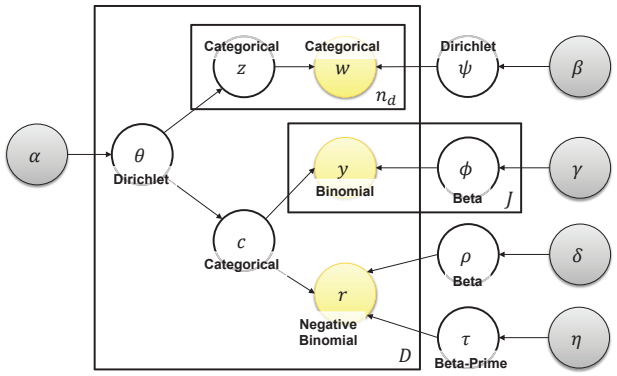
- r_d: The number of "helpful" ratings by viewers
  - a natural number greater than or equal to zero
  - assume a negative-binomial distribution $NegBin(\cdot)$.

$$r_d \sim NegBin(\tilde{\tau}_d, \tilde{\rho}_d)$$

$$\tilde{\rho}_d = \prod_{k=1}^{K} \rho_k^{c_{dk}}, \tilde{\tau}_d = \prod_{k=1}^{K} \tau_k^{c_{dk}}$$

- c_{dk}: a parameter that indicates the topic assignment of document d,
- Note: The negative binomial distribution of $\tau_k = 1$ is equivalent to the geometric distribution.

Model (Directed Acyclic Graph)



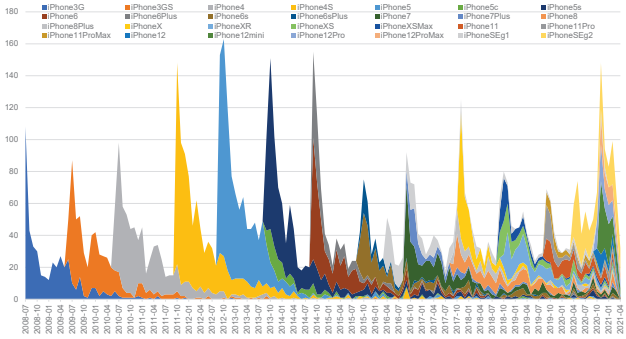
Data Description

- Data: collected from Kakaku.com (https://kakaku.com/)
- A consumer review site mainly written in Japanese with a long history
  - The website launched in 2000
  - The **global No. 1** in the "E-commerce and Shopping > price comparison" category in September 2021 (Similarweb, https://www.similarweb.com/),
  - The ranking of all websites, ranks **328th in the world**, and **33rd in Japan.**
  - Tanked **441st in the world** and **36th in Japan** (Alexa ranking, https://www.alexa.com/).

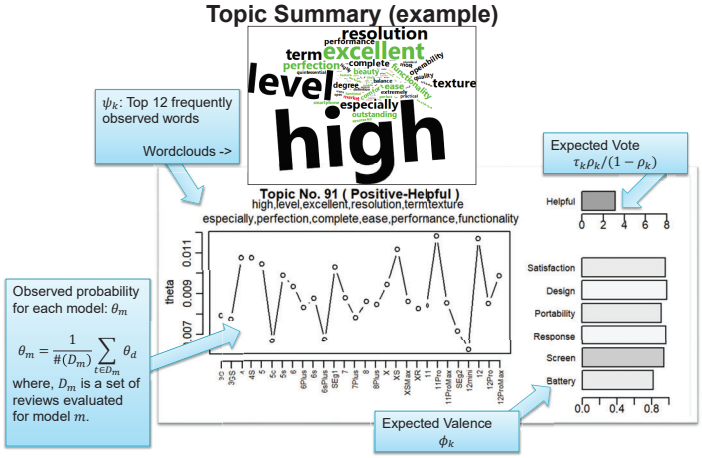
- Reliability of the Source
  - A market survey has shown that online review platforms have fewer manipulated posts than shopping platforms
  - Trustworthiness of reviews is one of the major issues, such as fake reviews and manipulated reviews
    - (e.g., Chen, Guo and Huang 2021; Mayzlin, Dover, and Chevalier 2014)

Review Texts

Apple iPhone Reviews ($D_{iPhone} = 6922$)



Topic Summary (example)



Regression for Salient Topics

- Examine the factors affecting θ_d
- Problems in assuming
  - Constraint and Efficiency of Estimation
    - θ_{dk} is the probability of topic affiliation, so there is a constraint that $\sum_{k=1}^{K} \theta_{dk} = 1$.
      - Mimno and McCallum (2008), Blei and Lafferty (2006) Büschken and Allenby (2016), Liu and Toubia (2018), Toubia, et al. (2019)
  - Compatibility
    - Cannot not determine in advance which words will be classified into which topics
  - Quality of topics
    - A certain percentage of topics are considered poor and have low informational value (Mimno, 2011)

- Focus only on the **helpful Salient Topics**
  - Based on the aggregate demand model (Berry 1994)
  - S: set of helpful salient topics,
  - S^c: set of other topics,
  - $\theta_{d0} = \sum_{k \in S^c} \theta_{dk}$: the base topics
  - $\log(\theta_{dk}) - \log(\theta_{d0})$: Dependent variable

Regression for Salient Topics

- Explanatory variables
  - *LatestModel*: 1 if the model to be reviewed is the latest and 0 otherwise, the
  - *Writers'Experience*: the number of submissions within the past three years by review writers (add 1 and take the logarithm)
  - *LatestModel*: 1 if the review is about a new model
  - *#ofWords, #ofWords*^{2} : the number of words in the review, curve linearly related to helpfulness for readability (Agnihotri and Bhattacharya 2016)
  - In addition to these variables, incorporate the dummy variables of the models (products) as a random effects
- Model:

$$\log(\theta_{dk}) - \log(\theta_{d0}) = \lambda_{00k} + \sum_{m=1}^{M} ModelDummy_{dm}\,\lambda_{2mk} + LatestModel_d \lambda_{3k} + Writers'Experience_d \lambda_{4k} + \#ofWords_d \lambda_{4k} + \#ofWords_d^2 \lambda_{5k} + \varepsilon_{dk}, k \in S$$

Result of Regression: Apple iPhone Series

topic No.	iPhone series: Frequent words	LatestModel	Writers' Experience	#ofWords	#ofWords^2
Negative-Helpful					
92	bad, nothing, say, anything, should, wrong, opinion, useless, special, else, particularly, think(-1)	-0.001 (-0.076)	**-0.048 *** **(-2.016)**	0.015 (0.183)	**-0.030 **** **(-3.061)**
53	apple, support, replacement, store, tell, care, break, say, ask, repair, replace, again	**0.028 ***** **(4.086)**	**-0.042 †** **(-1.781)**	0.057 (0.687)	**-0.042 ***** **(-4.210)**
29	sometimes, start, end, moment, freeze, seem, then, happen, issue, frequently, fix, stop	**0.022 **** **(3.206)**	**-0.049 *** **(-2.069)**	-0.114 (-1.365)	-0.016 (-1.599)
57	performance, price, late, series, cheap, camera, high, enough, spec, expensive, low, size	**0.040 ***** **(5.872)**	**0.227 ***** **(9.624)**	**-0.182 *** **(-2.180)**	-0.009 (-0.960)
18	camera, image, mode, lens, wide, night, angle, video, digital, shoot, even, can	-0.007 (-1.016)	**0.040 †** **(1.695)**	**-0.673 ***** **(-8.058)**	**0.067 ***** **(6.748)**
Positive-Helpful					
8	3gs, shape, much, become, well, folder, fast, design, ios5, make, flat, ios4	**-0.060 ***** **(-8.738)**	**0.061 *** **(2.562)**	**0.175 *** **(2.090)**	**-0.073 ***** **(-7.412)**
91	high, level, excellent, resolution, term, texture, especially, perfection, complete, ease, performance, functionality	0.001 (0.215)	0.037 (1.549)	**-0.454 ***** **(-5.434)**	**0.030 **** **(3.036)**
72	switch, docomo, decide, mnp, get, lose, worry, decision, glad, end, family, factor	0.005 (0.800)	**-0.084 ***** **(-3.554)**	**0.396 ***** **(4.743)**	**-0.091 ***** **(-9.174)**
77	touch, ipod, use, panel, compare, convenient, own, nano, 4th, dock, 5th, intuitive	-0.011 (-1.579)	**-0.046 †** **(-1.955)**	**-0.288 ***** **(-3.443)**	0.008 (0.797)
95	can, apps, app, also, lot, own, many, fun, enjoy, useful, make, customize	**0.016 *** **(2.309)**	**-0.081 ***** **(-3.413)**	**0.166 *** **(1.987)**	**-0.045 ***** **(-4.558)**

Note)
The model incorporating the intercept, and model (device) random effects.
†: p<10%, *: p<5%, **: p<1%, ***: p<0.1%.
Model summary statistics: R^2 = 0.170, Adj. R^2 = 0.166, and F-value = 44.15 (p<0.001).

Finding Influential Users by Topic in Unstructured User-generated Content

Mirai Igarashi* Kunpeng Zhang† P.K. Kannan‡ Nobuhiko Terui§

February 17, 2022

1 Introduction

This study addresses two critical limitations in the extant research on social influence, which is the concept that consumers effect each other's behavior. First, despite a vast amount of literature on social influence, few studies have developed models for unstructured user-generated content (UGC), such as text and images, influenced by others' contents. Most of the research has examined the structure of consumer behavior influenced by social interaction in terms of the numerical aspects of consumer behavior, such as the volume or binary of the focal behavior. However, since people get a lot of information from nonnumerical content in social media, analyzing how the generated unstructured content influences on other users and how the generated content changes while being influenced by other users are important research topics. Second, in this study, social influence is estimated heterogeneously for each edge and topic of content. Previous studies have examined various moderators that could change the level of influence such as types of connections (Iyengar et al., 2011) and consumer attributes (Wang et al., 2013). However, assuming different social influence for each edge is the strongest assumption and includes all the possible moderators in the previous studies. In addition, we consider that the influence on the behavior generating content varies with the topic of the content, just as the influence on purchasing products varies with the product characteristics (Park et al., 2018). These assumptions allow us to determine who has a strong influence on the network at the level of edges and by topic.

*Faculty of Business Science, University of Tsukuba
†Decision, Operations and Information Technologies, Robert H. Smith School of Business, University of Maryland
‡Marketing, Robert H. Smith School of Business, University of Maryland
§Graduate School of Economics and Management, Tohoku University

2 Model

The observed UGC data is assumed to be decomposed into multisets of the smallest unit constructs (e.g., words in a text and objects in an image) disregarding the order of the unit but keeping multiplicity. In the following, we denote the UGC data as $W = \{w_{ut}\}$, where u and t refer user and time points, respectively, and $w_{ut} = (w_{ut1}, \ldots, w_{utN_{ut}})^\top$ (N_{ut} denotes the number of constructs). We also observe the following relationships among social media users, and the set of users that user u follows is represented by $\mathcal{F}_u$.

First, we introduce the generative model for the elements of UGC following the conventional topic model. The proportion of topics included in the contents generated by the user u at time t is represented by a topic distribution, θ_{ut}. The latent topic behind w_{uti} is thought to be determined from a categorical distribution, $z_{uti} \sim \mathit{categorical}(\theta_{ut})$. When the topic assignment is given, the corresponding w_{uti} is also generated from a categorical distribution, $w_{uti} \mid z_{uti} = k \sim \mathit{categorical}(\phi_k)$. ϕ_k is a element distribution, which represents the generative probability of elements in the context of topic k.

Next, we incorporate the dynamic changes in users' interests and the influence of friend-generated contents on these interest levels. The topic distribution, θ_{ut}, has the following hierarchical structure.

$$\theta_{utk} = \frac{\exp(\eta_{utk})}{\sum_{k'} \exp(\eta_{utk'})}, \qquad \eta_{utk} \sim N(\lambda_{utk}, 1) \tag{1}$$

$$\lambda_{utk} = \alpha_k \cdot \eta_{ut-1k} + \sum_{f \in \mathcal{F}_u} \beta_{ufk} \cdot \eta_{ft-1k} + \gamma_{tk} + \delta_{uk} \tag{2}$$

β_{ufk} represents the lagged social influence of friend f whom user u follows on the interest of user u in topic k, and it varies depending on the topic as well as the network edge. γ_{tk} and δ_{uk} are time-topic and user-topic fixed effects controlling for common shocks at a specific time and each user's intrinsic preference for a specific topic, respectively.

References

Iyengar, R., Van den Bulte, C., and Valente, T. W. Opinion leadership and social contagion in new product diffusion. *Marketing Science*, 30(2):195–212, 2011.

Park, E., Rishika, R., Janakiraman, R., Houston, M. B., and Yoo, B. Social dollars in online communities: The effect of product, user, and network characteristics. *Journal of Marketing*, 82(1):93–114, 2018.

Wang, J., Aribarg, A., and Atchadé, Y. F. Modeling choice interdependence in a social network. *Marketing Science*, 32(6):977–997, 2013.

Does Deregulation of Health Claims on Functional Food Package Design Change Consumers' Purchase Behaviors?

Hiroshi Onishi[1], Tokyo University of Science

Masakazu Ishihara, New York University

Abstract

Japanese government deregulated health claim legislation for functional food products in April 2015. A new healthy food category, called "Kinousei (Functional-food)" brands, was introduced in addition to the incumbent category of "Tokuho (Special-health)" brands which requires more rigid scientific evidences of health effects for an official approval. This research investigates the effects of the health claim deregulation for the functional food products to consumers' purchase behaviors, for example, whether people have bought more volumes and switched to the healthy claimed products, or they have not changed their purchase behaviors, and which types of consumers are vulnerable to the health claims of food products. In addition, this study also examines what kinds of product package designs are perceived as healthy image by consumers.

We use a purchase panel dataset of functional yogurt products in Japanese market to investigate our research questions. We apply the deep learning framework of the BrandImageNet model (Liu et al. 2020) in order to evaluate the health image of product packages among more than one thousand of the yogurt items. Then, we estimate the difference-in-difference regression with cluster-robust inference by Cameron & Miller (2010) to investigate the effect of the health claim deregulation for the Kinousei brands and the consumers' healthy image perception to the product package designs on consumers'

[1] All correspondence may be addressed to the first author at hohnishi@rs.tus.ac.jp.

functional yogurt purchase behaviors by comparing the Tokuho brands as a control group between pre and post periods when the deregulation was in force in Japanese market.

We find the positive impacts of the health claim deregulation and healthy package design images among only consumers whose wellness consciousness is low, which results in switching their purchase products from normal brands to health claimed the Kinousei brands. Whereas, consumers with high wellness consciousness have not changed their purchase behaviors to keep buying the Tokuho brands. We then also discuss managerial implications for functional food manufacturers.

Keywords: *Health Claims; Package Design; Functional Foods; Deep Learning; Difference-in-Difference (DiD); Error Correction Models*

Statistical model for article citation network in Web of Science

Yuichiro Yasui
The Graduate University for Advanced Studies, SOKENDAI, Tokyo, Japan

Junji Nakano
Chuo University, Tokyo, Japan

Abstract

We propose a stochastic generative model to represent a directed graph constructed by citations among academic papers, where nodes and directed edges represent papers with discrete publication time and citations respectively. The proposed model assumes that a citation between two papers occurs with a probability based on the type of the paper, importance of cited paper, and difference between their publication times. We consider that the importance and type of a paper are approximated by the in- and out-degrees of the node. In our model, we adopt three functions: a logistic function for illustrating the numbers of papers published in discrete time, a generalized Pareto distribution for describing the out-degree distribution, and an inverse Gaussian probability distribution function to express the aging effect based on the difference between publication times. The model is initiated by analyzing statistics papers assembled from the Web of Science database. By using the proposed model, we can generate graphs and demonstrate that they are similar to the original data concerning the in-degree, out-degree, and triangular distributions. In addition, we analyze two other citation networks derived from physics papers in the arXiv database and verify the effectiveness of the model.

Estimation of High Dimensional Vector Autoregression via Sparse Precision Matrix

Benjamin Poignard*
Graduate School of Economics, Osaka University and Riken-AIP, Japan

Manabu Asai†
Faculty of Economics, Soka University, Japan

Abstract

We consider the problem of estimating sparse vector autoregression (VAR) processes via penalized precision matrix. Such matrix is the output of the underlying directed acyclic graph of an indirect form of the VAR process, whose zero components correspond to zero coefficients in the indirect form. The precision matrix estimators are deduced from the class of Bregman divergences and regularized by the SCAD, MCP and LASSO penalties. Under suitable regularity conditions, we derive error bounds for the regularized precision matrix for each Bregman divergence. Moreover, we establish the support recovery property, including the case when the penalty is non-convex. These theoretical results are supported by empirical studies.

*E-mail address: bpoignard@econ.osaka-u.ac.jp (corresponding author)
†E-mail address: m-asai@soka.ac.jp

市場ノイズを含む場合のジャンプ・ブラウン運動の汎関数の局所SIML推定
(Local SIML Estimation of Some Brownian and Jump Functionals Under Market Microstructure Noise)

January 2022

国友直人 (Naoto Kunitomo, Institute of Statistical Mathematics)
and
佐藤整尚 (Seisho Sato, University of Tokyo)

Abstract

To estimate some Brownian and jump functionals from high-frequency financial data under market microstructure noise, we introduce a new local estimation method of the integrated volatility and higher-order variation of Ito's semi-martingale processes. Although it is straight-forward to extend the realized volatility (RV) estimation to more general cases without micro-market noise, it may not be straight-forward to estimate Brownian and Jump functionals in the presence of micro-market noise. We develop the local SIML (LSIML) method, which is an extension of the separating information maximum likelihood (SIML) method proposed by Kunitomo, Sato and Kurisu (2018) and Kunitomo and Kurisu (2021). The new method is simple and the LSIML estimator has some desirable asymptotic properties as well as reasonable finite sample properties.

Key Words

High-Frequency Data Analysis, Integrated Volatility, Separating Information Maximum Likelihood (SIML), Higher-Order Brownian and Jump Functionals, Stable Convergence, Local SIML Estimation

References

[1] Ait-Sahalia, Y. and J. Jacod (2014), *High-Frequency Financial Econometrics*, Princeton University Press.

[2] Barndorff-Nielsen, O., P. Hansen, A. Lunde, and N. Shephard (2008), "Designig Realized Kernels to Measure the Ex Post Variation of Equity Prices in the Presence of Noise," *Econometrica*, 76-6 (2008) 1481-1536.

[3] Hausler, E. and Luschgy, H. (2015), *Stable Convergence and Stable Limit Theorems*, Springer.

[4] Jacod, J., Y. L., Per A. Mykland, M. Podolskijc, and M. Vetter (2009), "Microstructure noise in the continuous case: The pre-averaging approach," *Stochastic Processes and their Applications*, 119 (2009) 2249?2276.

[5] Ikeda, N. and S. Watanabe (1989), *Stochastic Differential Equations and Diffusion Processes*, 2nd Edition, North-Holland.

[6] Jacod, J. and P. Protter (2012), *Discretization of Processes*, Springer.

[7] Kunitomo, N. and Kurisu, D. (2017), "Effects of Jump and Noise in High-Frequency Financial Econometrics," *Asia-Pacific Financial Markets*, Springer.

[8] Kunitomo, N., S. Sato and D. Kurisu (2018), *Separating Information Maximum Likelihood Method for High-Frequency Financial Data*, Springer.

[9] Kunitomo, N. and Sato, S. (2021), "Local SIML Estimation of Some Brownian and Jump Functionals Market Microstructure Noise," MIMS-RBP Statistics & Data Science Series (SDS-21),
http://www.mims.meiji.ac.jp/publications/datascience.html

[10] Kunitomo, N. and Kurisu, D. (2021), "Detecting Factors of Quadratic Variation in the Presence of Market Microstructure Noise," Japanese Journal of Statistics and Data Science (JJSD), 4(1), 601-641, Springer.

[11] Malliavin, P. and M. E. Mancino (2009), "A Fourier Transform Method for Nonparametric Estimation of Multivariate Volatility," *Ann. Statist.*, 37, 1993-2010.

[12] Mancino, Maria Elvira, Recchioni, Maria Cristina, and Sanfelici, Simona (2017), "Fourier-Malliavin Volatility Estimation Theory and Practice," Springer.

ビッグミクロデータのプライバシー保護

星野伸明 *

2022 年 2 月 17 日

「ミクロデータ」とは、特定個体の属性変数の実現値の組で、これを一行（レコード）として扱うのがミクロデータセットと考えられる。社会科学において方法論的個人主義を採用するならば、ミクロデータの分析によって得られる知見は一義的な価値を持つ。特に経済学では方法論的個人主義が席巻し、一昔前は集計表の分析にとどまっていた社会科学者たちは、ミクロデータを分析するようになった。

松田 (2000) によれば 1940 年代から社会科学者のミクロデータ分析は始まっており、労働経済学など一部の分野では 1960 年代までに定着している。当時は簡単な手続きでミクロデータが利用出来たと言われている。しかし 1960 年代後半から、欧米諸国においてプライバシーが意識されるようになった。宇賀 (2012) によれば、プライバシーを「自己情報をコントロールする権利として理解する立場が有力」である。また「この考えの影響を受けた個人情報保護立法は 1970 年代からみられるようになった」。この時期以降、ミクロデータの情報保護研究は本格化する。

実は集計表の情報保護は遅くとも 1940 年代には米国センサス局で始まっており (Cox, 2001)、統計学的なミクロデータの保護は、集計表保護概念の影響を強く受けている。集計表は magnitude の集計結果の場合もあるが、ミクロデータの保護と強く関係するのは度数の集計表、すなわち分割表である。

分割表の保護における目標は、度数が小さいセルにおいて真の度数を隠す[1]こと、と考えられてきた。例えば小売業の事業所が一つしか存在しない地方の二元分割表で（小売業, 売上一千万円）のセル度数が 1 と分かれば、この事業所の売上は露見する。もし小売業の事業所が多くある地方だったとしても、売上が高い事業所は限られるだろう。つまり分割表においてセル度数が小さければ、そこに所属する個体群を絞り込めることはあり得る。

営業秘密が暴露されるような統計調査に協力は期待できないので、データ主体の秘密保護については真剣に研究されている。研究の蓄積により分野として確立しており、統計的開示制限 (statistical disclosure limitation) ないし管理 (control) と呼ばれる。なお厳密にはプライバシーは個人の権利である。法人も含めて秘密を保護する概念は “confidentiality” と呼ばれる。

さて、ミクロデータは集計表とは別物と理解されているかもしれないが、（離散観測の）ミクロデータセットは分割表と一対一に対応させることが出来る。つまりミクロデータの属性実現値の組は、分割表のセルを一意に示す。目視のための分割表は低次元にならざるを得ないが、分割表とは個体の属性空間を分割して各セルに所属する個体数を示すものであり、目視の制約を外せば属性空間の次元は高くても問題ない。実際、計算機による処理が前提のデータベースでは、多次元分割表（データキューブ）を処理の高速化に用いる。

従って度数が小さいセルを隠すということは、同一属性の個体数が少ないレコードを隠すということと同値である。集計表の秘匿と同様に考えればレコードを削除することになるが、ミクロ

*金沢大学経済学経営学系

[1]例えば平成 28 年経済センサスでは度数が 2 以下のセルに x を代入している。太田・北原 (2019) を見よ。

データセットには周辺度数がないこともあり、好まれない。最もよく使われるのは「再符号化」ないし「一般化」である。この方法は属性の分類を粗くし、セルを併合することで小さい度数のセルを減らす。例えば 1 歳刻みで記録されている年齢を 5 歳刻みに再符号化すれば、セルの度数は単調非減少である。全てのレコードについて同じ（大域的）符号化をするなら結果は “grouped data” であり、これを社会科学者は扱いなれているので受容されたと考えられる。

ところが再符号化は、いわゆるビッグデータの保護において主役ではあり得ない。ビッグデータという言葉が 2010 年代に流行した背景には、多くの人がインターネット上で活動履歴を膨大なデータとして残すようになった社会状況が存在する。そのようなミクロデータセットの属性変数の種類（フィールド数）は桁違いに多いものがある。例えば通販サイトではユーザーの各時点の挙動がそれぞれ変数（フィールド）である。フィールドが増えれば、分割表のセル度数はほとんど 0、正でも 1 ばかりということになる。このような状況で再符号化をすると、度数が非ゼロの遠いセル同士を併合することになり、データに分析する価値が残らない。

ビッグミクロデータの場合、度数を不確実にすることによっても度数が小さいセルが隠れると考えるべきである。非確率的な方法も存在するが、ランダムネスによる情報保護が「疑似データ」の生成や情報学的なプライバシー保護では主流である。例えばデータ（経験分布）をそのまま公開するのではなく、経験分布からの iid（ブートストラップ）標本を公開すれば、度数は不確実になる。

しかし iid 標本はビッグミクロデータの保護に向かない。iid の場合、標本サイズに比例してフィッシャー情報量が増えるため、ビッグデータでは母集団（ブートストラップでは経験分布）が正確に推定出来てしまう。従って Fienberg(1994) の言うように経験分布を平滑化した母集団モデルからの標本を公開することが考えられるが、情報保護目的と母集団記述目的のモデリングが混交し最適化を妨げる。あくまで母集団モデリングは母集団記述を目的とし、情報保護はサンプリングで行うべきである。Hoshino (2009) の条件付き複合ポアソン分布族のメンバーは、性質が良い非 iid 標本を生成する。また母数によってフィッシャー情報量のオーダーを管理可能である。

参考文献

[1] Cox, L. (2001) *Disclosure risk for tabular economic data, Confidentiality, disclosure, and data access*, Doyle, P. et al. eds., pp.167–184, Elsevier, Amsterdam.

[2] Fienberg, S.E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Technical report, Department of Statistics, Carnegie Mellon University.

[3] Hoshino, N. (2009). The quasi-multinomial distribution as a tool for disclosure risk assessment. *Journal of Official Statistics*, **25**, 269–291.

[4] 松田 芳郎 (2000) 『統計調査制度とミクロ統計の開示』, 松田 芳郎他編, pp.19–23, 日本評論社.

[5] 太田 将彰・北原 昌嗣 (2019) 『平成 28 年経済センサス - 活動調査における秘匿処理について』, 統計研究彙報, 第 76 号, pp.115–134.

[6] 宇賀 克也 (2012) 『情報法』, 宇賀 克也・長谷部 恭男編，chapter 6, 有斐閣.

無情報事前分布の理解：公平な分配と未知母数から

柳 本 武 美: 統計数理研究所

1. 動機

ベイス法は基本的に既存の事前情報を有効に取り込む手法である。しかし、解析結果は利用した既存の事前情報に依存する。従って、異なる情報が正しいと信じる人とは解析結果を共有出来ない。経験的な観測にも基づいた推論を重視する、論理実証主義の観点からは受け入れられない。この問題を解決する試みが、無情報事前分布あるいは客観的事前分布の利用である。文字通り無情報事前分布であれば、それを利用しても解析結果は共有出来るはずである。一様分布・Jeffreys prior・reference prior などが提案され、他にも様々な具体的な事前分布が提案されている。

望ましい無情報事前分布を考察すると、経済学における公平な分配との関連に気付く。もし、多くの人が受け入れる無情報事前分布が定義出来るならば、公平な分配に対して新しい視点が得られると期待出来そうである。逆に言えば、無情報事前分布の定義は難しい問題であることを意味している。この問題について改めて考察したので報告する。

2. 無情報事前分布

標本モデルを $\mathcal{P} = \{p(\boldsymbol{y}|\theta); \boldsymbol{y} \in \mathcal{X}(\subset R^n), \theta \in \Theta(\subset R^p)\}$ とする。母数空間 Θ 上の一様分布が最も素朴な無情報事前分布である。しかし、この定義では母数の変数変換に依存するので不変になる Jeffreys prior が提案された。より洗練された提案が Bernardo (1979, JRSS B) による reference prior である。しかし、定義は厳密であるが適用範囲に制約があり、実際的なモデルでは漸近的性質を適用するしかない。

無情報事前分布を理解するために、$c \in \Theta$, $d(\cdot,\cdot)$ を Θ 上の (擬) 距離として、事前密度

$$\pi(\theta) \;=\; C \exp\{-\delta_0 d(\theta.c)\} b(\theta) \tag{1}$$

但し C は規格化定数、δ_0 は所与の正数、を考える。この定式化では、$b(\theta)$ が無情報事前関数と見なされる。実際、δ_0 が非常に小さいと、推論上は $b(\theta)$ を仮定した場合と殆ど変わらない。

3. 公平な分配と未知母数

公平な分配は効率的な管理・運営と並んで経済学の重要な概念である。また社会学的に見ても基本的である。公平な分配に関して異論が少ないのは次の二つの場合である。

1) 分配する対象の (少) 人数に応じて均等割にする

2) 特定の一人が独り占めする

である。1) が素朴な公平な分配であり 2) はその反対概念である。これらは各々、要素が有限個の空間 Θ の等確率分布、$c \in \Theta$ に対する一点分布 $\pi(\theta) = \delta_D(\theta - c)$ で表現出来る。

しかし、素朴な概念では均等分配では公平感は限られる。多数の勤労者への配分であれば、労働時間・能力・技能などに差があるから不公平感が生じる。事業の成果の配分では、投資額・投資リスクなどに違いがあるからやはり不公平感が生じる。不公平感を除くためには、ここの状況に応じて対応する。その手法を事前分布の視点から見れば、事前情報を最大限に利用することに対応する。労働技能とか投資リスクを定量化することは難しいが、事前の情報に応じて決めることになる。要するに、ベイズ的視点では事前情報の積極的利用に対応する。

公平な分配をその枠内で議論を進めても方向感に欠ける。より現実的には社会の人々が受け入れるかどうかといった委ねるしかない。結果として人々の価値観に依存する。

尤度論とか頻度論で多用される、未知母数・既知母数の概念は内容が薄弱である。既知母数の概念は 2) の独り占めに対応する。しかし、未知母数の概念はどうしようもなく意味が不明である。ベイズ法における、無情報/客観的事前分布の概念は未知母数の概念の一つの表現と見られる。逆に、未知母数の概念が公平な分配と無情報事前分布の理解を助ける視点は見当たらない。

4. 統計量としてのベイズ推定量

二つの無情報事前分布 $\pi_1(\theta), \pi_2(\theta)$ があったときに、どちらがより無情報であるかを議論することは難しい。もし二つの事前分布が誘導する推測法、例えば推定量、のどちらが優れているかであれば、議論が進み得る。頻度論では推定量はリスクにより評価される。

もしモデル $\mathcal{P}$ と観測値 $\boldsymbol{x}$ のみから構成される事前関数を用いて推定量を導出すれば、その事後平均・事後モードは統計量になる。言い換えると推定量である。その推定量の性能が良ければ良い推定量である。事前密度 (1) を仮定した場合には、c に依存するので統計量ではなくなる。推定量が良くなったのは取りも直さずに事前関数が望ましいからとしか考えようがない。こうした一般的な事前関数が無情報事前分布かどうかの議論は難しい。しかし、モデルと観測値からのみ構成したから、解析者の主観と価値は入る余地がない。この視点では強い制約がないので、幅広い事前関数が議論出来る。

無情報事前分布を狭く捉えれば、観測値に依存した事前分布は許容出来ないと思われる。しかし、モデルと観測値から構成したとすれば問題はない。要は導出された推定量の性能が良いかどうかである。一つの試みとして、$p(\boldsymbol{x}|\theta), \theta = (\psi, \lambda)$ が $pm(t|, \lambda)pc(\boldsymbol{x}|t, \psi)$ と分解出来る t が存在すると仮定する。二つの事前関数を $\pi_m(\psi) = 1/pm(t|\psi, \hat{\lambda}_{ML})$, $\pi_M(\psi, \lambda) = \pi_m(\psi)\pi_J(\psi, \lambda)$ 但し $\pi_J(\psi, \lambda)$ は Jeffreys prior と定義する。これらの下での事後平均/モードは統計量になる。これらの事前関数が無情報であるかを議論することは難しいが、幾らかの例では $\pi_M(\psi, \lambda)$ は reference prior に一致する。良い性能を示すことが観察される。

[本節の議論は宮田庸一氏 (高崎経済大学)、小椋透氏 (三重大学病院) との共同研究である]

パラメータ識別可能なグラフィカルモデルによる欠測を含んだ 2×2 分割表の対称性の検定

糸洲弘[1]，中川智之[2]，田畑耕治[2]

[1] 東京理科大学大学院理工学研究科，[2] 東京理科大学理工学部

1. はじめに

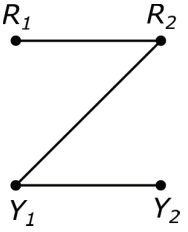
縦断研究やケース・コントロール研究，前後テスト分析等で発生するデータは正方分割表にまとめられる．これらの分割表においては，行の周辺分布と列の周辺分布が同等かどうかに関心がある．特に，2×2 分割表においては，対称性の検定によってこれを調べる（McNemar, 1947）．しかし，調査データには様々な理由から欠測が生じる．通常，欠測値を含むデータを分析するとき，欠測メカニズムを仮定する．欠測メカニズムは MCAR，MAR，NMAR の 3 つに分類され (Little and Rubin, 2020)，NMAR の場合には完全に観測された部分のみで解析を行うとパラメータ推定値にバイアスが生じる．そのため，無視不可能な欠測と呼ばれている．Ma *et al.* (2003) は，無視不可能な欠測を含む 2×2 分割表の欠測メカニズムをグラフィカルモデルを用いて示し，パラメータの識別可能性について議論した．

本報告では，無視不可能な欠測を含む 2×2 分割表の対称性の検定について議論する．Ma *et al.* (2003) で提案されたパラメータ識別可能なグラフィカルモデルに対して，対称性の制約を導入した仮説検定を提案する．適合度検定に対数尤度比統計量を用いることから，最尤推定値の計算には EM アルゴリズムを用いた．また，提案法の第 1 種の誤り確率と検出力をモンテカルロ・シミュレーションにより評価した．

2. 提案手法

行変数と列変数をそれぞれ Y_1 と Y_2 とし，対応する変数の欠測指標を R_1 と R_2 とする．欠測値を含んだ分割表に対して，同時分布 $P(Y_1 = y_1, Y_2 = y_2, R_1 = r_1, R_2 = r_2)$（以後，簡単のため $P(y_1, y_2, r_1, r_2)$ と記す）を考える．ただし，$y_t = j$ $(t = 1, 2; j = 1, 2)$，$r_t = k$ $(t = 1, 2; k = 0, 1)$ であり，0 を観測，1 を欠測とする．Ma *et al.* (2003) は，パラメータ識別可能なモデルの一つとして，図 1 のモデル (d) を提案した．

図 1: パラメータ識別可能なグラフィカルモデルの例（モデル (d) ）



欠測値を含んだ分割表に対して，対称性は

$$\sum_{r_1}\sum_{r_2} P(1, 2, r_1, r_2) = \sum_{r_1}\sum_{r_2} P(2, 1, r_1, r_2)$$

と表される．モデル (d) において，同時分布は

$$P(y_1, y_2, r_1, r_2) = P(y_1, y_2)P(r_2|y_1)P(r_1|r_2) = \pi_{y_1 y_2}\alpha_{r_2|y_1}P(r_1|r_2)$$

と表せるから，対称性の制約は $\pi_{12} = \pi_{21}$ となることに注意する．図 1 のモデル (d) に対称性（$\pi_{12} = \pi_{21}$）の制約を加えたもとでの最大対数尤度を求めるために EM アルゴリズムを用いる．求めたいパラメータは $\boldsymbol{\theta} = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, \alpha_{0|1}, \alpha_{1|1}, \alpha_{0|2}, \alpha_{1|2})$ である．E-step における Q 関数は

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \propto \sum_{l=0}^{1}\sum_{i=1}^{2} n_{\alpha_{l|i}}^{(t)} \log \alpha_{l|i} + \sum_{i=1}^{2}\sum_{j=1}^{2} n_{\pi_{ij}}^{(t)} \log \pi_{ij}$$

で与えられる．ただし，

$$\begin{cases} n_{\alpha_{0|i}}^{(t)} = \displaystyle\sum_j \left(n_{00,ij} + n_{10,+j}\frac{\alpha_{0|i}^{(t)}\pi_{ij}^{(t)}}{\sum_i \alpha_{0|i}^{(t)}\pi_{ij}^{(t)}} \right), \\ n_{\alpha_{1|i}}^{(t)} = \displaystyle\sum_j \left(n_{01,i+}\frac{\pi_{ij}^{(t)}}{\sum_j \pi_{ij}^{(t)}} + n_{11,++}\frac{\alpha_{1|i}^{(t)}\pi_{ij}^{(t)}}{\sum_{i,j} \alpha_{1|i}^{(t)}\pi_{ij}^{(t)}} \right), \\ n_{\pi_{ij}}^{(t)} = n_{00,ij} + n_{01,i+}\dfrac{\pi_{ij}^{(t)}}{\sum_j \pi_{ij}^{(t)}} + n_{10,+j}\dfrac{\alpha_{0|i}^{(t)}\pi_{ij}^{(t)}}{\sum_i \alpha_{0|i}^{(t)}\pi_{ij}^{(t)}} + n_{11,++}\dfrac{\alpha_{1|i}^{(t)}\pi_{ij}^{(t)}}{\sum_{i,j} \alpha_{1|i}^{(t)}\pi_{ij}^{(t)}}, \end{cases}$$

であり，$n_{00,ij}$ は完全観測部分のセル度数，$n_{10,+j}$ と $n_{01,i+}$ は一方のみ観測できたセル度数，$n_{11,++}$ は両方観測できなかったセル度数である．M-step の最大化点は，陽に求められ

$$\pi_{ij} = \frac{n_{\pi_{ij}}^{(t)} + n_{\pi_{ji}}^{(t)}}{2N}, \quad \alpha_{l|i} = \frac{n_{\alpha_{l|i}}^{(t)}}{n_{\alpha_{0|i}}^{(t)} + n_{\alpha_{1|i}}^{(t)}}$$

で与えられる．ただし，$N = \sum_{i,j} n_{\pi_{ij}}^{(t)}$ である．これらの結果を用いた検定方式，モデル (d) 以外のパラメータ識別可能なグラフィカルモデルを用いた場合の結果，シミュレーションによる性能評価については当日報告する予定である．

参考文献

Little, J. A. and Rubin, D. B. (2020). *Statistical Analysis with Missing Data*, 3rd ed., Wiley, Hoboken, New Jersey.

Ma, WQ., Geng, Z. and Li, XT. (2003). Identification of nonresponse mechanisms for two-way contingency tables. *Behaviormetrika* **30**, 125–144.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157.

連絡先: 糸洲弘・中川智之・田畑耕治 *
〒 278-8510 千葉県野田市山崎 2641
東京理科大学理工学部情報科学科
*E-mail: kouji_tahata@rs.tus.ac.jp

Deriving Lifetime Value of a Customer Who Exhibits Non-Poisson Purchase Behavior： Application to Marketing Intervention for Customer Retention

Makoto Abe
The University of Tokyo

In non-contractual CRM where customer withdrawal cannot be observed, the withdrawal is inferred by using BTYD (Buy Till You Die) models, such as Pareto/NBD. One advantage of a Pareto/NBD model --- the benchmark for customer-base analysis --- is that it requires minimal purchasing data from customers, namely recency and frequency. The tradeoff is the model posits strong assumptions on consumer behavior. Of those, the most restricting assumption is the Poisson purchase, which imposes a memoryless purchase process, implying that purchase occurs independent of the previous interpurchase time. The assumption is appropriate when purchases involve a wide variety of categories, such as purchases at a department store or an online shopping mall, since transactions tend to occur randomly. However, when transactions occur cyclically, for example, by focusing on a narrower product category, this assumption may not be satisfied. Furthermore, Pareto/NBD assumes that the two gamma distributions for the purchase rate and the withdrawal rate are independent. This ignores the association between the purchase behavior and the withdrawal behavior.

This research proposes a general purchase model that can accommodate either memoryless or cyclic pattern, which is then applied to customer-base analysis. Similar to the Pareto / NBD model, which has abundant previous research, it is assumed that the withdrawal is random and the amount per transaction follows a lognormal distribution within a customer, while a logistic threshold model captures the purchase periodicity. The proposed model (1) provides customer-specific measures for transaction cyclicity and lifetime value, (2) permits more accurate estimation of individual probabilities of being active, (3) results in a more reliable estimate for customer-base.

In the empirical analysis, customer transaction data from a hair salon was used to compare the individual model assuming random purchase and the aggregate model

(Pareto / NBD) based on the mixture distribution. The proposed model was superior to these two models in terms of prediction and parameter stability. Furthermore, as an application to customer retention strategies, the optimal intervention level that maximizes the increase in lifetime value was calculated for each customer.

Keywords: Customer Lifetime Value (CLV), BDTY Model, Random Purchasing, Periodicity, Pareto / NBD, Poisson

Discrete Choice in Marketing through the Lens of Rational Inattention

Sergey Turlo, Matteo Fina, Johannes Kasinger, Arash Laghaie, and Thomas Otter
Goethe University, Frankfurt

Models derived from random utility theory represent the work-horse methods to learn about consumer preferences from discrete choice data collected in experimental and observational settings. However, a large body of literature documents a variety of behavioral patterns that cannot be captured by basic random utility models, and that require different non-unified adjustments to accommodate these patterns. In this review article, we illustrate how a discrete choice model rooted in rational inattention (RI) theory nests a significant set of these patterns. We present illustrative simulations, discuss extant empirical work using experiments and observational data, and suggest how to develop an RI model for the analysis of discrete choice among multiple alternatives described along multiple attributes as encountered in prototypical discrete choice experiments and choice-based-conjoint analysis in marketing and economics.

Service Economy and Data Science

Nobuhiko Terui
Tohoku University

I present what I believe as necessary and important approach on the data science in service-oriented economy (service economy in short),

where I focus on

I. Personalization - heterogeneity modeling

II. Large scale data - high-dimensional sparse modeling

III. Network and UGC data - community detection using network and user generated contents in social media.

I explain the statistical and marketing models on these topics by the retrospective review of our works and conclude by stating current and future research.

A Choice Model of Utility Maximization and Regret Minimization

Abstract

Consumers often try to achieve multiple goals when purchasing products and services, where choices are sought that maximize utility and other objectives such as to minimize regret. If the choice outcomes are associated with high level of uncertainty at the time of purchase, consumers worry that the alternative of interest may turn out to be not optimal and they want to avoid this when making a purchase. We first propose a new regret function and explore its properties. Then we propose a generalized framework of multiple goal pursuit and apply it to a utility maximization and regret minimization problem. Pareto optimal sets are an outcome of multiple goal optimization problems where there exists multiple alternatives that are non-dominated. Our proposed framework allows us to generalize a dual-goal problem as a constrained optimization problem where either utility is maximized subject to constraints on regret, or regret is minimized subject to utility constraints. The proposed model fits the data better, provides improved predictions and offers a tractable solution to a problem of utility maximization and regret minimization.

Keywords: Multiple Goal Pursuit, Performance Uncertainty, Consideration Sets

The Impact of Gig Economy on Product Quality through the Labor Market: Evidence from Ride-sharing and Restaurant Quality

Minkyu Shin[*] Jiwoong Shin[†] Soheil Ghili[†] Jaehwan Kim[‡]

City University of Hong Kong[*], Yale School of Management[†], Korea University Business School[‡]

The rapid growth of the gig economy in recent years has transformed many sectors of the economy. Airbnb has challenged the hotel industry; Uber and Lyft have challenged traditional taxi companies and curtailed ridership on public transportation. And while these effects of gig work on direct competitors is very important, more indirect effects also merit attention.

This paper seeks to demonstrate the impact of the gig economy on the local economy beyond directly related incumbent industries through the labor market. We look at the restaurant industry as a case study for service sector. We design our analysis around a natural experiment where, due to regulatory shifts, Uber and Lyft exited the market in Austin, Texas, in May 2016 and returned in May 2017. Leveraging this exogenous exit and reentry, we conduct a series of analyses to study the relationship between rideshare and restaurant quality. Specifically, we are interested in examining the following hypothesis: The presence of Uber and Lyft in a city provides individuals with gig-work opportunities.

We first establish the relationship between the presence of ride-sharing companies and restaurant quality by analyzing how the quality of restaurant service in Austin responds to the presence of Uber and Lyft. We employ text analysis and difference-in-difference(DiD henceforth) approach to determine whether the entry and exit of Uber and Lyft influenced customer satisfaction with local restaurants. Dallas serves as a control group. We use every Yelp review of restaurants in Austin and Dallas from 2014 to 2019 to measure quality. This entails text analysis of each review to capture restaurant quality along two dimensions: service and food. First, leveraging a difference-in-difference setting, we find that the quality of service *decreases* in Austin relative to Dallas with the presence of Uber and Lyft.

Secondly, we carry out our main analysis a second time looking at customer satisfaction with food quality, rather than service quality, as our dependent variable. We hypothesize that customer experience with the food quality would be less influenced by the presence of Uber and Lyft than is the service quality. Employees in charge of the food quality such as chefs working in the kitchen are not much attracted by the opportunity to drive for Uber and Lyft, relative to workers such as wait-staff mostly dealing with the service for customers. We demonstrate that the customer evaluation of *food* quality does *not change* before and after Uber and Lyft's re-entry to Austin.

Third, we divide restaurants into two tiers based on pricing labels provided by Yelp. One group consists of restaurants that are assigned a single dollar sign($) in Yelp, meaning they are cheaper. The rest of the restaurants, those with two or three dollar signs($$-$$$), comprise the second group. Workers in low-tier restaurants would be paid less, either because their base hourly wage is lower or their tipped income is lower. Therefore, we expect Uber and Lyft's impact to be

more pronounced for a single dollar sign restaurants whose service workers are more likely to be lured by gig-work opportunities. Our empirical analysis confirms this expectation. We find that the effect of Uber and Lyft in Austin on service quality is significant for *single-dollar sign restaurants*. In contrast, we do not find any significant effect for high-tier restaurants.

Next, we directly test our mechanism by examining turnover rates of staff at restaurants by leveraging a unique worker-level dataset of restaurants in Austin and Dallas from 2014 to 2019. We examine how the turnover rate of staff in Austin's restaurants changes with the local activity of Uber and Lyft relative to Dallas. We find that the turnover rate *increases* in Austin relative to Dallas after Uber and Lyft return. Additionally, in the same spirit as the dollar-sign analysis in Yelp Data, we use the restaurant category information in the dataset to examine whether we see a similar pattern in the turnover rates for different restaurant tiers. In our hypothesis, we expect the effect of Uber and Lyft on turnover rates to be stronger for low-end restaurant categories than relatively high-end restaurants. We conduct separate DiD analyses for each category and indeed find an increase in turnover only for *low-end restaurants*. In contrast, we do not see such a pattern for middle-end or high-end restaurants.

We then delve deeper into the analysis by decomposing the turnover rates into those for "back-of-house" staff and "front-of-house" staff. The latter group represents those who directly deal with customers and consists mainly of service staff, whereas the former includes higher-paid positions such as managers and chefs. The results are consistent with the Yelp review data analysis: the increase in turnover is observed only for *front-of-house workers*, while there is no significant effect for back-of-house staff.

Finally, we check our results by conducting several robustness checks and discuss other alternative explanations based on demand-side channels. One would expect some other channels through which the rideshare companies could have impacted the local economy, such as the demand changes due to the easier mobilization. While we cannot completely rule out all possible explanations, we show that these alternative accounts cannot fully explain the patterns observed in our data. We present other evidence suggesting that our findings are more likely to arise from the supply-side channel through the labor market rather than the demand-side channels.

Our paper also contributes to marketing literature, especially in service marketing and the role of employee for customer satisfaction, making a connection between employee turnover and customer satisfaction about a restaurant's service quality. This work shows that the expansion of the gig economy, by providing new work opportunities for low-wage, low-skilled workers, has far-reaching and significant ramifications on broader industries through the labor market. As the gig economy expands, as it is predicted to do, understanding these second-order effects will be critical for the development of effective regulatory policy. Our work focuses on the hospitality sector, but we consider it a telling case study for the service economy as a whole and hope it serves as a starting point for deeper study of how the gig work may shape the future economy.

A Multiple Duration Choice Model for Service Data

Takuya Satomura (Keio University)

Greg M. Allenby (Ohio State University)

The consumption of services often involves the presence of fixed costs that are shared among services when they are jointly consumed. Lawn care companies, for example, may offer a variety of yard work services (digging, weeding, fertilizing) at discounted rates if they can do the work during the same visit, and hair salons offer services that, when consumed together, allow consumers to minimize their cost of travel. Understanding the cost of access and consumption, therefore, requires models where purchase timing and choice are integrated into a common model of behavior, where the decision of when to purchase and what to purchase are driven by the same latent utilities. In this paper, the authors propose a competing risks model that simultaneously captures purchase timing and choice. The model assumes that consumers sequentially maximize utility and decide to make a purchase when the utility for one of the inside goods is greater than the utility of the outside, no-choice option. The results of the empirical analysis indicate that consumer utility for the goods not purchased continues to accumulate in value rather than resetting to zero, implying that latent, unmet demand is always present. Implications for maximizing service profitability are explored.

Keywords: Competing Risks Model, Utility Maximization, Fixed Cost, Service Consumption

Robust False Discovery Rate Control via Debiased Rank Lasso

Kazuma Sawaya and Yoshimasa Uematsu

Department of Economics and Management, Tohoku University

1 Model and the Rank Lasso

We consider the high-dimensional linear model

$$y = X\beta^0 + \varepsilon, \tag{1}$$

where $y = (y_1, \ldots, y_n)^\top$ is the vector of n independent responses, $\beta^0 \in \mathbb{R}^p$ is the sparse coefficient vector, $X = (X_1, \ldots, X_n)^\top$ is the empirically centered n independent observations of p features, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is the i.i.d. error vector with density function $f(\cdot)$. Let $S \subset \{1, \ldots, p\}$ denote the set of important variables; i.e., $S = \{j : \beta_j^0 \neq 0\}$. Our goal is to propose a robust method to yield "discoveries" $\widehat{S}$ of S with controlling the false discovery rate (FDR), $\mathbb{E}\left[|S^c \cap \widehat{S}|/|\widehat{S}|\right]$, below predetermined level.

Wang et al. (2020) propose the rank Lasso estimator

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n(n-1)} \sum_{i \neq j} \left| (y_i - y_j) - (X_i - X_j)^\top \beta \right| + \lambda \|\beta\|_1 \right\}, \tag{2}$$

which is robust to heavy-tailed error distribution such as Cauchy distribution. Using this estimator, we propose a new method for variable selection.

2 New Methodology

We start with removing the bias of $\widehat{\beta}$ to make it asymptotically normal in a similar way of the debiased Lasso estimator(Javanmard and Montanari, 2014). Define

$$\widehat{\beta}^d := \widehat{\beta} + \frac{1}{\int f^2(u)du} \frac{1}{2n(n+1)} \widehat{\Theta} X^\top (2r(\widehat{\varepsilon}) - (n+1)), \tag{3}$$

where $\widehat{\Theta}$ is a consistent estimator of the precision matrix $\Theta = \mathbb{E}[X_1 X_1^\top]^{-1}$, $\widehat{\varepsilon} = y - X\widehat{\beta}$, and $r(\cdot)$ is the rank. Thanks to the asymptotic normality of $\widehat{\beta}_j^d$ as shown below, some selection algorithms with this estimator work well. Specifically we employ the *data-splitting* method of (Dai et al., 2020); see Algorithm 1.

3 Theoretical Results

We justify that our Algorithm 1 controls the FDR under the preassigned level.

Theorem 1 *Under regularity conditions, we have for any $j \in \{1, \ldots, p\}$,*

$$\sqrt{n}(\widehat{\beta}_j^d - \beta_j^0) = Z_j + \Delta_j, \tag{5}$$

where $Z = [\sqrt{12n} \int f^2(u)du]^{-1} \widehat{\Theta}^\top S(\varepsilon)$ and Δ is some remainder such that

$$Z_j \xrightarrow{d} N\left(0, \frac{1}{12[\int f^2(u)du]^2} \Theta_{jj}\right), \quad \|\Delta\|_\infty = o_p(1).$$

Algorithm 1 FDR control via a single data split with the debiased rank Lasso

1. Split the data into two groups $(y^{(1)}, X^{(1)})$ and $(y^{(2)}, X^{(2)})$, independent to the response vector y.
2. Estimate (3) for each group as impact coefficients $\widehat{\beta}^{d(1)}$ and $\widehat{\beta}^{d(2)}$ with some $\widehat{\Theta}^{(1)}, \widehat{\Theta}^{(2)}$ on each part of the data.
3. Calculate the mirror statistics
$$M_j = \mathrm{sign}\big(\widehat{\beta}^{d(1)}\widehat{\beta}^{d(2)}\big) g\big(|\widehat{\beta}^{d(1)}|, |\widehat{\beta}^{d(2)}|\big),$$
where function $g(u, v)$ is non-negative, symmetric about u and v, and monotonically increasing in both u and v.
4. Given a designated FDR level $q \in (0, 1)$, calculate the cutoff τ_q as:
$$\tau_q = \min\left\{ t > 0 : \widehat{\mathrm{FDP}}(t) = \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\} \vee 1} \le q \right\}. \tag{4}$$
5. Select the features $\{j : M_j > \tau_q\}$.

Theorem 2 *For any designated FDR level $q \in (0, 1)$, assume that $s \log p \ll \sqrt{n}$ and that there exists a constant $t_q > 0$ such that $\mathbb{P}(FDP(t_q) \le q) \to 1$ as $p \to \infty$. Then, under regularity conditions, we have*

$$\limsup_{n,p\to\infty} FDR(\tau_q) \le q.$$

In the presentation, we will confirm the validity and robustness of our procedure through numerical experiments.

References

Dai, C., B. Lin, X. Xing, and J. S. Liu (2020). False discovery rate control via data splitting. *arXiv preprint arXiv:2002.08542*.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research 15*(1), 2869–2909.

Wang, L., B. Peng, J. Bradic, R. Li, and Y. Wu (2020). A tuning-free robust and efficient approach to high-dimensional regression. *Journal of the American Statistical Association*, 1–44.

Innovation algorithm of fractionally integrated ($I(d)$) process and applications on the estimation of parameters

Junichi Hirukawa* and Kou Fujimori
Niigata University and Shinsyu University

1 Introduction

The long memory phenomena frequently occur in the empirical studies of various fields. The fractionally integrated process is the one of the suitable candidate which appropriately represents the long memory property. There are two recursive algorithms for determining the one-step predictors of time series, that is, the Durbin-Levinson algorithm and the innovation algorithm. The Durbin-Levinson algorithm for the fractionally integrated process is well-known and widely used, which naturally derives the Cholesky factorization of the inverse matrix of the covariance matrix of the process. In this talk, we derive the innovation algorithm for the fractionally integrated process. The result is also applied to the derivations of the Cholesky factorization of the covariance matrix and the Gaussian likelihood of the process in the explicit forms. Moreover, the asymptotic theory of Gaussian maximum likelihood estimator (GMLE) is derived in terms of the innovation algorithm.

In this talk, we consider one of the long memory process so-called the fractionally integrated ($I(d)$) process defined by

$$(1-L)^d z_t = \varepsilon_t, \quad (t = 1, \ldots, n) \tag{1}$$

and $z_r = 0$, $(t \leq 0)$, where $d \in (-1/2, 1/2)$, $(d \neq 0)$, L is the lag operator and $\{\varepsilon_t\} \overset{i.i.d.}{\sim} (0, \sigma^2)$.

There are two recursive algorithms for determining the one-step predictors $\hat{x}_{n+1}, n \geq 1$ defined by

$$\hat{x}_{n+1} = \begin{cases} 0 & \text{if } n = 0, \\ P_{\mathcal{H}_n} x_{n+1} & \text{if } n \geq 1. \end{cases}$$

One of which is the Durbin-Levinson algorithm and the other is the innovation algorithm. The Durbin-Levinson algorithm for $I(d)$ process is well-known and widely used (see e.g., Hosking (1981), Brockwell and Davis (1991)).

On the other hand, the second recursion, which is so-called the innovation algorithm, for $I(d)$ and $ARIMA(p, d, q)$ processes have not been established.

While the Durbin-Levinson algorithm gives the coefficients of $x_1, \ldots, x_n$ in the representation

$$\widehat{x}_{n+1} = \sum_{j=1}^{n} \phi_{n,j} x_{n+1-j},$$

the innovation algorithm gives the coefficients of the "innovations", $u_{j-1}^{x} = (x_j - \hat{x}_j)$, $j = 1, \ldots, n$, in the orthogonal expansion

$$\hat{x}_{n+1} = \sum_{j=1}^{n} \theta_{n,j} \left(x_{n+1-j} - \hat{x}_{n+1-j} \right) = \sum_{j=1}^{n} \theta_{n,j} u_{n-j}^{x}.$$

Defining $\theta_{n,0} = 1$, we can also have an innovations representation of x_{n+1} itself, that is,

$$x_{n+1} = \sum_{j=0}^{n} \theta_{n,j} \left(x_{n+1-j} - \hat{x}_{n+1-j} \right) \sum_{j=0}^{n} \theta_{n,j} u_{n-j}^{x}.$$

Then, we have the Gaussian log-likelihood of $I(d)$ process for $\boldsymbol{\theta} = (d, \sigma^2)'$

$$l(\boldsymbol{\theta}) = l\left(d, \sigma^2\right) = -\frac{n}{2} \log\{2\pi\} - \frac{1}{2} \sum_{j=1}^{n} \log v_{j-1}(\boldsymbol{\theta}) - \frac{1}{2} \sum_{j=1}^{n} \frac{u_{j-1}(d)^2}{v_{j-1}(\boldsymbol{\theta})}, \tag{2}$$

where $u_n = z_{n+1} - \hat{z}_{n+1}$ is the zero mean uncorrelated process with $E(u_n^2) = v_n$.

Theorem 1. *Let $\{z_t\}$ is the Gaussian $I(d)$ process defined in (1) with $d \in (-1/2, 1/2)$, $(d \neq 0)$. And let $\widehat{\boldsymbol{\theta}} = \left(\widehat{d}, \widehat{\sigma}^2\right)'$ is the Gaussian MLE (GMLE) of $\boldsymbol{\theta} = (d, \sigma^2)'$ which maximizes the Gaussian log-likelihood (2). Then, the GMLE $\widehat{\boldsymbol{\theta}}$ has consistency, that is,*

$$\widehat{\boldsymbol{\theta}} \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_0,$$

where $\boldsymbol{\theta}_0 = (d_0, \sigma_0^2)'$ is the true value of $\boldsymbol{\theta}$.

References

BROCKWELL, P. J. AND DAVIS, R. A. (1991). *Time Series: Theory and Methods*. New York: Springer.

HOSKING, J. R. M. (1981). Fractional Differencing. *Biometrika* **1**, 165–176.

LI, W. K. AND MCLEOD, A. I. (1986). Fractional Time Series Modelling. *Biometrika* **73**, 217-221.

YAJIMA, Y. (1985). On Estimation of Long - Memory Time Series Models. *Australian Journal of Statistics* **27**, 303–320.

Machine Collaboration*

Qingfeng Liu† and Yang Feng‡

February 5, 2022

Ensemble learning has emerged and been extensively studied by many in the past few decades (e.g., Dasarathy and Sheela (1979), Schapire (1990), Ho (1995), and Breiman (1996)), with its popularity recently skyrocketing (e.g., Lu and Van Roy (2017), Yu et al. (2018), Qi et al. (2019), and Tian and Feng (2021)). Mendes-Moreira et al. (2012), Sagi and Rokach (2018), and Dong et al. (2020) are some recent comprehensive surveys. The general idea of ensemble learning is to combine the predictions obtained from different learning methods (hereafter, base machines), or predictions based on different subsamples or different feature spaces, in order to improve prediction performance. Bagging, stacking, and boosting are three prominent examples. In bagging (Breiman, 1996) and stacking, base machines are first run in parallel and independently, and then the final prediction is constructed as a simple/weighted average of the predictions from these base machines. In boosting (Schapire et al., 1998), the base machines work jointly in a top-down manner. In all three algorithms, the output from each base machine is fixed after being calculated. Like human collaboration, an idea that may yield potential improvement is to let different kinds of base machines communicate with each other and update their outputs after observing the predictions of the other base machines. Based on this idea, we propose the *Machine Collaboration* (MaC) ensemble learning framework with heterogeneous base machines, where the word heterogeneous stands for that the base machines are of different types (e.g. DT, DNN, Ridge Regression). Compared with bagging, stacking, and boosting, MaC has the following desirable features. Figure 1 provides the schematic for bagging, stacking, boosting, and MaC. As illustrated, bagging and stacking are parallel & independent, boosting is sequential & top-down, while MaC is *circular & recursive*. In the framework of MaC, base machines work in a circular manner. Further, the circulation goes multiple rounds. Valuable information is passed recursively through base machines around a "round table", but not top-down. In this process, the base machines update their structures and/or parameters once in each round, according to the information received from the other machines. We demonstrate that MaC can deliver competitive performance when compared with the base machines or other ensemble methods.

*The authors gratefully acknowledge the support of the Japan Society for the Promotion of Science through KAKENHI Grant No. JP19K01582 (Liu), the Nomura Foundation for Social Science Grant No. N21-3-E30-010 (Liu) and a National Science Foundation CA-REER Grant No. DMS-1554804 (Feng).

†Corresponding author. Department of Economics, Otaru University of Commerce, Otaru City, Hokkaido, Japan. Email: qliu@res.otaru-uc.ac.jp.

‡School of Global Public Health at New York University, NY, NY, USA. Email: yang.feng@nyu.edu.

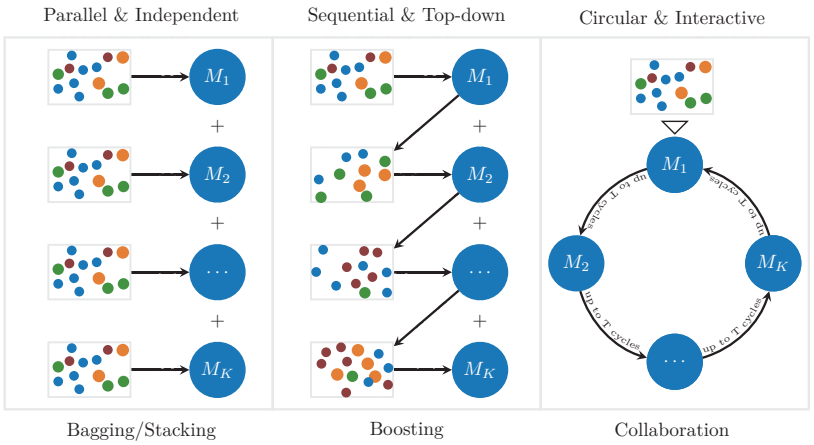


Figure 1: Bagging, boosting, and machine collaboration

References

Breiman, L., 1996. Bagging predictors. Machine learning 24, 123–140.

Dasarathy, B.V., Sheela, B.V., 1979. A composite classifier system design: Concepts and methodology. Proceedings of the IEEE 67, 708–713.

Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q., 2020. A survey on ensemble learning. Frontiers of Computer Science 14, 241–258.

Ho, T.K., 1995. Random decision forests, in: Proceedings of 3rd international conference on document analysis and recognition, IEEE. pp. 278–282.

Lu, X., Van Roy, B., 2017. Ensemble sampling, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 3259–3267.

Mendes-Moreira, J., Soares, C., Jorge, A.M., Sousa, J.F.D., 2012. Ensemble approaches for regression: A survey. Acm computing surveys (csur) 45, 1–40.

Qi, Y., Liu, B., Wang, Y., Pan, G., 2019. Dynamic ensemble modeling approach to nonstationary neural decoding in brain-computer interfaces, in: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 6087–6096.

Sagi, O., Rokach, L., 2018. Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8, e1249.

Schapire, R.E., 1990. The strength of weak learnability. Machine Learning 5, 197–227.

Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.S., 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. The annals of statistics 26, 1651–1686.

Tian, Y., Feng, Y., 2021. Rase: Random subspace ensemble classification. Journal of Machine Learning Research .

Yu, Z., Luo, P., Liu, J., Wong, H.S., You, J., Han, G., Zhang, J., 2018. Semi-supervised ensemble clustering based on selected constraint projection. IEEE Transactions on Knowledge and Data Engineering 30, 2394–2407.

コロナ禍における感染拡大防止対策の人流抑制効果に関する実証分析

東北大学　佐藤 宇樹
大阪市立大学　黒田 雄太
東北大学　松田 安昌

本研究の目的は、新型コロナウイルス感染症の感染拡大防止を目的とした人流抑制政策の効果を実証的に明らかにすることである。2020 年 1 月以降、新型コロナウイルス感染症の感染拡大は世界規模で起こっており、その感染拡大を抑制することは近年の重要な課題の一つとなっている。日本においても、政府や地方自治体が中心となり、これまでに緊急事態宣言やまん延防止等重点措置などの政策が人流の抑制を目的に行われてきた。宣言を行うことで人々の移動や行動を制限することが可能になり、新型コロナウイルスの感染拡大防止に効果があると考えられている。しかし、宣言時には外出制限や自粛による経済の停滞や、飲食店への補助金の支給などの経済コストが発生している。日本全国の人流データを用いて政策の人流抑制効果をデータから定量的に明らかにすることは、現在行われている補助金による所得の再分配が適切であるか否かを検証するためにも重要である。

本研究の特徴は以下の 3 点である。1 点目は欧米ではなく日本の人流データを使用している点である。データへのアクセスの容易さなどから、既存研究の多くは欧米の人流データを用いて人流抑制政策の効果を検証している。本研究では日本の人流データを使用することで、人流抑制政策の効果が欧米の国々と日本ではどのように異なるのかについて明らかにする。2 点目は 2021 年以降の新型コロナウイルス感染症の感染拡大の第 2 波以降の政策の効果も評価している点である。既存研究では感染拡大初期の 2020 年前半に行われた政策の評価を行っているものが多い。本研究では 2020 年 1 月から 2021 年 9 月までのデータを用いて分析を行うことで、2021 年以降に実施された 2 回目以降の緊急事態宣言の効果も検証し、政策の回数が増えるにつれて、政策の効果がどのように変化するのかについて明らかにする。3 点目は地域の特性を表す市民資本という概念を用いて、政策効果の地域差に関して分析している点である。先行研究ではその地域の人々の利他性や公共性を図る指標である市民資本を用いて、政策効果の地域差を明らかにしている。本研究においても、投票率から作成された各地域の市民資本を用いることで、日本においても市民資本の蓄積に応じて、政策効果の地域差が存在することを明らかにする。

本研究では NTT ドコモが提供しているモバイル空間統計データを用いて作成した人流データに、固定効果パネルデータモデルを応用することで、以下の実証分析結果を得た。まず、緊急事態宣言やまん延防止等措置の政策が行われている期間には政策が行われていない期間と比較して、人流が抑制されていることが明らかになった。また、宣言時以外の平時でも人流は 2020 年前半の新型コロナウイルス感染症の感染拡大以前の水準には戻っていないことも示された。次に、市民資本を用いた政策効果の地域差の比較から、1 回目や 2 回目などの緊急事態宣言の回数が少ない時期にはどの地域でも政策の効果はほぼ同じだが、3 回目や 4 回目と宣言の回数が多くなるにつれて、市民資本の高い地域よりも低い地域の人々の流出を抑制する効果が低くなってきていることが明らかになった。また、飲酒に関する政策の違いを用いた分析から、市民資本の高い地域では制限の強い政策をとるほどその地域の人々の他地域への流出を抑制する効果が強くなるが、市民資本の低い地域の人々に関しては制限を強くしても効果が変わらないことが示された。

実証分析結果から得られる考察は次の 2 点である。 1 点目として、市民資本の低い地域の人々の流出を抑制するには飲酒の禁止や提供時間の短縮だけだけでなく、飲食店の営業禁止など、より強い政策が必要であると考えれる。2 点目として、人流抑制政策を行うと、市民資本の高い地域の人々の流出は抑えられるが、その地域へ他の地域の人々が移動する流入は抑えることができない。そのため、流入を減らすためには、市民資本の低い地域だけでなく、高い地域に関してもこれまで以上に制限の強い政策を同時に行う必要があると考えられる。

購入型クラウドファンディングの資金調達パターン分析

内田彬浩（筑波大学大学院）・伴正隆（筑波大学）

近年、購入型クラウドファンディングの市場規模が拡大しており、同市場のメカニズムの理解が必要とされているが、新興分野であるため学術・実務ともにその理解は進んでいない。内田・伴（2022, 日本ベンチャー学会誌）では購入型クラウドファンディングプロジェクト（以下、PJ）の最終的な購買件数と、PJ が掲げる目標金額を超える資金を調達出来たか否かの PJ 成否について、潜在クラス回帰モデルによって理解を深めた。

本研究では PJ の資金調達パターンについて、Bass モデルへの適用によって理解を深める。Bass モデルが対象とする現象は、頻繁には購入されない製品における初回購買による製品普及である。購入型クラウドファンディングで扱う出資への対価（リターン）は、一般的な小売店では購入できない製品や、市販される前の製品であり、さらに PJ 終了までリターンを手に取ることは出来ず、PJ 期間内にリピート購買が出来ないものであり、Bass モデルが対象とする現象との類似点があると考える。また、Bass モデルが想定するイノベーターとイミテーターについても、PJ 閲覧時の資金調達具合を考慮せずに意思決定する出資者と、そうではない出資者の存在としてそれぞれ援用することが出来ると考える。しかし、一般的な製品の普及と異なり、PJ の実施期間が有限であるせいか、Bass モデルでは想定されない PJ 実施期間末期での出資（購買）の集中があり、購入型クラウドファンディング資金調達パターンを分析するのに適したモデルが必要であり、本研究で提案する。

PJ の資金調達時系列データは、PJ 開始直後から終了まで等時間間隔で記録された、各時点の購買件数データである。時間を $t(t = 1, \dots, T)$とし、n_t を t 期における購買件数とすると、クラウドファンディングの変数で表現する Bass モデルは

$$n_t = uf_t = up\left(1 - \frac{N_t}{u}\right) + uq\frac{N_t}{u}\left(1 - \frac{N_t}{u}\right) = up(1 - F_t) + uqF_t(1 - F_t) \qquad (1)$$

となる。ここで u は PJ 終了時点の潜在的な総購買件数パラメータ、n_t と N_t はそれぞれ t 期における購買件数と t 期までの累積購買件数、p と q はそれぞれイノベーター係数とイミテーター係数である。f_t と F_t はそれぞれ n_t と N_t に対応する割合データである。Bass(1969)にあるように、一般的な解法では、(1)式の微分方程式を解いて、時間を説明変数とする非線形回帰によってパラメータを推定する。

本研究では、Bass モデルを基本モデルとして、PJ の実施期間が有限であることと、クラウドファンディング資金調達パターンの特徴である、PJ 末期に購買が集中する傾向を組み込んだモデルの提案を行う。具体的には、(1)式に通常の Bass モデルでは表現しきれない PJ 末期の購買の集中を表現する関数

$$g_t = rt^k(t - s), \text{ ただし } s = \frac{k+1}{k+2} \qquad (2)$$

を追加することである。PJ 終了間際のかけこみ購買がどこからやってくるのかを考えたとき、PJ を認知しつつも、購買意思決定を保留し先送りする出資者の存在が想定される。(2)式の g_t はそのような、t 期の先送り出資者数（購買数）、つまり t 期の先送りした購買と先送りから購買に転じた人数の合計を表現する。r と k はそれぞれパラメータである。毎期存在するであろう先送りと先送りから出資に転じる購買は、時点 s において、先送りする出購買

数よりも出資に転じる購買数が超過する構成になっている。つまり先送り購買について、区間$0 \le t < s$においては$g_t < 0$として先送り超過を表現し、$s \le t \le T$で$g_t \ge 0$として購買超過を表現する。(2)式の挙動は図 1 のようになる。

一方で、時点 s を挟んで最終的に先送り超過の購買件数と購買超過の購買件数は等しくなる必要があり、$s \le t \le T$の購買件数と$0 \le t < s$における購買件数が等しいという制約を置く。つまり$|\int_0^s rt^k(t-s)dt| = \int_s^T rt^k(t-s)dt$を$s$について解き、(2)式にある$s = \frac{k+1}{k+2}$ を得る。また、(2)式で表現する PJ 末期の購買件数の集中 Sは

$$S = \left|\int_0^s g_t\,dt\right| = \int_S^{T(=1)} g_t dt = r\frac{(k+1)^{k+1}}{(k+2)^{k+3}} \tag{3}$$

となる。区間と Sの関係は図 2 である。

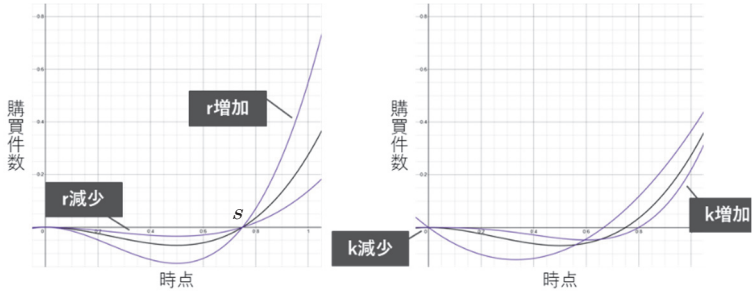


図 1 先送り関数g_tの挙動

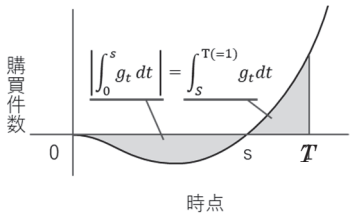


図 2 区間と Sの関係

Bass モデルに(2)式を組み込み、有限期間 Bass モデルとして

$$n_t - g_t = n_t - rt^k(t-s) = uf_t^* = up(1-F_t^*) + uqF_t^*(1-F_t^*) \tag{4}$$

を提案する。ここでf_t^*とF_t^*はそれぞれ、g_tによって PJ 末期$s \le t \le T$における購買の集中が$0 \le t < s$に振り戻された、t期における購買件数割合と、その累積割合である。

本提案モデルを異なる視点から解釈すると、Bass モデルでは各期の購買件数分布について単峰性の分布しか表現できない。したがって、PJ の初期と末期の 2 回購買が集中する傾向にあるクラウドファンディングの資金調達パターンを表現することが困難であり、提案モデルでは PJ 末期の購買をより早い時点（$0 \le t < s$）に振り戻すことで、Bass モデルの pと qのパラメータで表現しやすい形に資金調達パターンのグラフを変換している。

提案モデルは Bass モデルの微分方程式を利用して

$$n_t = uf_t^* + rt^k(t-s) = u\frac{p^*(p^*+q^*)^2\exp\{-(p^*+q^*)t\}}{[(p^*+q^*)\exp\{-(p^*+q^*)t\}]^2} + rt^k(t-s) \tag{5}$$

とし、R の optim 関数を用いた非線形回帰によって推定する。ただし(2)に含まれるrt^k項において rと kの識別性が低いため、0.75≦ s ≦0.975 ($<T$)の範囲で 0.025 間隔でのグリッドサーチを行う。

実証分析では日本の大手クラウドファンディングプラットフォームである Campfire に掲載された ALL-in 方式で実施された PJ の資金調達時系列推移データを使用する。PJ ごとに募集期間が異なり、$m\,(m = 1, \dots, M)$を各 PJ の日次暦上の各時点とすると、推定上の問題から、iを PJ として以下のように標準化してモデル推定する。

$$t_i = \frac{m_i}{M_i},\ \text{つまり}\quad t_i = \frac{1}{M_i}, \frac{2}{M_i}, \dots, T_i = \frac{M_i}{M_i} = 1 \tag{6}$$

モデルを 6,901 件の PJ データに適用した実証結果から、Bass モデルよりも提案モデルの方が適合度の高い結果を得たが不都合な結果もあり、詳細は当日報告する。

BTYD モデルによるユーザー投稿行動の分析

法政大学 長谷川翔平

はじめに

我々普段が利用するサービスには契約型と非契約型があり，契約型は顧客の契約終了によって企業側が顧客離脱を観測できるのに対し，非契約型は顧客離脱が観測できない。例えば，有料動画配信サービスは契約型に該当し，顧客が月額料金の支払いを停止すると運営企業は顧客の離脱を観測できる。一方で，EC サイトは非契約型に該当し，顧客がサービスの利用を止めても運営企業は顧客の離脱を観測できない。このような非契約型サービスの消費者行動を分析するモデルとして，BTYD（Buy Till You Die）モデルがある。

本研究では，経営科学系研究部会連合協議会主催の令和 3 年度データ解析コンペティションで提供されたママリのデータを用いる。ママリは，コネヒト株式会社が運営する妊娠から子育てまでの悩みを解決するママ向けアプリで，ユーザーが相互に質問・回答を行う Q&A サービスである。ママリの Q&A サービスを利用するにはユーザー登録を行う必要があるが，離脱顧客の多くは退会手続きを行わず，顧客離脱を観測できない場合が多いと考えられ，非契約型サービスの特徴を持つ。提供データにはユーザーの質問・回答の投稿日時と投稿文章が含まれており，本研究ではこれらのデータを用いてユーザーの投稿・離脱行動と関心の関係を分析する。

モデル

ユーザーの投稿・離脱行動は，BTYD モデルの代表的なモデルである Pareto/NBD モデル (Schmittlein et al. 1987) によって分析する。Pareto/NBD モデルでは，購買はランダムに発生すると仮定するが，本研究のデータには 1 日に複数回の投稿を行うユーザーも存在するため，モデル化する変数 x は投稿回数ではなく投稿日数とする。また，消費者の異質性を考慮するため，Abe (2009) で提案された階層ベイズモデルを採用する。

$$P[x_i|\lambda_i] = \begin{cases} \frac{(\lambda_i T_i)^{x_i}}{x_i!} e^{-\lambda_i T_i} & (\tau_i > T_i) \\ \frac{(\lambda_i \tau_i)^{x_i}}{x_i!} e^{-\lambda_i \tau_i} & (\tau_i \leq T_i) \end{cases}, \qquad f(\tau_i) = \mu_i e^{-\mu_i \tau_i}$$

ここで，$i(=1,\dots,N)$ は消費者，T_i は観測期間，τ_i は生存時間である。ユーザーの関心はそのユーザーが投稿した質問または回答の文章から LDA (Blei et al. 2003) によって抽出する。

$$\begin{aligned} z_{im} &\sim Multi\left(\theta_i\right), & \theta_i &\sim Dir(\alpha) \\ w_{im} &\sim Multi\left(\phi_{z_{im}}\right), & \phi_k &\sim Dir(\rho) \end{aligned}$$

ここで，z_{im} はユーザー i の投稿文章に含まれる単語 $w_{im}(m=1,\dots,M_i)$ のトピック $(k=1,\dots,K)$ である。投稿文章から LDA によって推定されるトピック分布 θ_i は K 個のトピックに対するユーザー i の関心と捉えることができる。本研究では，Pareto/NBD モデルのパラメータ

$\eta_i = (\log\lambda_i, \log\mu_i)'$ を目的変数，LDA のトピック分布 θ_i を説明変数とした階層モデルから，ユーザーの投稿・離脱行動と関心の関係を分析する。図 1 は本研究の分析モデルのグラフィカルモデルである。

$$\begin{bmatrix}\log\lambda_i \\ \log\mu_i\end{bmatrix} \equiv \eta_i = \beta\theta_i + e_i, \quad e_i \sim N(0, \Sigma)$$

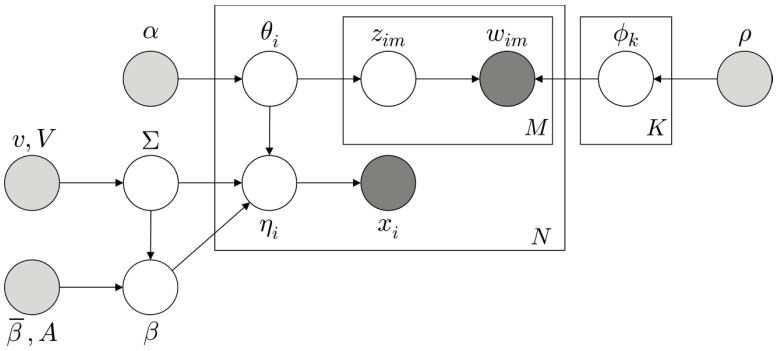


図 1　グラフィカルモデル

推定結果

質問投稿データからは「妊娠・出産」や「家族付き合い」などのトピックが抽出され，階層モデルのパラメータ β からは「子育て不安・不満」に関心の高いユーザーは投稿回数が多く，離脱も遅いことが示された。また，「妊娠・出産」に関心の高いユーザーは投稿日数が多いものの，離脱が早く，出産後に離脱している可能性が示された。ユーザーの出産前後に企業側からフォローを行うことで生存時間を延ばすことができると考えられる。回答投稿データからも「妊娠・出産」のトピックが抽出され，質問投稿と同様にこのトピックに関心の高いユーザーは投稿日数は多いものの離脱が早いことがことが示された。回答行動からの離脱を防ぐために各ユーザーの関心に合わせたインセンティブを導入することで，顧客の維持率を高められる可能性がある。

参考文献

1. Abe, M. (2009). “Counting Your Customers” One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model. *Marketing Science*, 28(3), 541–553.
2. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3 (Jan), 993–1022.
3. Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting Your Customers: Who-Are They and What Will They Do Next? *Management Science*, 33(1), 1–24.

Product Embedding for Large-Scale Disaggregated Sales Data

Yinxing Li and Nobuhiko Terui

Marketing data are expanding in several modes nowadays, as the number of variables explaining customer behavior has greatly increased, and automated data collection in the store has also led to the recording of customer choice decisions from large sample sizes. Thus, high-dimensional models have recently gained considerable importance in several areas, including marketing. Although some models, such as Prod2Vec, involve various marketing variables such as price and customer demographic data, the role of the variables in forecasting is still not discussed. In light of the limitations mentioned above, our study not only aims to propose a model with better forecasting precision but also to reveal how customer demographics affect customer behavior. we propose a Bayesian Word2Vec based framework that incorporates marketing variables and environment by considering following situations in our model.

When considering the market basket, our study incorporates the **receipt vector** into the model as the prior information of each purchased product in a basket, which means a preferred purchasing pattern for a certain shopping. It assumes that the customer will consider the whole purchasing context before choosing a product. We also assume a **state space model** for the receipt vector through the trips for each customer. We use the weekday, promotion information as the data for the state space model for higher interpretability of the prior structure such as the purchasing scenarios of each customer.

Besides, we consider the purchasing probability of a certain product conditional on an existing market basket is influenced by the following three factors - 1) **The compatibility with the marketing basket**, which is represented by the inner product of product vectors, 2) **customer utility** for the product, which incorporate the customer heterogeneity structure, and 3) **Thinking ahead** algorithm, which represents one-step ahead forecasting before purchasing the product.

Our proposed model contributes both to higher precision for forecasting by incorporating the marketing environment and customer heterogeneity into the model, and better interpretability. We use receipt data from a retailer for our empirical analysis, containing the information of customer demographic, promotion and other marketing information. We show not only the effectiveness of marketing environment for the forecasting by using the Hit Rate@K for the hold-out sample comparing to the several benchmark models , but also the high interpretability of our proposed model in the empirical study.

氏名・所属：〇武富 尚吾 1,2、　石垣 司 1、　花塚泰史 2、　森徹平 2
1.東北大学大学院経済学研究科
2.株式会社ブリヂストン

題目：待ち時間予測モデルを利用した *p*MP モデリングによる緊急ロードサービス施設配置最適化

要旨：
効率的で安全なモビリティ社会の実現のためには、突発的なトラブルにより走行不能となった車両を現場に急行して修理する緊急ロードサービスが重要である。このサービスレベルを向上させるためには、ユーザーからのレスキュー要請を受けてから復旧完了までの時間を短縮することが望ましい。この待ち時間の短縮を達成するためには、サービス拠点となる店舗の位置最適化が有効な手段となり得る。類似の問題として、需要地点と最寄りの施設間の移動距離/時間の総和または平均を最小化するために、複数の候補施設から p 個の施設を特定する p-median 問題（*p*MP）が様々な分野で検討されている。ただし、緊急ロードサービスの場合、車両のトラブル状況や、店舗特性（保有修理能力や営業形態）による出動準備時間や移動時間（両者の合計を待ち時間と呼ぶ）への影響を考慮する必要がある。そこで、本研究では、緊急ロードサービスの問題に特有の要因を考慮した緊急ロードサービス拠点の施設配置最適化法を提案する。株式会社ブリヂストンの緊急ロードサービスの実データに対して、ベイズ線形回帰モデルによる待ち時間予測モデルと *p*MP モデルを組み合わせることで、ユーザーの待ち時間を短縮する店舗配置の最適化を試みた。その結果、移動距離のみを考慮した従来手法と比較して、提案手法では待ち時間が最大 7%短縮されることが確認された。