

2023 年度科学研究費シンポジウム

統計科学・機械学習・情報数学の最前線

- 科学研究費補助金 基盤研究(A) 20H00576「大規模複雑データの理論と方法論の革新的展開」研究代表者：青嶋 誠(筑波大学)によるシンポジウム
- 日時：2024 年 1 月 26 日(金)～1 月 27 日(土)
- 場所：東北大学 青葉山キャンパス 情報科学研究科棟 2 階大講義室
- 開催責任者：荒木 由布子(東北大学)
- 内容・目的：複雑で多様な形態，サイズ，性質およびドメインを持つデータが観測・測定されるようになり，データ科学に基づく研究は多様な分野で急速に進化しています。その成果は学术界だけではなく，例えば根拠に基づく意思決定の手法の1つとしても，広く社会に浸透し始めています。本シンポジウムでは，データ駆動型研究でも特に，統計科学・機械学習・情報数学における新たな方法論の提案や最近の研究動向の紹介，教育に関連する報告，注目されている手法の解説とともに，応用統計学，バイオ統計学，材料科学，生命科学，経済学，ファイナンス，認知科学，社会科学，自然科学など，多岐にわたる科学分野における応用例の紹介や問題提起などを，広く対象としています。分野横断型のシンポジウムとすることで，参加者が各分野の情報を共有し，幅広い視点からの意見交換により，問題解決や新たな研究の発展につながることを期待します。

講演プログラム

1月26日(金)

10:00 ~ 10:05 オープニング

10:05 ~ 10:35 大石 峰暉 (東北大学)

Kick-one-out 法に基づいた回帰におけるクラスタリング

10:35 ~ 11:05 永井 勇 (中京大学)

GMANOVA モデルにおける様々な仮定が不要な罰則無推定法とその解釈

11:05 ~ 11:35 松井 秀俊 (滋賀大学), 寺田 吉壱 (大阪大学/理研 AIP)

可変ドメイン関数データに対する軌道の動的予測と気象データ分析への応用

11:35 ~ 13:00 昼休み

13:00 ~ 14:00 ポスターセッション

- ー 若山 智哉 (東京大学), 菅澤 翔之助 (慶應義塾大学), 小林 弦矢 (明治大学)
Similarity-based Random Partition Distribution for Clustering Functional Data
- ー 竹下 和慶 (大阪大学), 寺田 吉壱 (大阪大学/理研 AIP)
深層学習を用いた function-on-scalar 回帰とその理論的性質
- ー 岡崎 彰良 (九州大学), 川野 秀一 (九州大学)
頑健クラスタリングに基づくマルチタスク学習回帰
- ー Huang, Hongming (大阪大学), 鈴木 譲 (大阪大学)
An Efficient Procedure for Computing the Bayesian Network Structure Learning
- ー Liu Lirui (大阪大学), 鈴木 譲 (大阪大学)
An Information Criterion Improving WBIC and SBIC
- ー 池上 孝則 (放送大学)
データ量子統計の提案
- ー 松野 大輔 (千葉大学), 内藤 貫太 (千葉大学)
Semiparametric Density Estimation with Bregman Divergence
- ー 蔣 思澄 (千葉大学), 内藤 貫太 (千葉大学)
一般化 Kendall's tau 行列による主成分分析
- ー 森本 直紀 (東京理科大学), 八木 文香 (東京理科大学), 瀬尾 隆 (東京理科大学)
3-step 単調欠測データにおける T^2 型検定統計量を用いた平均ベクトルの検定
- ー Yao Sai (東北大学), 荒木 由布子 (東北大学)
Modal Interval Regression for Visualizing Data-frequent Areas
- ー Bat-Erdene, A. (Iwate University), Kawasaki, S (Iwate University)
Characterization of Asymptotic Distribution in Martingale Convergence of the Supercritical Branching Processes with Poissonian Offsprings

14:00 ~ 14:10 休憩

14:10 ~ 15:00 **特別講演 1** 山口 類 (愛知県がんセンター, 名古屋大学)

タンパク質結合能予測のためのマルチモーダル機械学習モデルの開発

15:00 ~ 15:30 竹内 努 (名古屋大学, 統計数理研究所), 本武陽一 (一橋大学), 西澤 淳 (岐阜聖徳学園大学, 名古屋大学), 村上 広椰 (名古屋大学), 加納 龍生 (名古屋大学), 馬 海霞 (名古屋大学), 内田 舜也 (名古屋大学), Suchetha COORAY (国立天文台科学研究部), 河野 海 (名古屋大学)

Application of the Topological Data Analysis to Cosmology (位相的データ解析の宇宙論への応用)

15:30 ~ 15:45 休憩

15:45 ~ 16:15 吉羽 要直 (東京都立大学)

代表的な接合関数の裾従属係数と裾次数

16:15 ~ 16:45 江村 剛志 (統計数理研究所)

コピュラ・グラフィック推定量—理論と応用—

16:45 ~ 17:35 **特別講演 2** 服部 聡 (大阪大学)

医学研究における不完全データの感度解析

18:00 ~ 20:00 懇親会

(東北大学青葉山キャンパス内 Buddy's Table)

1月27日(土)

9:30 ~ 10:00 石井 一夫 (公立諏訪東京理科大学), 金打 聖菜 (公立諏訪東京理科大学)

野菜の価格と気象との関連性について

10:00 ~ 10:30 本田 敏雄 (一橋大学)

Sparse Quantile Regression via ℓ_0 penalty

10:30 ~ 10:40 休憩

10:40 ~ 11:30 **特別講演 3** 今泉 允聡 (東京大学/理化学研究所)

インコンテキスト学習の統計学的解析

11:30 ~ 12:50 昼休み

12:50 ~ 13:20 徳田 悟 (九州大学)

対数周辺尤度とその陰的正則化のスケーリング挙動

13:20 ~ 13:50 川久保 友超 (千葉大学), 各務 和彦 (名古屋市立大学)

所得不平等度の分解の推定

13:50 ~ 14:00 休憩

14:00 ~ 14:30 中北 誠（理化学研究所 AIP）, 鳥谷部 智規（慶應義塾大学）, 中妻 照雄（慶應義塾大学）

Bayesian Analysis of Stochastic Conditional Duration Models with Intraday and Intra-deferred Future Seasonalities in High-frequency Commodity Market

14:30 ~ 15:00 渡名喜 庸蔵（大阪大学）

微小ノイズをもつ空間 2 次元線形放物型確率偏微分方程式モデルの
係数パラメータの推定

15:00 ~ 15:10 休憩

15:10 ~ 16:00 **特別講演 4** 松田 安昌（東北大学）

VAR Models for Surface Time Series

16:00 ~ 16:05 **クロージング**

Kick-one-out 法に基づいた回帰におけるクラスタリング

大石 峰暉

東北大学 データ駆動科学・AI 教育研究センター

n 次元目的変数ベクトル $\mathbf{y}$, $n \times k$ 説明変数行列 $\mathbf{Z}$ に対する以下のモデルを考える.

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \quad (1)$$

ここで, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_k)'$ は k 次元回帰係数ベクトル, $\boldsymbol{\varepsilon}$ は平均ベクトル $\mathbf{0}_n$, 共分散行列 $\sigma^2 \mathbf{I}_n$ を持つ n 次元誤差ベクトルであり, $\text{rank}(\mathbf{Z}) = k \leq n$ を仮定する. Kick-One-Out 法 (Zhao *et al.*, 1986; Bai *et al.*, 2018) に基づいた k 個の説明変数のクラスタリング法を提案する.

まずは通常の KOO 法について説明する. KOO 法は変数選択を行う手法で, フルモデル (すべての説明変数を用いたモデル) を基準として, 説明変数ごとにその変数が必要かどうかを判定する. c_0 をフルモデルの良さ, c_j ($j \in \{1, \dots, k\}$) をフルモデルから第 j 説明変数を除いたモデルの良さとする. ここで, モデルの良さはモデル選択規準などで測ることができる. このとき, 第 j 説明変数は $c_j > c_0$ であれば必要, $c_j \leq c_0$ であれば不要と判断する. つまり, KOO 法で選ばれる説明変数は以下のように定義される.

$$\{j \in \{1, \dots, k\} \mid c_j > c_0\}.$$

KOO 法では $(k+1)$ 回の計算で変数選択ができるため, 説明変数が多いときにも適用可能である. また, 様々な設定の下で変数選択の一致性が理論的に保証されている (e.g., Bai *et al.*, 2018; Oda & Yanagihara, 2020).

次に, 提案手法について説明する. 先に述べた通り, KOO 法はフルモデルと変数を 1 つ除いたモデルを比較し, その変数が必要かどうかを判定した. 提案手法では, フルモデルと 1 組の変数を結合したモデルを比較し, その結合を受理するかどうかを判定することでクラスタリングを行う. ここでのフルモデルとは, どの変数も結合されていないモデルを意味する. フルモデルにおける $\boldsymbol{\gamma}$ の最小二乗推定量, および当てはめ値はそれぞれ以下で与えられる.

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}, \quad \hat{\mathbf{y}} = \mathbf{Z}\hat{\boldsymbol{\gamma}}.$$

例えばフルモデルの良さを一般化 C_p 規準 (Atkinson, 1980) で測るとき, フルモデルの良さ c_0 は以下で与えられる.

$$c_0 = \frac{1}{s^2} \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \alpha k, \quad s^2 = \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{n - k}.$$

ここで, α はモデルの当てはまりに対する罰則の強さを表す正の値である. 次に結合の候補を考える. E を結合の候補となる添え字の組の集合とする. つまり, E は $\{1, \dots, k\}^2$ の部分集合で, $(j, \ell) \in E \Rightarrow j < \ell$ とする. 任意の $(j, \ell) \in E$ に対し, 第 j, ℓ 変数を結合したモデルとは, モデル (1) に $\gamma_j = \gamma_\ell$ の制約を課したモデルを意味する. この制約の下では, モデル (1) は以下のように書き換えられる.

$$\mathbf{y} = \mathbf{Z}_{j\ell}\boldsymbol{\gamma}_{j\ell} + \boldsymbol{\varepsilon}.$$

ここで, $\mathbf{Z}_{j\ell}$ は $\mathbf{Z}$ から第 ℓ 列ベクトルを除き第 j 列に加えた $n \times (k-1)$ 行列, $\boldsymbol{\gamma}_{j\ell}$ は $\boldsymbol{\gamma}$ から第 ℓ 成分を除いた $(k-1)$ 次元ベクトルである. このモデルに対して, $\boldsymbol{\gamma}_{j\ell}$ の最小二乗推定量, および当てはめ値はそれぞれ以下で与えられる.

$$\hat{\boldsymbol{\gamma}}_{j\ell} = (\mathbf{Z}_{j\ell}'\mathbf{Z}_{j\ell})^{-1}\mathbf{Z}_{j\ell}'\mathbf{y}, \quad \hat{\mathbf{y}}_{j\ell} = \mathbf{Z}_{j\ell}\hat{\boldsymbol{\gamma}}_{j\ell}.$$

よって1組の変数 (j, ℓ) を結合したモデルの良さ $c_{j\ell}$ は、一般化 C_p 規準を用いると以下で与えられる.

$$c_{j\ell} = \frac{1}{s^2} \|\mathbf{y} - \hat{\mathbf{y}}_{j\ell}\|^2 + \alpha(k-1).$$

以上のように定義された $c_0, c_{j\ell}$ に対し、最適な結合を以下で定義する.

$$\hat{E} = \{(j, \ell) \in E \mid c_{j\ell} \leq c_0\}.$$

モデル選択規準はその値が小さいほど良いモデルであると判断できる. したがって, $c_{j\ell} < c_0$ であれば2つの変数の結合がモデルを改善することからその結合を受理し, $c_{j\ell} > c_0$ であればモデルが改悪されることからその結合を受理しない. このようなルールに従って変数の組ごとにそれらを結合するかどうかを判定し, その結果を統合することで最終的なクラスタリング結果を得ることができる.

本発表では、特に真のクラスターの選択に関する一致性に着目し、提案手法と Ohishi *et al.* (2021) による一般化 fused Lasso を用いた手法を数値的に比較した. 一般化 C_p 規準の α として以下を用いたとき、提案手法は真の結合の選択に関する一致性を有することが数値的に確認された.

$$\alpha = \frac{n-k}{n-k-2} + \beta, \quad \beta = \frac{(n-k)\sqrt{n-k-3}}{(n-k-2)\sqrt{n-k-4}} \cdot k^{1/4} \log n.$$

上記の α は Oda & Yanagihara (2020) で示された変数選択の一致性を有する値である.

参考文献

- Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika*, **67**, 413–418.
- Bai, Z. D., Fujikoshi, Y. & Hu, J. (2018). Strong consistency of the AIC, BIC, C_p and KOO methods in high-dimensional multivariate linear regression. Technical Report TR-No. 18-09. Hiroshima Statistical Research Group. Hiroshima.
- Oda, R. & Yanagihara, H. (2020). A fast and consistent variable selection method for high-dimensional multivariate linear regression with a large number of explanatory variables. *Electron. J. Stat.*, **14**, 1386–1412.
- Ohishi, M., Fukui, K., Okamura, K., Itoh, Y. & Yanagihara, H. (2021). Coordinate optimization for generalized fused Lasso. *Comm. Statist. Theory Methods*, **50**, 5955–5973.
- Zhao, L. C., Krishnaiah, P. R. & Bai, Z. D. (1986). On detection of the number of signals in presence of white noise. *J. Multivariate Anal.*, **20**, 1–25.

GMANOVA モデルにおける 様々な仮定が不要な罰則無推定法とその解釈

中京大学 教養教育研究院 永井 勇

n 個の各個体に対して、時間と共に測定されたデータは経時測定データと呼ばれ、非常に多くの分野で分析されている。このような経時測定データにおいて、全ての個体で p 回測定しその測定時点が揃っている場合、Potthoff and Roy (1964) により提案された次の一般化多変量分散分析 (GMANOVA) モデルでの分析がよくされている;

$$Y = \mathbf{1}_n \mu' X' + A \Xi X' + \mathcal{E}, \quad (1)$$

ここで、 Y は各行が各個体の経時測定データからなる $n \times p$ 既知行列、 $\mathbf{1}_n$ は全ての要素が 1 の n 次元のベクトル、 μ は q 次元未知ベクトル、 X は個体内計画行列と呼ばれる測定時点から作られる $p \times q$ 既知行列 (詳細は後述)、 A は個体間計画行列と呼ばれる測定時点に無関係な各個体の特徴を表す説明変数からなる $n \times k$ 既知行列で各列で中心化されているとし、 Ξ は $k \times q$ 未知行列であり、 $\mathcal{E}$ は $E[\mathcal{E}] = \mathbf{0}_n \mathbf{0}'_p$ で $\text{Var}[\text{vec}(\mathcal{E})] = \Sigma \otimes I_n$ の $n \times p$ 誤差行列であり、 $\mathbf{0}_r$ は全ての要素が 0 の r 次元ベクトル、 Σ は未知の $p \times p$ 正則行列である。

Y の経時的な変動 (経時変動) を推定する際には、個体内計画行列 X の i 行目として、 i 番目の測定時点 t_i ($i = 1, \dots, p; t_1 < \dots < t_p$) の関数からなる行列が用いられる。例えば X の i 行目として $(t_i^{q-1}, \dots, t_i^0)$ を用いて未知の μ や Ξ を推定することで、全ての個体で共通の経時変動 や 各説明変数に対応する経時変動 の切片と各次数の係数がそれぞれ推定でき、それらの結果として目的の経時変動を測定時点 t_i の $(q-1)$ 次多項式で推定できる。また、 X の i 行目として t_i の別の関数からなるものを用いて、 μ や Ξ を推定することで、用いた関数の重み付き和の形で全体の経時変動を推定できる。

このモデル (1) において、 μ や Ξ の推定は次のリスクを最小にすることで行われる;

$$\text{tr}\{ (Y - \mathbf{1}_n \mu' X' - A \Xi X') \Sigma^{-1} (Y - \mathbf{1}_n \mu' X' - A \Xi X')' \}. \quad (2)$$

このリスクを最小にする μ や Ξ を求める際に、従来は $\text{rank}(A) = k$ かつ $\text{rank}(X) = q$ を仮定している。これらの仮定は暗に $k \leq n$ や経時変動の推定に用いる t_i の関数への制限をしている。一方で、これらの仮定が満たされない場合、以下のような手法がある:

- リスク (2) に罰則を加えたものを最小化する罰則付き推定法 (例えば Katayama & Imori (2014) 参照) を用いる
- Yanagihara ら (2023) などの情報量規準を用いて、 A や X からいくつかの列を選択し、選択したものだけを推定に用いる
- Moore-Penrose 型一般逆行列 (例えば Srivastava (2002) A.6) を用いる

本講演では、永井 (2021) のアイデアを用い、 $\text{rank}(A) < k$ かつ $\text{rank}(X) < q$ の下で、リスク (2) を最小にする推定量を提案する。さらに、本講演で提案する推定量の解釈などについても触れる。数値実験を通じた比較なども講演で触れる予定である。

引用文献:

- [1] Katayama, S. & Imori, S. (2014) Lasso penalized model selection criteria for high-dimensional multivariate linear regression analysis. *J. Multivariate Anal.*, **132**, 138–150.
- [2] 永井 勇 (2021) 高次元小標本における多変量線形回帰モデルでの推定法, 2021 年度統計関連学会連合大会 講演.
- [3] Potthoff, R. F. & Roy, S. N. (1964) A generalized multivariate analysis of variance model usefule specially for growth curve problems. *Biometrika*, **51**, 313–326.
- [4] Srivastava, M. S. (2002) *Methods of Multivariate Statistics*, Wiley-Interscience.
- [5] Yanagihara, H., Nagai, I., Fukui, K. & Hijikawa, Y. (2023) Modified C_p criterion in widely applicable models. *Smart Innov. Syst. Tec.*, **352**, Intelligent Decision Technologies 2023: Proceedings of the 15th KES International Conference on Intelligent Decision Technologies KES-IDT-2023 (eds. I. Czarnowski, R. J. Howlett & L. C. Jain), 173-182.

可変ドメイン関数データに対する軌道の動的予測と 気象データ分析への応用

滋賀大学 データサイエンス学部 松井 秀俊
大阪大学 大学院基礎工学研究科 / 理化学研究所 AIP 寺田 吉彦

1. 概要

観測個体がそれぞれ経時的に測定されたデータを関数化処理し、得られた関数データ集合を対象として分析を行う方法は関数データ解析とよばれ、医学や工学などさまざまな分野で応用されている (Kokoszka & Reimherr, 2017). 関数データ解析では、観測される関数データの定義域は観測個体間で同じという状況が多い。これに対して本研究では、定義域の終端となる時点が個体ごとに異なるデータを扱う。このようなデータを、ここでは可変ドメイン関数データとよぶ。たとえば、日本付近を通過した一部の台風についての発生から消滅までの中心気圧の推移は台風が消滅するまでの時間が台風毎に異なるため、従来の関数データとして捉えることができない (図 1)。また、

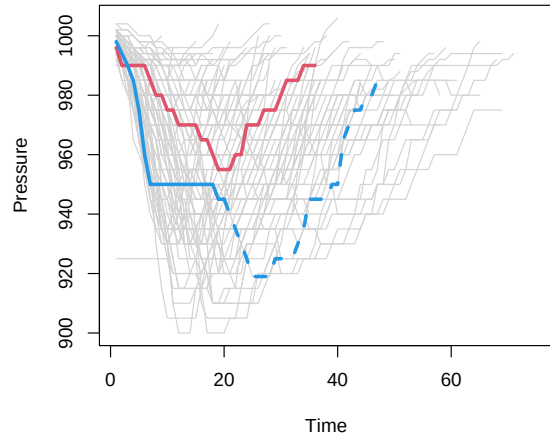


図 1: 台風の中心気圧 (出典：デジタル台風)。
台風の発生時点を 0 で揃えている。

一般的な傾向として、台風は消滅までの時間が長いものほど中心気圧の最小値は低く、短いものほど高い。可変ドメイン関数データは、その特性が定義域の長さに応じて異なることから、定義域の長さを考慮に入れた分析が行われている (Geller et al., 2012; Johns et al., 2019).

本研究では、1 つの特徴についての可変ドメイン関数データが得られているとき、新たな観測個体について経時測定データが途中まで観測されている状況で、その先の軌道および終端の時点を予測する方法について検討する。図 1 の例では、これまでに赤や灰の曲線に対応する台風が観測され、さらに青の曲線に対応する台風が実線部分まで観測されたとき、その先の破線部分の軌道と、その終端時点を予測するという問題を考える。そのために、可変ドメイン関数データに対して従来の関数データに対する動的予測の考え方 (Chiou, 2012) を適用する。提案手法の有効性を、人工データおよび実データの分析を通して検証する。

2. 可変ドメイン関数データの動的予測

いま、始点を 0、終点を確率変数 $M \in (0, 1]$ としたランダム区間 $\mathcal{I} = [0, M]$ 上のランダム関数 $Y(t)$ ($t \in \mathcal{I}$) を考える。終端 $M = m$ が与えられた下で、 $Y(t)$ は区間 $[0, m]$ 上の平均二乗連続な確率過程となることを仮定する。さらに、 $\mathbf{X}$ を共変量を表す確率変数ベクトル、その実現

値を $\mathbf{x}$ とする．このとき， $Y(t)$ の平均関数および共分散関数を次で表す．

$$\begin{aligned}\mu(t \mid m, \mathbf{x}) &= E[Y(t) \mid M = m, \mathbf{X} = \mathbf{x}], \\ \gamma(t, s \mid m, \mathbf{x}) &= E[\{Y(t) - \mu(t \mid m, \mathbf{x})\} \{Y(s) - \mu(s \mid m, \mathbf{x})\} \mid M = m, \mathbf{X} = \mathbf{x}].\end{aligned}$$

あるランダム関数 Y の軌道が，ある時点 $\tau \in \mathcal{I}$ について $\mathcal{S} = [0, \tau]$ の区間でのみ観測された状況を考える． $\mathcal{S}$ の区間での軌道を表すランダム関数を $Y_{\mathcal{S}}$ ，実際に観測された軌道を $y_{\mathcal{S}}$ と記す．このとき， τ まで観測されているランダム関数 Y の終端 M を，次の条件付き期待値で予測する．

$$\widehat{M} = E[M \mid M > \tau, Y_{\mathcal{S}} = y_{\mathcal{S}}, \mathbf{X} = \mathbf{x}].$$

実際には，終端が $m > \tau$ なる観測データに対して，区間 $\mathcal{S}$ での関数データ $y(t)$ および共変数 $\mathbf{x}$ を説明変数，対応するデータの終端 m を目的変数として，ノンパラメトリック回帰などを用いる．そして， $\mathcal{S}$ 上でのみ観測されている Y の終端 M を予測する．その予測値を $\widehat{m}$ と記す．

次に， τ より先の区間 $\mathcal{T} = (\tau, \widehat{m}]$ における $Y(t)$ の軌道 $Y_{\mathcal{T}}$ を，次で予測する．

$$\widehat{Y}_{\mathcal{T}} = E[Y_{\mathcal{T}} \mid M = \widehat{m}, Y_{\mathcal{S}} = y_{\mathcal{S}}, \mathbf{X} = \mathbf{x}].$$

この条件付き期待値を計算するにあたり， $M = m$ と $\mathbf{X} = \mathbf{x}$ が与えられた下でランダム関数 $Y(t)$ が平均関数 $\mu(t \mid m, \mathbf{x})$ ，共分散関数 $\gamma(t, s \mid m, \mathbf{x})$ のガウス過程に従うと仮定する．このとき， $Y(t)$ を離散時点で評価し，区間 $\mathcal{S}, \mathcal{T}$ で分割したベクトル $\mathbf{y} = (\mathbf{y}_{\mathcal{S}}^{\top}, \mathbf{y}_{\mathcal{T}}^{\top})^{\top}$ は次の平均ベクトル，分散共分散行列をもつ多変量正規分布に従うとする．

$$\begin{pmatrix} \mathbf{y}_{\mathcal{S}} \\ \mathbf{y}_{\mathcal{T}} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \boldsymbol{\mu}_{\mathcal{S}}(\widehat{m}, \mathbf{x}) \\ \boldsymbol{\mu}_{\mathcal{T}}(\widehat{m}, \mathbf{x}) \end{pmatrix}, \begin{pmatrix} \Gamma_{\mathcal{S}\mathcal{S}}(\widehat{m}, \mathbf{x}) & \Gamma_{\mathcal{S}\mathcal{T}}(\widehat{m}, \mathbf{x}) \\ \Gamma_{\mathcal{T}\mathcal{S}}(\widehat{m}, \mathbf{x}) & \Gamma_{\mathcal{T}\mathcal{T}}(\widehat{m}, \mathbf{x}) \end{pmatrix} \right).$$

そして， $Y_{\mathcal{T}}(t)$ を離散時点で評価したベクトル $\mathbf{y}_{\mathcal{T}}$ を，次の事後平均によって予測する．

$$\widehat{\mathbf{y}}_{\mathcal{T}} = \boldsymbol{\mu}_{\mathcal{T}}(\widehat{m}, \mathbf{x}) + \Gamma_{\mathcal{T}\mathcal{S}}(\widehat{m}, \mathbf{x}) \Gamma_{\mathcal{S}\mathcal{S}}^{-1}(\widehat{m}, \mathbf{x}) \{\mathbf{y}_{0,\mathcal{S}} - \boldsymbol{\mu}_{\mathcal{S}}(\widehat{m}, \mathbf{x})\}.$$

平均ベクトルや分散共分散行列は，Johns et al. (2019) の方法を利用して推定する．

提案した手法を台風のデータ分析に適用し，途中まで観測された台風の中心気圧の推移から，その先の軌道と，台風が消滅する時点を予測する．結果については，当日報告する．

参考文献

- Chiou, J. M. (2012). Dynamical functional prediction and classification, with application to traffic flow prediction. *Ann. Appl. Statist.* **6**, 1588–1614.
- Gellar, J., Colantuoni, E., Needham, D. M., and Crainiceanu, C. (2014). Variable-domain functional regression for modeling ICU data. *J. Amer. Stat. Assoc.* **109**, 1425–1439.
- Johns, J. T., Crainiceanu, C., Zipunnikov, V., and Gellar, J. (2019). Variable-domain functional principal component analysis. *J. Comput. Graph. Stat.* **28**, 993–1006.
- Kokoszka, P. and Reimherr, M. (2017). *Introduction to functional data analysis*. CRC Press, Boca Raton.

Similarity-based Random Partition Distribution for Clustering Functional Data

若山智哉
東京大学経済学研究科

菅澤翔之助
慶應義塾大学経済学部

小林弦矢
明治大学商学部

1 はじめに

過去数十年にわたるモバイルデバイスの急速な普及により、人口データは現代社会の理解や政策立案において重要な役割を担っている。具体的には、都市計画、交通計画、医療サービスの提供、広範囲な政策立案など、多岐にわたる分野に影響を及ぼしている [1, 3, 5]。この発表では、人口データを深く理解するための確率分割を提案し、その応用について説明した。

2 データと背景

クラスタリングは人口データから洞察を引き出すために有効である。クラスタ内の共通点を特定し、地域特性を明らかにすることで、ライドシェアサービスや街頭広告などの産業と、都市工学や人文科学などの研究領域の両方に貢献することができる。特に、モデルベースのクラスタリングは、ノンパラメトリックベイズ手法と相性が良い。ディリクレ過程 (DP) に基づくクラスタリングは、自動的にクラスタ数を決定し、空間構造を組み込むことも容易である。[2] の研究では、Pitman–Yor 過程の中で既存のクラスターを新しい項目に割り当てる際に地理的な近接性を統合しており、広く応用されている。特に、人口データ解析でも隣接情報が重要であり、図 1 に示すように、隣接する地区間には高い相関が見られる。

にもかかわらず、我々のデータセットへの既存のクラスタリング手法の適用には、その多変量性と時間的情報の 2 つの特性が障壁となっている。多変量クラスタリング問題に取り組む際、クラスタ数を固定しない手法を単純に適用することはリスクを伴う。高次元データはその複雑性から、容

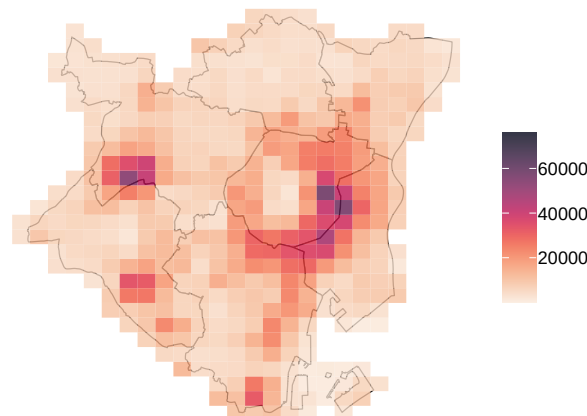


図 1 2019 年 1 月 29 日 14 時の東京都の中心地区における人口マップ

易に分類可能であり、これは教師あり分類タスクでは役立つ特徴である [4]。一方で、この高い分類可能性は、教師なしクラスタリングではクラスタ数が過剰になる可能性がある。特に、この性質は DP が小さいクラスターサイズを好む傾向を推進し、余計なクラスターを作ることになり、データの理解を妨げることになる。また、人口データの解析では、時間的な構造も重要である。地域によっては平日と休日では人口の増減が大きく異なるため、グループ構造は日毎に変化しており、典型的な平日の傾向だけに注目すると重要な洞察を見逃してしまう可能性がある。

3 方針

本研究では一般化ディリクレ過程 (GDP) を基にした新たな確率分割手法、SGDP を提案する。この手法は、観測された時間毎の人口フローを関数（確率過程の実現値）としてクラスタリングすることを目的としている。SGDP では、一般化パラメトリゼーションを用いることで過剰なクラスタの形成リスクを低減し、同時に地理的隣接性を考慮しており、この方法の空間的な性質についても理論的な裏付けがある。さらに、観測値とクラスタ平均との間の差異をガウス過程でモデル化する。これは、ある時刻での観測された人口がクラスタ平均から逸脱する場合、その影響が連続する時刻にも及ぶという考え方に基づいており、単なる独立したノイズとして処理するよりも、過剰なクラスタ形成を抑制する効果がある。また、この方法は時間的な変動にも柔軟に対応できるため、各地域の時間によって変わるパターンをより詳細に捉え、理解することが可能になる。

4 結果

データ解析の結果や手法の有効性はポスターセッションで議論した。

参考文献

- [1] Amir Ahmadi-Javid, Oded Berman, and Pooya Hoseinpour. Location and capacity planning of facilities with general service-time distributions using conic optimization. *arXiv preprint arXiv:1809.00080*, 2018.
- [2] David B Dahl, Ryan Day, and Jerry W Tsai. Random partition distribution indexed by pairwise information. *Journal of the American Statistical Association*, 112(518):721–732, 2017.
- [3] Antonio Páez and Darren M Scott. Spatial statistics for urban analysis: A review of techniques with examples. *GeoJournal*, 61:53–67, 2004.
- [4] Tomoya Wakayama and Masaaki Imaizumi. Fast convergence on perfect classification for functional data. *Statistica Sinica*, 34(4), 2024.
- [5] Mingshu Wang and Lan Mu. Spatial disparities of uber accessibility: An exploratory analysis in atlanta, usa. *Computers, Environment and Urban Systems*, 67:169–175, 2018.

深層学習を用いた function-on-scalar 回帰とその理論的性質

竹下 和慶¹, 寺田 吉彦^{1,2}

¹ 大阪大学大学院基礎工学研究科, ² 理化学研究所革新知能統合研究センター

1. はじめに

ある領域や区間上で連続的に観測されたデータをランダムな関数の実現値として捉えたデータ解析は関数データ解析 (Functional Data Analysis; FDA) と呼ばれている. 本研究では, 入力データ (予測変数) が従来の多変量データで, 出力 (応答変数) が関数データであるような Function-On-Scalar (FOS) 回帰と呼ばれる関数回帰分析を考える.

多くの既存研究は, 単純で扱いやすい線形な FOS 回帰に焦点を当てており, 非線形な FOS 回帰についての研究結果は少ない. 例えば, Scheipl et al. (2015) は予測変数に関して非線形な関数型加法モデル (Functional Additive Mixed Model; FAIMM) を提案している. FAIMM では, 応答変数が予測変数の一部を引数とした非線形関数の和として表現できると仮定している. この手法は, 従来の線形 FOS 回帰よりも高い柔軟性を有するモデルである. しかし, 仮定した加法モデルがデータに適合しない場合には推定精度が低下してしまうという問題がある.

Luo et al. (2023) は Scheipl et al. (2015) の問題点を解消するため, より柔軟なモデリングが可能となるニューラルネットワークに基づく手法を提案した. また, 従来の万能近似定理を拡張し, 予測変数に関して連続な関数に対してこのモデルの万能近似性を示している. この手法の利点は, 従来の FAIMM が基底関数を用いて真の関数を直接的に推定するのに対し, この手法では予測変数の次元数に基づいた複数の二変数関数を推定するのみで良いという点である. このことによって, FAIMM よりも計算コストが低く, 予測変数の次元が比較的大きい場合でも適用可能である.

しかし, この方法には, 2つの問題点がある. 1つ目は, 真の関数が予測変数に関する連続性を仮定している点である. 2つ目は, 提案された推定関数自身が積分形で記述されているため実用上厳密にモデル化することが不可能な点である. そのため, 実際に手法を適用する際には, 時間と積分に関して離散化し基底関数を用いて推定している. Luo et al. (2023) の Theorem 3 では, 推定量の収束レートが導出されているが, 証明に重大なミスがあり, 正しい収束レートを導出できていない. 本発表では, 深層ニューラルネットワークを用いたより自然な FOS 回帰分析を提案し, その推定量の収束レートを導出する.

2. 提案手法と主結果

本研究では, Luo et al. (2023) の問題点を解決した深層ニューラルネットワークを用いたより直接的な推定量を提案する. また, 真の回帰関数 f^* が非等方的 Besov 空間 (Anisotropic Besov space) に含まれるという仮定の下で, 推定量の理論的性質を導出する. 本研究で扱う非等方的 Besov 空間は, 方向によって滑らかさが異なるような複雑な関数を含む関数空間であり, Hölder 空間などを特別な場合として含んでいる. 従って, Luo et al. (2023) よりも柔軟な設定で議論することが可能である. また, 非等方的 Besov 空間において深層ニューラルネットワークは適応的な近似能力を有し, 通常のノンパラメトリック回帰問題の設定において, 深層ニューラルネットワークを用いた最小二乗型推定量は, minimax 最適性をもつことが知られている (Suzuki and Nitanda, 2021).

予測変数の空間は d 次元とする. このとき, 入力層のニューロンが $d+1$ 個, 層の個数が L 個, 各層のニューロンが W 個, 出力層のニューロンが 1 個の順伝搬深層 ReLU ニューラルネットワーク

のクラス $\Phi(L, W, S, B)$ を用いて f° を推定することを考える.

$$\Phi(L, W, S, B) := \left\{ (\mathcal{W}^{(L)}\eta(\cdot) + b^{(L)}) \circ \cdots \circ (\mathcal{W}^{(1)}x + b^{(1)}) \mid \mathcal{W}^{(L)} \in \mathbb{R}^{1 \times W}, b^{(L)} \in \mathbb{R}, \right. \\ \left. \mathcal{W}^{(1)} \in \mathbb{R}^{W \times (d+1)}, b^{(1)} \in \mathbb{R}^W, \mathcal{W}^{(l)} \in \mathbb{R}^{W \times W}, b^{(l)} \in \mathbb{R}^W (1 < l < L), \right. \\ \left. \sum_{l=1}^L (\|\mathcal{W}^{(l)}\|_0 + \|b^{(l)}\|_0) \leq S, \max_l \|\mathcal{W}^{(l)}\| \vee \|b^{(l)}\|_\infty \leq B \right\}$$

ここで, $\|\cdot\|_0$ は非ゼロ成分の個数を表し, $\|\cdot\|_\infty$ は各成分の絶対値の最大値とする. また, S はスパース定数, B はノルム定数である.

予測変数と応答変数の訓練データ $D_n = \{(\mathbf{X}_l, Y_l)\}_{l=1}^n$ が与えられたもとで, f° に対する推定量 $\hat{f}_n \in \Phi(L, W, S, B)$ を次の目的関数の最小化によって定める.

$$\hat{f}_n = \operatorname{argmin}_{f \in \Phi(L, W, S, B)} \frac{1}{n} \sum_{l=1}^n \int_0^1 \{Y_l(t) - \bar{f}(\mathbf{X}_l, t)\}^2 dt$$

ここで, $\bar{f}$ は f のクリッピングであり, ある定数 $F > 0$ に対して, $\bar{f} = \min\{\max\{f, -F\}, F\}$ と定められる.

ニューラルネットワークの大きさを適切に定めることで, 従来の非線形回帰問題 (Suzuki and Nitanda, 2021) と同様に, 次の推定誤差レートが得られる.

$$E_{D_n} \left[\int_0^1 \left\| \hat{f}_n(\mathbf{X}, t) - f^\circ(\mathbf{X}, t) \right\|_{L^2(P_{\mathbf{X}})}^2 dt \right] \lesssim n^{-\frac{2\tilde{\beta}}{2\tilde{\beta}+1}} (\log n)^3$$

ここで, $\|\cdot\|_{L^2(P_{\mathbf{X}})}$ は予測変数の訓練データが従う確率密度関数で重みづけた L^2 ノルム, $E_{D_n}[\cdot]$ は訓練データ D_n の分布に対する期待値を表す. また, $\tilde{\beta}$ は非等方的 Besov 空間の各軸方向における滑らかさパラメータの調和平均である. この結果により, 収束レートは $\log$ 項を無視すると $n^{-\frac{2\tilde{\beta}}{2\tilde{\beta}+1}}$ であり, 滑らかさのパラメータ $\tilde{\beta}$ が大きければ収束レートが早くなることがわかる. さらに, この収束レートは直接的に次元 d に依存していないため, 次元の呪いを緩和できる可能性が示唆される. 具体的には, 真の関数が予測変数の空間で複雑であったとしても, 時間方向に対して十分な滑らかさを有する場合には, 良い収束レートが期待できる. 提案手法とその理論的性質の詳細や数値実験は当日報告する.

参考文献

- Scheipl, F., Staicu, A.-M. and Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2), 477–501.
- Luo, Ruiyan and Qi, Xin (2023). Nonlinear function-on-scalar regression via functional universal approximation. *Biometrics*, 79(4), 3319–3331
- Taiji Suzuki and Atsushi Nitanda (2021). Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. *Advances in Neural Information Processing Systems* 34, 3609–3621

頑健クラスタリングに基づくマルチタスク学習回帰

九州大学 大学院数理学府 岡崎 彰良
九州大学 大学院数理学府 川野 秀一

1. はじめに

マルチタスク学習とは、タスクと呼ばれる複数の関連していると考えられるデータ集合が得られている際に、それらのタスクそれぞれに異なるモデルを設定し、モデル間に共通した構造の推定を行うことでモデル全体の推定精度を改善する統計手法である (Caruana, 1997). マルチタスク学習を実行する際、データによってはタスク間に共通する構造が単一ではないと考えられる場合が存在する (Xue et al., 2007). このような状況におけるマルチタスク学習は、共通構造を共有するタスク集合の分類と共通成分の推定を同時に行うため、モデルのパラメータに対するクラスタリングにより実現される. また、マルチタスク学習では、用いられるタスクすべてが何らかの関連性を有していることを仮定したもとで実行されるが、実際にはどのタスクとも関連性を有さない外れ値タスクが含まれている状況が考えられる (Gong et al., 2012).

Okazaki and Kawano (2023) では、各タスクに設定された回帰モデルに関する損失関数に対し、凸クラスタリングに基づく正則化項を加えることで、凸最適化の枠組みにおいてクラスタリングによるマルチタスク学習を提案した. しかし、この手法ではすべてのタスクがいずれかのクラスに属するという仮定のもと推定が行われるため、外れ値タスクがタスク集合に含まれている場合にクラスタリングの精度が悪化する問題が存在する.

そこで本報告では、凸クラスタリングによるマルチタスク学習 (Okazaki and Kawano, 2023) において、外れ値タスクの影響を軽減すると同時に外れ値タスクの検出を行う手法を提案する. 各タスクに対して外れ値を表すパラメータを導入し、グループスパース正則化 (例えば Yuan and Lin (2006)) によりそのパラメータの選択を行う. 提案手法は、タスクのクラスタリングを担う凸クラスタリングにおける損失関数を、外れ値に対して頑健な多変量損失関数 (例えば Hampel et al. (1986)) で置き換えた手法として解釈でき、タスクが構成するクラス推定において外れ値タスクの影響に対して頑健であることが期待できる. 提案手法への有効性を、モンテカルロ・シミュレーションおよび実データへの適用を通して検証する.

2. 凸クラスタリングに基づくマルチタスク学習回帰モデル

いま、 T 個の関連すると思われるタスクが与えられているとし、 m ($m = 1, \dots, T$) 番目のタスクに対し、 n_m 個の説明変数に関する p 次元データ $\{\mathbf{x}_{mi}; i = 1, \dots, n_m\}$ および n_m 個の目的変数に関するデータ $\{y_{mi}; i = 1, \dots, n_m\}$ の組 $\{(y_{mi}, \mathbf{x}_{mi}); i = 1, \dots, n_m\}$ が独立に得られているとする. また、 m 番目のタスクに関する目的変数ベクトルを $\mathbf{y}_m = (y_{m1}, \dots, y_{mn_m})^\top \in \mathbb{R}^{n_m}$, 計画行列を $\mathbf{X}_m = (\mathbf{x}_{m1}, \dots, \mathbf{x}_{mn_m})^\top \in \mathbb{R}^{n_m \times p}$ とおく. これに対し、 T 個のタスクの組を $\{(\mathbf{y}_m, \mathbf{X}_m); m = 1, \dots, T\}$ とする.

次に、 T 個のタスクに対して、目的変数と説明変数の関係を次の T 個の線形回帰モデル

$$\mathbf{y}_m = \mathbf{X}_m \mathbf{w}_m + \boldsymbol{\epsilon}_m, \quad m = 1, \dots, T \quad (1)$$

により捉える. ここで、 $\mathbf{w}_m = (w_{m1}, \dots, w_{mp})^\top$ は m 番目のタスクに対応する回帰係数ベクトルであり、 $\boldsymbol{\epsilon}_m$ は各要素がそれぞれ独立に $N(0, \sigma^2)$ に従う観測誤差ベクトルである. (1) 式に対して、Okazaki and Kawano (2023) では、次の最小化問題

$$\min_{\mathbf{w}_m, \mathbf{u}_m \in \mathbb{R}^p, m=1, \dots, T} \left\{ \sum_{m=1}^T \frac{1}{2n_m} \|\mathbf{y}_m - \mathbf{X}_m \mathbf{w}_m\|_2^2 + \frac{\lambda_1}{2} \sum_{m=1}^T \|\mathbf{w}_m - \mathbf{u}_m\|_2^2 + \lambda_2 \sum_{(m,l) \in \mathcal{E}} r_{m,l} \|\mathbf{u}_m - \mathbf{u}_l\|_2 \right\} \quad (2)$$

を提案した. ここで、 $\mathbf{u}_m \in \mathbb{R}^p$ は m 番目のタスクが属するクラスタの重心に関するパラメータ、 $\mathcal{E}$ はタスクの組の集合、 λ_1, λ_2 は正の値を取る正則化パラメータ、 $r_{m,l}$ は m 番目と l 番目のタスク間の関係を表す係数であり 0 以上の値を取る. 目的関数の第二項目および第三項目は凸クラスタリング (Hocking et al., 2011) に基づく正則化項であり、 $\mathbf{w}_m$ に関してクラスタリングを実行する. 同じクラスに属するタスクは $\mathbf{u}_m$ が等し

い値として推定される．すなわち， $\mathbf{u}_m$ は各タスクが属するクラスタにおける回帰係数ベクトルの共通成分を表す．

3. 外れ値タスクに頑健な凸クラスタリング回帰モデル

(1) 式に対して，次の最小化問題

$$\min_{\mathbf{w}_m, \mathbf{u}_m, \mathbf{o}_m \in \mathbb{R}^p, m=1, \dots, T} \left\{ \sum_{m=1}^T \frac{1}{2n_m} \|\mathbf{y}_m - \mathbf{X}_m \mathbf{w}_m\|_2^2 + \frac{\lambda_1}{2} \sum_{m=1}^T \|\mathbf{w}_m - \mathbf{u}_m - \mathbf{o}_m\|_2^2 + \lambda_2 \sum_{(m,l) \in \mathcal{E}} r_{m,l} \|\mathbf{u}_m - \mathbf{u}_l\|_2 + \lambda_3 \sum_{m=1}^T P(\mathbf{o}_m, \eta) \right\} \quad (3)$$

を考える．ここで， $\mathbf{o}_m \in \mathbb{R}^p$ は m 番目の外れ値タスクに関するパラメータ， λ_3 は正の値を取る正則化パラメータである．第四項目は $\mathbf{o}_m$ に関して変数選択を行う正則化項であり， η は縮小される値を調整するパラメータである．第四項目により， $\mathbf{o}_m \neq \mathbf{0}$ として推定された場合， m 番目のタスクは他タスクと共通性を有さない外れ値タスクとして解釈される．第四項目にグループ lasso 正則化 (Yuan and Lin, 2006) を用いた場合，最小化問題 (3) は次の最小化問題と等価となる．

$$\min_{\mathbf{w}_m, \mathbf{u}_m \in \mathbb{R}^p, m=1, \dots, T} \left\{ \sum_{m=1}^T \frac{1}{2n_m} \|\mathbf{y}_m - \mathbf{X}_m \mathbf{w}_m\|_2^2 + \lambda_1 \sum_{m=1}^T h_{\lambda_3}(\mathbf{w}_m - \mathbf{u}_m) + \lambda_2 \sum_{(m,l) \in \mathcal{E}} r_{m,l} \|\mathbf{u}_m - \mathbf{u}_l\|_2 \right\}. \quad (4)$$

ここで， $h_{\lambda_3}(\cdot)$ は

$$h_{\lambda_3}(\mathbf{a}) = \begin{cases} \frac{1}{2} \|\mathbf{a}\|_2^2 & \|\mathbf{a}\|_2 \leq \lambda_3, \\ \lambda_3 \|\mathbf{a}\|_2 - \frac{\lambda_3^2}{2} & \|\mathbf{a}\|_2 > \lambda_3, \end{cases} \quad (5)$$

で定義される多変量フーバー損失関数 (Hampel et al., 1986) である．すなわち，提案手法は回帰係数に対してクラスタリングを実行する第二項目および第三項目を頑健化した手法として考えることができる．

最小化問題に含まれるパラメータの推定は，最小化問題 (3) に対する交互方向乗数法により実行される．数値実験の結果については当日紹介する．

参考文献

- Caruana, R. (1997). Multitask learning. *Machine Learning*, **28**, 41–75.
- Gong, P., Ye, J., and Zhang, C. (2012). Robust multi-task feature learning. *KDD '12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 895–903.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons.
- Hocking, T. D., Joulin, A., Bach, F., and Vert, J. P. (2011). Clusterpath an algorithm for clustering using convex fusion penalties. In *Proceedings of the 28th International Conference on Machine Learning*, 745–752.
- Okazaki, A. and Kawano, S. (2023). Multi-task learning regression via convex clustering. arXiv:2304.13342.
- Xue, Y., Liao, X., Carin, L., and Krishnapuram, B. (2007). Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, **8**(2), 35–63.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **68**(1), 49–67.

An Efficient Procedure for Computing Bayesian Network Structure Learning

Hongming Huang¹ and Joe Suzuki¹

¹Osaka University, Graduate School of Engineering Science, Department of Systems Innovation

1 Introduction

In recent years, there has been active progress in the Bayesian Network Structure Learning (BNSL). Finding the optimal Bayesian network structure is an NP-hard problem[2], so the proposed techniques are only applicable to small networks.

Tomi Silander and Petri Myllymäki proposed a straightforward score-based method for finding the optimal Bayesian network[1]. This state-of-the-art method continuously compares the scores of various structures to select the most favorable outcome. However, as shown on the left side of Figure 1, it needs to calculate and record both the local scores Q and the optimal local solutions R , which requires two passes over all possible local structures. Then, based on these information, a calculation for the globally optimal structure is performed, making it impossible to eliminate redundant data during the computation process.

In this paper, we introduce an efficient method for layered computation to reduce the space complexity. Differing from existing methods, it only requires a single pass through the possible local structures as shown on the right side of Figure 1 and has the capability to eliminate redundant information of the completed levels during the computational process. For example, when computing the topmost red layer, it only requires Q and R from the immediately lower yellow layer, and the information from the green layer and any layers below can be completely discarded as shown in Figure 2. This method retains only the crucial data necessary for calculations at higher structural levels, eliminating the need to hold onto redundant data.

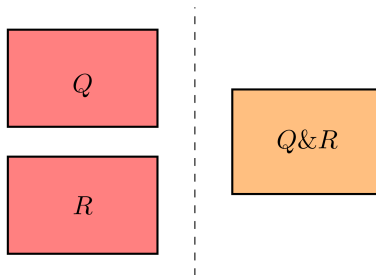


Figure 1: The left side of the dividing line represents the existing method where Q and R need to traverse possible local structures separately. The right side of the dividing line represents the proposed method where Q and R can be updated together, requiring only a single traversal.

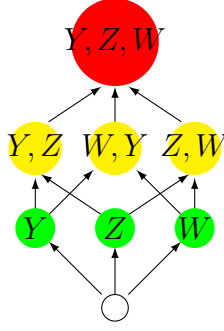


Figure 2: An example for proposed method. The red nodes represent the cases to be computed, the yellow ones indicate the data information required to compute the yellow nodes, and the green ones denote the nodes and data information that can be deleted.

2 Background

2.1 Marginal Likelihood

In the realm of statistical analysis, the precision of parameter estimation is paramount. Traditional methods often fall short when parameters are unknown, leading to uncertainty in the results. To overcome this, we leverage Bayesian inference, specifically the marginal likelihood $Q(X)$, which considers parameter uncertainty and integrates prior knowledge effectively.

We consider a Bayesian statistical model characterized by a set of realized values $\mathcal{X}$ for a random variable X , and a parameter set $\Theta \subseteq \mathbb{R}^d$. Given a sample size n and i.i.d. observations $x_1, \dots, x_n \in \mathcal{X}$, the model is described by a family of probability distributions $\{p(\cdot|\theta)\}_{\theta \in \Theta}$ and a prior distribution φ over Θ . The marginal likelihood is defined as:

$$Q(X) := \int_{\Theta} \varphi(\theta) \prod_{i=1}^n p(x_i|\theta) d\theta,$$

which synthesizes the likelihood across all parameter values, weighted by the prior.

2.2 Bayesian Network Structure Learning

Suppose that given n tuples of examples

$$X^{(1)} = x_{i,1}, \quad X^{(2)} = x_{i,2}, \quad \dots, \quad X^{(p)} = x_{i,p},$$

where $i = 1, 2, \dots, n$, with respect to p variables $X^{(1)}, X^{(2)}, \dots, X^{(p)}$.

A fundamental step in structure learning is finding the optimal parent sets for each node within the network. The joint probability of the variables $X^{(1)}, \dots, X^{(p)}$ in a Bayesian network is factorized according to the network structure as follows:

$$P(X^{(1)}, \dots, X^{(p)}) = P(X_1)P(X_2|\pi_2) \cdots P(X_i|\pi_i) \cdots P(X_p|\pi_p) = \prod_{i=1}^p P(X_i|\pi_i),$$

where X_i denotes the vertex at i -th position in the structure and π_i denotes the parent set of the vertex X_i .

For a Directed Acyclic Graph, a key fact is that due to the absence of cyclic structures, there must exist a universally bottom-most node. This node is defined as the network's sink node.

Assuming S is a non-empty set of nodes, it contains a bottom-most node X . The optimal parent set for X lies within $S \setminus \{X\}$, and identifying a new sink node and its optimal parent set continues recursively within this subset until completion. Set $V = \{X^{(1)}, \dots, X^{(p)}\}$, with the marginal likelihood $Q(X)$, the purpose is to find a BN structure with the maximum posterior probability:

$$R(X^{(1)}, \dots, X^{(p)}) = \max_{\pi_i} \prod_{i=1}^p \frac{Q(X_i, \pi_i)}{Q(\pi_i)}, \quad \pi_i \subseteq V \setminus \{X_p, \dots, X_i\}.$$

3 Existing method

For the state-of-the-art method proposed by Silander and Myllymäki, they firstly compute $Q(S)$ for all subsets $S \subseteq \{X^{(1)}, \dots, X^{(p)}\}$. Then, for each subset S , they will compute the optimal parent set for each node in it:

$$\frac{Q(X, \pi(X, S))}{Q(\pi(X, S))} = \max_{T \subseteq S} \frac{Q(X, T)}{Q(T)} = \max \left\{ \frac{Q(X, S)}{Q(S)}, \max_{Y \in S} \frac{Q(X, \pi(X, S \setminus Y))}{Q(\pi(X, S \setminus Y))} \right\},$$

where $\pi(X, S)$ is the parent set of $X \notin S$ in $S \subseteq \{1, \dots, p\}$. Then, they constructed an ordered graph to store the parent sets $\{\pi(X, S)\}_S$ for each $X \in \{1, \dots, p\}$.

In the second stage, they combine the results obtained from above and the process for finding the sink node:

$$R(S) = \begin{cases} \max_{X \in S} R(S \setminus X) \frac{Q(X, \pi(X, S \setminus X))}{Q(\pi(X, S \setminus X))}, & |S| \geq 1 \\ 1, & |S| = 0 \end{cases},$$

which recursively obtain the optimal parent set and the global optimal structure. To this end, they constructed an additional ordered graph to store and recursively obtain the $\{R(S)\}$ values.

4 Proposed Method

The proposed procedure seeks the same score more efficiently, constructing an ordered graph just once. We claim our method can dramatically save the memory as shown in Table 1.

For every subset S where $|S| \leq k$, we assume that $Q(S)$, $R(S)$, and $\pi(X, S \setminus X)$ for each X in S have already been computed, the process is then repeated for each subset S where $|S| = k + 1$. The following is the description of the algorithm steps:

1. We firstly compute $Q(S)$ for the calculation in Step 2.
2. For each $X \in S$, we define $\pi(X, S \setminus X)$ by $S \setminus X$ and $\pi(X, S \setminus \{X, Y\})$, $Y \in S \setminus X$, when

$$\frac{Q(S)}{Q(S \setminus X)} \quad \text{and} \quad \frac{Q(X, \pi(X, S \setminus \{X, Y\}))}{Q(\pi(X, S \setminus \{X, Y\}))}$$

are maximized, respectively. In this step, we will update the optimal parent set for each node in S . Then we store the information and value for Step 3 and upper level iterations.

3. In Step 3, we will find the $X \in S$ that maximizes $R(S \setminus X) \frac{Q(X, \pi(X, S \setminus X))}{Q(\pi(X, S \setminus X))}$ to obtain $R(S)$. Through this way, the optimal global structure is always obtained for each subset S .

	Silander-Myllymaki	Proposed method
Computation	$O(p^2 2^p)$	$O(p^2 2^p)$
memory	$O(p 2^p)$	$O(2^p / \sqrt{p})$

Table 1: Comparison of computation and memory requirements

5 Summary

We plan to conduct experiments in the future to verify that the spatial complexity of the proposed method is lower than that of the compared methods in practical applications. Furthermore, although we believe that the time complexities of both methods are similar, it has been suggested that the Branch & Bound technique was employed to reduce computational time in the compared method[3]. Therefore, our future work will also include the use of optimization techniques to decrease the time complexity of our proposed method.

References

- [1] Silander, T., & Myllymaki, P. (2012). A simple approach for finding the globally optimal Bayesian network structure. arXiv preprint arXiv:1206.6875.
- [2] Chickering, M., Heckerman, D., & Meek, C. (2004). Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5, 1287-1330.
- [3] Suzuki, J., & Kawahara, J. (2017). Branch and Bound for Regular Bayesian Network Structure Learning. In *UAI*.
- [4] Kitson, N. K., Constantinou, A. C., Guo, Z., Liu, Y., & Chobtham, K. (2023). A survey of Bayesian Network structure learning. *Artificial Intelligence Review*, 1-94.
- [5] Rantanen, K., Hyttinen, A., & Järvisalo, M. (2017). Learning chordal Markov networks via branch and bound. *Advances in neural information processing systems*, 30.
- [6] Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1), 3741-3782.
- [7] Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

An Information Criterion improving WBIC and SBIC

LIU LIRUI and Joe Suzuki

Mathematical Science Department, Graduate School of Engineering Sciences, Osaka University

We introduce a novel Information Criterion (IC) aimed at enhancing the functionality of WBIC and sBIC, designed to be effective without the regularity constraints and yet remain stable. ICs are essential in model selection, balancing model fit and complexity. Many existing criteria, like the singular Bayesian information criterion (sBIC), presuppose regularity in the true distribution relative to the statistical model. sBIC, despite its theoretical advantages in capturing large sample behavior (Drton et al., 2017), faces limitations in broader applications due to its dependency on maximum likelihood estimates and regularity assumptions (Suzuki, 2023). This led to the development of criteria such as the Watanabe Bayesian information criterion (WBIC) (Watanabe, 2013), which functions in non-regular scenarios but struggles with large sample sizes and the redundant estimation of known learning coefficients. Addressing these gaps, our proposed Information Criterion enhances WBIC and sBIC's utility, offering a flexible and robust approach to model selection without the regularity constraints.

We denote by $\hat{\theta}_n$ the maximum likelihood estimate, i.e., the θ that maximizes the likelihood $\prod_{i=1}^n p(x_i | \theta)$ obtained from $x_1, \dots, x_n \in \mathcal{X}$, and assume that the learning coefficient λ has been obtained somehow. The Singular Bayesian Information Criterion (Drton et al., 2017) is defined by

$$SBIC := \sum_{i=1}^n -\log p(x_i | \hat{\theta}_n) + \lambda \log n.$$

Assume that p, q are regular. When Θ is compact, the estimator $\hat{\theta}_n$ converges in probability to θ_* as $n \rightarrow \infty$ (Suzuki, 2023). Note that the performance of sBIC is guaranteed only in the regular case, and sBIC cannot be applied to general cases.

We introduce a parameter $\beta > 0$, called the inverse temperature, and assume that it decreases with n as $\beta = \beta_0 / \log n$ for some constant $\beta_0 > 0$. Given $x_1, \dots, x_n \in \mathcal{X}$, we define the posterior mean of a function $f : \Theta \rightarrow \mathbb{R}$ with respect to β by

$$\mathcal{E}_\beta[f(\theta)] := \frac{\int_{\Theta} f(\theta) \prod_{i=1}^n p(x_i | \theta)^\beta \varphi(\theta) d\theta}{\int_{\Theta} \prod_{i=1}^n p(x_i | \theta')^\beta \varphi(\theta') d\theta'} \quad (1)$$

The Watanabe Bayesian Information Criterion (Watanabe, 2013) is defined as

$$WBIC_n := \mathcal{E}_\beta \left[\sum_{i=1}^n -\log p(x_i | \theta) \right].$$

From (Watanabe, 2013), WBIC behaves as

$$WBIC_n = \sum_{i=1}^n -\log p(x_i | \theta_*) + \frac{\lambda \log n}{\beta_0} + U_n \sqrt{\frac{\lambda \log n}{2\beta_0}} + O_p(1), \quad (2)$$

where $U_1, U_2, \dots$ are random variables that converge in law to a normal distribution with mean zero.

We define the predictive distribution and the empirical loss (Watanabe, 2010) given $x_1, \dots, x_n \in \mathcal{X}$ as follows:

$$r(x | x_1, \dots, x_n) = \mathcal{E}_1[p(x | \theta)], \quad x \in \mathcal{X},$$

and

$$T_n := \frac{1}{n} \sum_{i=1}^n \{-\log r(x_i | x_1, \dots, x_n)\},$$

respectively. We propose a novel information criterion given by

$$\text{newIC} := nT_n + \lambda \log n. \quad (3)$$

In comparison with WBIC and sBIC, we argue the merits of newIC. First, unlike sBIC that relies on maximum likelihood estimates, newIC employs empirical loss. Assuming $\beta_0 = 1$, equation (2) becomes

$$WBIC_n = \sum_{i=1}^n -\log p(x_i | \theta_*) + \lambda \log n + O_p(\sqrt{\log n}),$$

where, in (2), the third term divided by $\sqrt{\log n}$ is $O_P(1)$. As per Suzuki (2023), T_n can be expressed as

$$T_n = \frac{1}{n} \sum_{i=1}^n \{-\log p(x_i | \theta_*)\} + \frac{1}{n} \left(\lambda - \frac{1}{2} \mathcal{E} \left[\sqrt{t} \xi_n(u) | x_1, \dots, x_n \right] \right) - \frac{1}{2} \mathbb{E}_X[\mathcal{V}(X)] + o_P\left(\frac{1}{n}\right),$$

where $\mathcal{E}[\cdot]$ and $\mathcal{V}[\cdot]$ are the posterior expectation and variance in equation (1) with $\beta = 1$ respectively. Noting that λ is constant and that $n\mathbb{E}_X[\mathcal{V}(X)]$ and $\mathcal{E}[\sqrt{t}\xi_n(u) | x_1, \dots, x_n]$ are $O_p(1)$ due to convergence to their expected values by the law of large numbers, and substituting them into the definition (3) of newIC, we obtain

$$\text{newIC} = \sum_{i=1}^n \{-\log p(x_i | \theta_*)\} + \lambda \log n + O_P(1). \quad (4)$$

If λ is not known, it can be estimated by $\hat{\lambda}$, obtained as

$$\hat{\lambda} = \frac{WBIC_1 - WBIC_2}{1/\beta_1 - 1/\beta_2} = \lambda + O_P(1/\sqrt{\log n}), \quad (5)$$

where this formula is derived from equation (2). Substituting λ with $\hat{\lambda}$, newIC can be approximated as

$$\text{newIC} \approx \sum_{i=1}^n \{-\log p(x_i | \theta_*)\} + \hat{\lambda} \log n + O_p(\sqrt{\log n}). \quad (6)$$

Comparing equations (2), (4), and (6), we provide the following statements:

- (I) If λ is not known, newIC and WBIC perform almost identically as $n \rightarrow \infty$.
- (II) If λ is known, newIC outperforms WBIC.

We applied newIC and WBIC to reduced-rank regression and Gaussian mixture models. Specifically, in former experiments using $\hat{\lambda}$ for newIC calculations, WBIC and newIC showed similar performance, with newIC consistently larger by a constant. For Gaussian mixture models, employing λ for newIC. Note that newIC is independent of β , it does not change with different β_0 and remained accurate, while WBIC was incorrect for $\beta_0 = 10$.

Our paper introduces the innovative newIC for model selection. Distinguishing itself from sBIC, newIC operates effectively beyond the confines of regularity, making it applicable in a wider range of scenarios. When contrasted with WBIC, newIC demonstrates enhanced performance, especially in cases where the learning coefficient is known. This advancement solidifies newIC's position as a versatile and reliable tool for model selection in diverse statistical contexts.

References

- [1] Suzuki, J. (2023). WAIC and WBIC with R Stan: 100 Exercises for Building Logic. Springer Nature.
- [2] Drton, M., Plummer, M. (2017). A Bayesian information criterion for singular models. Journal of the Royal Statistical Society Series B: Statistical Methodology, 79(2), 323-380.
- [3] Watanabe, S. (2013). A widely applicable Bayesian information criterion. The Journal of Machine Learning Research, 14(1), 867-897.
- [4] Watanabe, S. (2009). Algebraic geometry and statistical learning theory (Vol. 25). Cambridge university press.
- [5] Watanabe, S., Oppen, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of machine learning research, 11(12).
- [6] Aoyagi, M., Watanabe, S. (2005). Stochastic complexities of reduced rank regression in Bayesian estimation. Neural Networks, 18(7), 924-933.

データ量子統計の提案

放送大学大学院 池上 孝則

1. 最尤法の信頼性

外れ値や異常値を多く含むデータから統計量を安定的に算出するには、その方法論を適格に選択する必要がある。二次元データにおいて回帰式を決定する場合を例に挙げると、残差に着目した演算である最小二乗法や最小絶対値法は外れ値等の影響を受けやすく処理に細心の配慮が必要であるが、母数の尤もらしさを示す尤度に着目した処理である最尤法は、求める情報に対する直接的な処理であって残差に着目した演算ではないため、不確かさを内包するデータ群の処理においても有効に機能する可能性がある。

しかし最尤法には、こうした期待感がある反面、数理の常識に照らして腑に落ちない点も少なくない。その一例として、データが $N(\mu, \sigma^2)$ に従うと仮定し、その確率密度関数 $p(x|\theta)$ の逆関数を尤度と解釈した場合を考える。ここに、データ群 $(x_1, x_2, \dots, x_n)$ における x_i の尤度 $L(\theta; x_i)$ の同時確率によって μ を推定する尤度関数 L を定義した場合、簡単のために係数を無視すると、

$$\begin{aligned} L(\theta : x_1, x_2, \dots, x_n) &= \prod_{i=1}^n \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} = \exp\left\{-\frac{n\mu^2 - 2(x_1 + x_2 + \dots + x_n)\mu + (x_1^2 + x_2^2 + \dots + x_n^2)}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{n(x_1^2 + x_2^2 + \dots + x_n^2) - (x_1 + x_2 + \dots + x_n)^2}{2n\sigma^2}\right\} \exp\left\{-\frac{(\mu - \frac{x_1 + x_2 + \dots + x_n}{n})^2}{\frac{2\sigma^2}{n}}\right\} \end{aligned} \quad (1)$$

となる。式(1)は、その第1項は定数であることから、 $(x_1 + x_2 + \dots + x_n)/n$ を中心とした分散が σ^2/n の正規分布であることを表している。すなわち、当該尤度関数の最尤推定値は標本平均に一致するが、その結果は平均値の妥当性を再確認するという消極的な情報を提供しているに過ぎない。また、尤度関数の分散が σ^2/n であるということはデータが増す毎に分散が減少することを意味しているが、その一方でデータには普遍的に σ^2 程度のばらつきがあるとする統計量が付随しており、数理的に整合性がとれない。

2. 重ね合わせによる尤度関数の検討

最尤法におけるこうした疑義の原因が積統合によって構築する尤度関数にあると推定し、ここでは和結合、すなわち尤度を表す関数の重ね合わせによる情報統合の可能性を検討する。ここに、母集団の分布が $N(0, \sigma^2)$ と仮定し、完全な重ね合わせを行うという意味で正規分布関数 $f(x)$ 同士の畳み込み積分を行なうと、畳み込み関数 $h(x)$ は

$$\begin{aligned} h(x) &= \int_{t=-\infty}^{\infty} f(t)f(x-t)dt = \int_{t=-\infty}^{\infty} \exp\left(-\frac{t^2}{2\sigma^2}\right) \exp\left(-\frac{(x-t)^2}{2\sigma^2}\right)dt = \int_{t=-\infty}^{\infty} \exp\left(-\frac{2t^2 - 2xt + x^2}{2\sigma^2}\right)dt \\ &= \exp\left(-\frac{x^2}{4\sigma^2}\right) \int_{t=-\infty}^{\infty} \exp\left\{-\frac{(t - \frac{x}{2})^2}{\sigma^2}\right\}dt = \sqrt{\pi}\sigma \exp\left(-\frac{x^2}{4\sigma^2}\right) \end{aligned} \quad (2)$$

となる。

式(2)は $h(x)$ が分散 $2\sigma^2$ の正規分布であることを示しているが、それと同時に同時に $h(x)$ を二乗すると母集団の確率密度関数 $f(x)$ に回帰するということをも意味している。この事実は、量子論において「波動関数の絶対値の二乗は電子の存在確率を表す」とする「ボルンの確率解釈」に通じるものがある。さらに、尤度を虚数と仮定すれば尤度関数という虚の世界から確率密度関数という実の世界に回帰することになり、量子論の解釈に重ねることができる。こうした連想が以下に述べるデータ量子統計の骨格を作っている。

3. データ量子統計の世界観

データ量子統計とは、前述の正規分布同士の畳み込み積分の結果と量子論との相似性を踏まえ、尤度の処理を目的として構築した方法論である。ここに、量子論は「エネルギーには最小単位が存在し、その整数倍で分配される」とする革命的なアイデアに基づいて構成されているが、「データ量子」は「データは本

来的にその信頼性に関して上限を有する情報」と捉えて導入した量であり、「個々のデータが有する不確かさの程度はそれらが属するデータ群の分散 σ^2 程度」と解釈する。

また、前述の畳み込み積分の結果を踏まえ、データ量子を表す関数としてデータ群を近似する正規分布関数の逆関数を対応させる。この際、尤度には規格化条件が課されないことから、正規分布関数における規格化の為の定数は省略する。更に、量子論における「波動関数の確率解釈」との整合性を図るという観点に加え、確率と尤度を明確に区別するために、尤度は純虚数として位置付ける。

すなわち、 $N(\mu, \sigma^2)$ を仮定したデータ群 $(x_1, x_2, \dots, x_n)$ において、データ x_j の尤度を表すデータ量子を φ_j とすると、データ量子関数 φ を

$$\varphi = \sum_{j=1}^n \varphi_j = i \sum_{j=1}^n \exp\left\{-\frac{(x_j - \mu)^2}{2\sigma^2}\right\} \quad (3)$$

と定義する。ちなみに、データ量子関数は最尤法における尤度関数に相当する関数である。また、データ量子関数の絶対値の二乗で構成される関数 Ψ 、すなわち

$$\Psi = |\varphi|^2 \quad (4)$$

を定義し、それを「データ密度関数」と呼ぶこととする。データ密度関数 Ψ は母世界における尤度の状態を表すデータ量子関数を実世界におけるデータの存在確率を表す関数に変換したものであり、離散的なデータを微積分が可能な連続関数で処理できることから、記述統計や検定の分野での活用が期待される。

4. データ量子統計に基づく統計量

データ量子統計における代表値は最尤法やベイズ推定と同様にデータ量子関数が最大となる最尤推定値を代表値とするが、分散の定義は統計学とは異なる。すなわち、データ量子関数 φ の最大値が $\varphi_{\max}$ であるとき、 φ と $y = \varphi_{\max} \cdot \exp(-\frac{1}{4})$ （データ密度関数の場合 Ψ と $y = \Psi_{\max} \cdot \exp(-\frac{1}{2})$ ）との交点を σ_L 、 σ_H とすると、

$$\sigma^2 = \left\{ \frac{\sigma_L + \sigma_H}{2} \right\}^2 \quad (5)$$

で定義し、混乱を避けるために「データ量子分散」と呼ぶことにする。なお、最尤推定値とデータ量子分散は収束演算によって同時に決定され、共に外れ値や異常値に対して頑健な統計量である。

相関係数は、まず二次元データにおける x 方向、 y 方向の最尤推定値およびデータ量子分散 (μ_x, σ_x^2) 、 (μ_y, σ_y^2) を求め、次に共分散を求めるデータ量子関数 φ 、すなわち

$$\varphi = i \sum_{j=1}^n \exp\left\{-\frac{(x_j - \mu_x)(y_j - \mu_y)}{2\sigma^2}\right\} \quad (6)$$

に基づいて「データ量子共分散」 σ_{xy} を算出し、相関係数 $r = \sigma_{xy} / (\sigma_x \sigma_y)$ を求める。

回帰式は、 $y = a + bx$ である場合には回帰係数は $b = \sigma_{xy} / \sigma_x^2$ 、 $a = \mu_y - b\mu_x$ で決定する。一般的な回帰式の場合は、まずデータ群と回帰式の評価するデータ量子関数 φ を設定する。すなわち、点 $p(x_i, y_i)$ から回帰式 $f(x)$ に下した x 軸方向・ y 軸方向の残差を δ_{ix} 、 δ_{iy} とすると、両者に比例関係がない場合は

$$\varphi = \sum_{i=1}^n \left\{ \exp\left(-\frac{\delta_{ix}^2}{2\sigma_x^2}\right) + \exp\left(-\frac{\delta_{iy}^2}{2\sigma_y^2}\right) \right\} = \sum_{i=1}^n \left[\exp\left\{-\frac{\{x_i - f^{-1}(y_i)\}^2}{2\sigma_x^2}\right\} + \exp\left\{-\frac{\{y_i - f(x_i)\}^2}{2\sigma_y^2}\right\} \right] \quad (7)$$

でデータ量子関数 φ を定義し、 φ が最大になるように回帰係数を決定する。

5. まとめ

データ量子統計は、尤度は確率とは異質の情報との基本認識に基づき、量子論を手掛かりとして確立した方法論である。処理対象とするデータに統計モデルを仮定する必要がなく、不確かさを含むデータであっても画一的な手続きによって実数で信頼性の高い結果を安定的に導くことができる等の特徴を有するが、論理性に関する検証はまだ不十分な段階であり、またデータ処理の大半は収束演算を多用する数値演算であることから、先ずは方法論の実効性を重視する理工系の研究者での活用をお勧めしたい。

Semiparametric density estimation with Bregman divergence

千葉大・融合理工学府 松野 大輔

千葉大・理学研究院 内藤 貫太

はじめに：密度推定におけるセミパラメトリックな推定量を考える．提案する推定量は，パラメトリック密度推定量とノンパラメトリックに構成した調整項との積として定義される．調整項は，関数間の距離尺度 Bregman ダイバージェンスを局所化した基準の最小化を通して導出される．このようにして得た推定量の漸近的性質と実装について報告した．

Bregman ダイバージェンス：非負値関数 f, g 間における Bregman ダイバージェンスは，狭義凸関数 $U: [0, \infty) \rightarrow \mathbb{R}$ ，その凸共役 $U^*(s) := \sup_{z \in \mathbb{R}} \{zs - U(z)\}$ ， $u := U'$ を用いて

$$\begin{aligned} D_U(f, g) &:= \int_{\mathbb{R}^d} [U(f(t)) - U(g(t)) - u(g(t))\{f(t) - g(t)\}] dt \\ &= \int_{\mathbb{R}^d} [U^*(u(g(t))) - U^*(u(f(t))) - f(t)\{u(g(t)) - u(f(t))\}] dt \end{aligned}$$

と定義される．例えば， $U(x)$ を $x^2/2$ や $x \log x - x$ とすれば， $D_U(f, g)$ は f, g 間の 2 乗距離や Kullback-Leibler (KL) ダイバージェンスとなる．また， U を β -function

$$U_\beta(x) := \frac{x^{\beta+1}}{\beta(\beta+1)} - \frac{x}{\beta}, \quad \beta \in (0, 1] \quad (1)$$

とすると，パラメータが $\beta = 1$ と $\beta \searrow 0$ となるとき， $D_{U_\beta}(f, g)$ はそれぞれ 2 乗距離と KL ダイバージェンスに一致する．この Bregman ダイバージェンスを，密度の推定点 $\mathbf{x} \in \mathbb{R}^d$ で

$$D_U(f, g | \mathbf{x}) := \int_{\mathbb{R}^d} K_H(\mathbf{t} - \mathbf{x}) [U^*(u(g(\mathbf{t}))) - U^*(u(f(\mathbf{t}))) - f(\mathbf{t})\{u(g(\mathbf{t})) - u(f(\mathbf{t}))\}] dt$$

のように局所化する．ここで， $K_H(\mathbf{z}) = |H|^{-1/2} K(H^{-1/2}\mathbf{z})$ ， $K: \mathbb{R}^d \rightarrow [0, \infty)$ は原点对称な核関数，バンド幅行列 H は $d \times d$ 正定値対称行列である．

設定： n 個の $\mathbb{R}^d$ 値観測値 $\mathbf{X}_1, \dots, \mathbf{X}_n$ が独立に密度 $f(\mathbf{x})$ に従うとする．この密度 $f(\mathbf{x})$ を観測値に基づいて推定する．提案する密度推定量は次の 3 段階で構成される：

〈Step 1〉パラメトリックモデル $g(\mathbf{x}, \boldsymbol{\theta})$ とパラメータの推定量 $\hat{\boldsymbol{\theta}}$ を与える．ただし， $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ ， $p \geq 1$ とする．

〈Step 2〉調整項 $\hat{\xi}(\mathbf{x}) := \arg \min_{\xi} L_n(\mathbf{x}, \xi, \hat{\boldsymbol{\theta}})$ を導入する．ここで， $L_n(\mathbf{x}, \xi, \hat{\boldsymbol{\theta}})$ は， $D_U(f, g(\cdot, \hat{\boldsymbol{\theta}})\xi | \mathbf{x})$ の推定量の ξ に依存する項

$$L_n(\mathbf{x}, \xi, \hat{\boldsymbol{\theta}}) := \int_{\mathbb{R}^d} K_H(\mathbf{t} - \mathbf{x}) U^*(u(g(\mathbf{t}, \hat{\boldsymbol{\theta}})\xi)) dt - \frac{1}{n} \sum_{i=1}^n K_H(\mathbf{X}_i - \mathbf{x}) u(g(\mathbf{X}_i, \hat{\boldsymbol{\theta}})\xi)$$

である．

〈Step 3〉セミパラメトリック推定量 $\hat{f}(\mathbf{x}) := g(\mathbf{x}, \hat{\boldsymbol{\theta}}) \cdot \hat{\xi}(\mathbf{x})$ を得る．

漸近的結果： $g(\mathbf{x}, \boldsymbol{\theta})$ による f の最良近似を与えるパラメータを $\boldsymbol{\theta}_0$ とする. $n \rightarrow \infty$ において $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = O_P(1/\sqrt{n})$ とし, $g_0(\mathbf{x}) := g(\mathbf{x}, \boldsymbol{\theta}_0)$ と略記する. $\hat{\boldsymbol{\theta}}$ 以外に, 推定対象の $f(\mathbf{x})$ と, 提案推定量 $\hat{f}(\mathbf{x})$ を定める g, K_H, U に対しても, 収束や滑らかさ, 可積分性についての仮定を置いた下で, $\hat{f}(\mathbf{x})$ は次のような漸近正規性を持つ.

$$\sqrt{n|H|^{1/2}} \left\{ \hat{f}(\mathbf{x}) - f(\mathbf{x}) - \frac{\mu_2}{2g_0(\mathbf{x})^2} \text{tr}[HB(\mathbf{x})] \right\} \xrightarrow{d} N\left(0, \|K\|_2^2 f(\mathbf{x})\right), \text{ as } n \rightarrow \infty, H \rightarrow O.$$

ここで

$$B(\mathbf{x}) := g_0(\mathbf{x}) \mathcal{H}_{g_0 \cdot f}(\mathbf{x}) - f(\mathbf{x}) \mathcal{H}_{g_0^2}(\mathbf{x}) + 2 \cdot \frac{u''(f(\mathbf{x}))}{u'(f(\mathbf{x}))} f(\mathbf{x}) g_0(\mathbf{x})^2 \cdot \frac{\partial}{\partial \mathbf{x}} \left\{ \frac{f(\mathbf{x})}{g_0(\mathbf{x})} \right\} \cdot \frac{\partial}{\partial \mathbf{x}^\top} g_0(\mathbf{x})$$

であり, $\mathcal{H}_h$ は関数 h の Hesse 行列とする.

実装：提案推定量 $\hat{f}(\mathbf{x})$ の実装にあたって, g を正規密度, $\hat{\boldsymbol{\theta}}$ を平均と共分散行列の最尤推定量, K をガウシアンとする. U を (1) の U_β として導かれた密度推定量

$$\hat{f}_\beta(\mathbf{x}) := g(\mathbf{x}, \hat{\boldsymbol{\theta}}) \cdot \frac{n^{-1} \sum_{i=1}^n K_H(\mathbf{X}_i - \mathbf{x}) g(\mathbf{X}_i, \hat{\boldsymbol{\theta}})^\beta}{\int_{\mathbb{R}^d} K_H(\mathbf{t} - \mathbf{x}) g(\mathbf{t}, \hat{\boldsymbol{\theta}})^{1+\beta} d\mathbf{t}} \quad (2)$$

を考える. この $\hat{f}_\beta(\mathbf{x})$ が有限個の正規密度の線型結合として表されることが示される.

シミュレーション：推定対象の密度は, 混合正規密度 (Wand and Jones (1993)) や歪正規密度 (Azzalini and Capitanio (2014)) とする. f がこのような混合正規密度, 歪正規密度の場合, $\hat{f}_\beta(\mathbf{x})$ の積分二乗誤差が陽に表現されることが示される. この事実に基づき, $\hat{f}_\beta(\mathbf{x})$ を, Hjort and Glad (1995) で提案された関連推定量

$$\hat{f}_{\text{HG}}(\mathbf{x}) := g(\mathbf{x}, \hat{\boldsymbol{\theta}}) \cdot \frac{1}{n} \sum_{i=1}^n \frac{K_H(\mathbf{X}_i - \mathbf{x})}{g(\mathbf{X}_i, \hat{\boldsymbol{\theta}})}$$

や, 従来のノンパラメトリックカーネル密度推定量

$$\hat{f}_{\text{KDE}}(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n K_H(\mathbf{x} - \mathbf{X}_i)$$

と MISE で比較するシミュレーションを行った. また, $\hat{f}_\beta(\mathbf{x})$ のバンド幅選択アルゴリズムについても提案した. このバンド幅選択アルゴリズムにより得られたバンド幅を用いた $\hat{f}_\beta$ の精度についても, シミュレーションにより検証した.

シンポジウムにおいては, (2) の $\hat{f}_\beta(\mathbf{x})$ の平均積分二乗誤差 (MISE) による推定精度を検証するシミュレーション結果について報告した.

参考文献

- Azzalini, A. and Capitanio, A. (2014). *The Skew-Normal and Related Families*, volume 3 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press, Cambridge.
- Hjort, N. L. and Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.*, **23**(3):882–904.
- Wand, M. P. and Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.*, **88**(422):520–528.

一般化 Kendall's tau 行列による主成分分析

千葉大・融合理工学府 蔣 思澄

千葉大・理学院 内藤 貫太

はじめに：ロバストな主成分分析の手法として、Kendall's tau 行列に基づく方法が知られている。この方法の理論的性質について調べた先行研究は数多く存在する。Kendall's tau 行列は、多変量 Spatial Sign によって構成されるが、近年この多変量 Spatial Sign の一般化の研究に進捗が見られた。本報告では、この多変量 Spatial Sign の一般化を反映した一般化 Kendall's tau 行列を新たに定義し、その性質について議論した。

設定： $\mathbf{X}, \mathbf{Y}, \mathbf{X}_1, \dots, \mathbf{X}_n$ は *i.i.d.* で期待値ベクトル $\boldsymbol{\mu} \in \mathbb{R}^d$ 、有限な正定値共分散行列 Σ を持つ分布 F に従うとする。 d 次元ベクトル $\mathbf{y}$ の Spatial Sign は

$$S(\mathbf{y}) = \begin{cases} \frac{\mathbf{y}}{\|\mathbf{y}\|} & , \mathbf{y} \neq \mathbf{0} \\ \mathbf{0} & , \mathbf{y} = \mathbf{0} \end{cases}$$

で定義される。ここで、 $\|\mathbf{y}\|^2 = \mathbf{y}^\top \mathbf{y}$ である。母集団での Kendall's tau 行列 $\mathbf{K} \in \mathbb{R}^{d \times d}$ は

$$\mathbf{K} = E \left[S(\mathbf{X} - \mathbf{Y}) S(\mathbf{X} - \mathbf{Y})^\top \right]$$

で定義される (Han and Liu (2018)). ξ を非負動径関数とし、これを重み関数とする。このとき、一般化 Kendall's tau 行列 $\mathbf{K}_\xi$ は

$$\mathbf{K}_\xi = E \left[\xi(\|\mathbf{X} - \mathbf{Y}\|)^2 (\mathbf{X} - \mathbf{Y})(\mathbf{X} - \mathbf{Y})^\top \right] = E[h(\mathbf{X}, \mathbf{Y})]$$

で定義される。ここで、 $h(\mathbf{X}, \mathbf{Y}) = \xi(\|\mathbf{X} - \mathbf{Y}\|)^2 (\mathbf{X} - \mathbf{Y})(\mathbf{X} - \mathbf{Y})^\top$ であり、 $\xi(r) = 1/r$ のとき、 $\mathbf{K}_\xi = \mathbf{K}$ である。 $\mathbf{K}_\xi$ のスペクトル分解を $\mathbf{K}_\xi = \Gamma \Lambda \Gamma^\top$ とする。ここで、

$$\Gamma = [\boldsymbol{\gamma}_1 \dots \boldsymbol{\gamma}_d], \Gamma \Gamma^\top = \Gamma^\top \Gamma = I_d, \Lambda = \text{diag}(\lambda_1, \dots, \lambda_d) = \begin{bmatrix} \lambda_1 & & O \\ & \ddots & \\ O & & \lambda_d \end{bmatrix}$$

であり、固有値については、単根性 $\lambda_1 > \lambda_2 > \dots > \lambda_d > 0$ を仮定する。 $\mathbf{K}_\xi$ の推定量 $\widehat{\mathbf{K}}_\xi$ は

$$\widehat{\mathbf{K}}_\xi = \frac{1}{n(n-1)} \sum_{i,j}^n \left[\xi(\|\mathbf{X}_i - \mathbf{X}_j\|)^2 (\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^\top \right]$$

で定義される。

問題：標本共分散行列 $\widehat{\Sigma}$ は外れ値の影響を受けやすいので、 $\widehat{\Sigma}$ による主成分分析はロバストではないことはよく知られている。一方、 F が楕円分布のとき、 Σ と $\mathbf{K}_\xi$ が同じ直交行列で対角化できることを示すことができる。このことから、 $\widehat{\mathbf{K}}_\xi$ の固有ベクトルが、 $\widehat{\Sigma}$ に基づく主成分分析に代わるロバストな主成分分析を提供することが期待される。本報告では Raymaekers and Rousseeuw (2019) で紹介された重み関数 ξ と $\mathbf{K}$ に対応する重み関数を含む 6 つの ξ の集合 Ξ に対して、 ξ の選択基準を構築し、どの ξ がどのような意味で良いのか調べた。特に、本報告では固有空間の類似度と $\mathbf{K}_\xi$ の固有ベクトルの影響関数に基づく ξ の選択基準を考えた。

固有空間の類似度に基づく選択基準： $d \times d$ 正定値行列 A と B に対して、 A と B のスペクトル分解を

$A = Y\Lambda_A Y^\top$ と $B = W\Lambda_B W^\top$ とし、直交行列を $Y = [\mathbf{y}_1 \cdots \mathbf{y}_d]$, $W = [\mathbf{w}_1 \cdots \mathbf{w}_d]$ とする. 部分行列を $Y_k = [\mathbf{y}_1 \cdots \mathbf{y}_k]$, $W_k = [\mathbf{w}_1 \cdots \mathbf{w}_k]$, $k = 1, \dots, d$ で定義する. k 次元までの A と B の固有空間の類似度を

$$S_k(A, B) = \frac{1}{k} \sum_{i=1}^k \sum_{j=1}^k (\mathbf{y}_i^\top \mathbf{w}_j)^2 = \frac{1}{k} \text{Tr}(Y_k Y_k^\top W_k W_k^\top)$$

で定義する (Marden (1999)). この定義より, $0 \leq S_k \leq 1$ である. または $S_k(A, B) = 1$ のとき, A と B が同じ固有空間を持つことがわかる. $\mathbf{K}_\xi$ と $\widehat{\mathbf{K}}_\xi$ の類似度について, $n \rightarrow \infty$ のとき,

$$S_k(\mathbf{K}_\xi, \widehat{\mathbf{K}}_\xi) = 1 - \frac{1}{nk} \left[\sum_{i=1}^k \sum_{j=k+1}^d \left(\frac{\gamma_j^\top W_1 \gamma_i}{\lambda_j - \lambda_i} \right)^2 \right] + o_P\left(\frac{1}{n}\right)$$

が得られる. ここで, $W_1 = (1/\sqrt{n}) \sum_{i=1}^n \{2E_Y[h(\mathbf{X}_i, \mathbf{Y})] - 2\mathbf{K}_\xi\}$ である. データに基づき ξ を選択する場合,

$$G_k(\mathbf{K}_\xi, \widehat{\mathbf{K}}_\xi) = \frac{1}{k} \sum_{i=1}^k \sum_{j=k+1}^d \left(\frac{\gamma_j^\top W_1 \gamma_i}{\lambda_j - \lambda_i} \right)^2$$

を考えることにする. 特に, $G_k(\mathbf{K}_\xi, \widehat{\mathbf{K}}_\xi)$ の変動を考慮し, k を固定したとき, 最適な重み $\widehat{\xi}_{opt}$ を

$$\widehat{\xi}_{opt} = \arg \min_{\xi \in \Xi} E[G_k(\mathbf{K}_\xi, \widehat{\mathbf{K}}_\xi)]$$

と定め, ξ を選択する.

影響関数に基づく選択基準: $\overline{\mathbf{X}}$ を $\mathbf{X}_1, \dots, \mathbf{X}_n$ から得られた標本平均となる. $M(\mathbf{y}) = (\mathbf{y} - \overline{\mathbf{X}})^\top \widehat{\Sigma}^{-1} (\mathbf{y} - \overline{\mathbf{X}})$ とする. このとき,

$$\mathbf{X}_{large} = \arg \max_{\mathbf{X} \in \{\mathbf{X}_1, \dots, \mathbf{X}_n\}} M(\mathbf{X})$$

と定める. このとき, ロバストネスの観点から, $\mathbf{K}_\xi$ の第一固有ベクトル γ_1 の影響関数 (Hampel et al. (2005)) である $IF_\xi(\mathbf{x}, \gamma_1, F)$ を $\mathbf{x} = \mathbf{X}_{large}$ で評価した $\|IF_\xi(\mathbf{X}_{large}, \gamma_1, F)\|$ を用いる. 特に,

$$\bar{\xi}_{opt} = \arg \min_{\xi \in \Xi} \|IF_\xi(\mathbf{X}_{large}, \gamma_1, F)\|$$

で, ξ を選択する.

適用例: ξ の選択に関連するシミュレーション実験の結果および実データへの適用結果については, シンポジウムにて報告した.

参考文献

- Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (2005). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley and Sons, Inc.
- Han, F. and Liu, H. (2018). Eca: High-dimensional elliptical component analysis in non-gaussian distributions. *Journal of the American Statistical Association*, **113**(521):252–268.
- Marden, J. I. (1999). Some robust estimates of principal components. *Statistics and Probability Letters*, **43**(4):349–359.
- Raymaekers, J. and Rousseeuw, P. (2019). A generalized spatial sign covariance matrix. *Journal of Multivariate Analysis*, **171**:94–111.

3-step 単調欠測データにおける T^2 型検定統計量を用いた 平均ベクトルの検定

東京理科大学大学院 理学研究科 森本 直紀
東京理科大学 理学部 八木 文香
東京理科大学 理学部 瀬尾 隆

1 はじめに

3-step 単調欠測データにおける平均ベクトルの検定問題 $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs. $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ について考えた。ただし、 $\boldsymbol{\mu}_0$ は既知ベクトルとした。平均ベクトルに対する T^2 型検定統計量 Q は

$$Q = (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)' \widehat{\text{Cov}}[\tilde{\boldsymbol{\mu}}]^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) = \sum_{i=1}^3 Q_i \quad (1)$$

と表され、 Q を構成する統計量 Q_1, Q_2, Q_3 の間に独立性を仮定せず相関を考慮することにより、 Q の正確な特性関数の漸近展開及び近似上側パーセント点を導出した。なお、 $\boldsymbol{\mu}_0 = \mathbf{0}$ としても一般性を失わない。(1)において、 $\hat{\boldsymbol{\mu}}$ は $\boldsymbol{\mu}$ の最尤推定量、 $\tilde{\boldsymbol{\mu}}$ は $\boldsymbol{\Sigma}$ を既知としたときの $\boldsymbol{\mu}$ の最尤推定量、 $\widehat{\text{Cov}}[\tilde{\boldsymbol{\mu}}]$ は $\text{Cov}[\tilde{\boldsymbol{\mu}}]$ の推定量である ($\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の最尤推定量については、Jinadasa, and Tracy (1992) を参照)。また、Fujikoshi (2000) に基づいて、 Q のカイ二乗近似を改良する変換統計量を提案する。単調欠測データにおける平均ベクトルの検定問題については、統計量の間の相関を考慮すると計算が複雑になるため、独立性を仮定したもとの研究が進められてきた。相関を考慮したときの検定統計量 Q の正確な帰無分布に関しては、Yagi, Seo, and Hanusz (2019) で 2-step の場合まで議論されており、本研究はその研究を 3-step の場合に拡張したもので、より正確な検定統計量の帰無分布を与えた。

3-step 単調欠測データを以下のように設定した。

$$\left\{ \begin{array}{ccc} \mathbf{x}'_{11} & \mathbf{x}'_{12} & \mathbf{x}'_{13} \\ \vdots & \vdots & \vdots \\ \mathbf{x}'_{N_1 1} & \mathbf{x}'_{N_1 2} & \mathbf{x}'_{N_1 3} \\ \mathbf{x}'_{N_1+1,1} & \mathbf{x}'_{N_1+1,2} & * \\ \vdots & \vdots & \vdots \\ \mathbf{x}'_{N_2 1} & \mathbf{x}'_{N_2 2} & * \\ \mathbf{x}'_{N_2+1,1} & * & * \\ \vdots & \vdots & \vdots \\ \mathbf{x}'_{N_3 1} & * & * \end{array} \right\} \begin{array}{l} \left. \begin{array}{c} \left. \begin{array}{c} \left. \begin{array}{c} \mathbf{x}'_{11} \\ \vdots \\ \mathbf{x}'_{N_1 1} \end{array} \right\} n_1 \\ \left. \begin{array}{c} \mathbf{x}'_{N_1+1,1} \\ \vdots \\ \mathbf{x}'_{N_2 1} \end{array} \right\} n_2 \\ \left. \begin{array}{c} \mathbf{x}'_{N_2+1,1} \\ \vdots \\ \mathbf{x}'_{N_3 1} \end{array} \right\} n_3 \end{array} \right\} p_1 \end{array} \right\} p_2 \end{array} \right\} p_3$$

$$\begin{array}{l} \mathbf{x}_{i1} \stackrel{\text{i.i.d.}}{\sim} N_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}), \quad i = 1, 2, \dots, N_3, \\ \mathbf{x}_{i2} \stackrel{\text{i.i.d.}}{\sim} N_{p_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}), \quad i = 1, 2, \dots, N_2, \\ \mathbf{x}_{i3} \stackrel{\text{i.i.d.}}{\sim} N_{p_3}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_{33}), \quad i = 1, 2, \dots, N_1, \\ \mathbf{x}_{iP_3} = (\mathbf{x}'_{i1}, \mathbf{x}'_{i2}, \mathbf{x}'_{i3})' \stackrel{\text{i.i.d.}}{\sim} N_{P_3}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, 2, \dots, N_1, \\ \mathbf{x}_{iP_2} = (\mathbf{x}'_{i1}, \mathbf{x}'_{i2})' \stackrel{\text{i.i.d.}}{\sim} N_{P_2}(\boldsymbol{\mu}_{12}, \boldsymbol{\Sigma}_2), \quad i = 1, 2, \dots, N_2, \\ P_j = \sum_{i=1}^j p_i, \quad N_j = \sum_{i=1}^j n_i, \quad j = 1, 2, 3, \end{array}$$

ただし、 p_i, n_i ($i = 1, 2, 3$) をそれぞれ単調欠測データの各 step の次元数及びデータ数とした。また、3-step 単調欠測データの平均ベクトルと分散共分散行列をそれぞれ

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_3 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{12} \\ \boldsymbol{\mu}_3 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} \\ \boldsymbol{\Sigma}'_{12} & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} \\ \boldsymbol{\Sigma}'_{13} & \boldsymbol{\Sigma}'_{23} & \boldsymbol{\Sigma}_{33} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_2 & \boldsymbol{\Sigma}_{12(3)} \\ \boldsymbol{\Sigma}'_{12(3)} & \boldsymbol{\Sigma}_{33} \end{pmatrix}$$

とした。

2 検定統計量について

統計量 Q_1, Q_2, Q_3 の間の相関を考慮すると Q の特性関数の漸近展開として

$$\phi(t) = (1 - 2it)^{-\frac{P_3}{2}} + \frac{1}{n_1} \sum_{j=0}^2 \beta_j (1 - 2it)^{-\frac{P_3}{2}-j} + O(n_1^{-\frac{3}{2}}) \quad (2)$$

が得られ、この結果を用いて以下の Q の近似上側 100 α % 点 $q_{AEC}(\alpha)$ と Q のカイ二乗近似を改良する変換統計量 Q_{YC}, Q_{ZC} を提案した.

$$q_{AEC}(\alpha) = \chi_{P_3}^2(\alpha) - \frac{2}{n_1 P_3} \left(\beta_0 - \frac{\beta_2}{P_3 + 2} \chi_{P_3}^2(\alpha) \right) \chi_{P_3}^2(\alpha), \quad (3)$$

$$Q_{YC} = (n_1 \gamma_1 + \gamma_2) \log \left(1 + \frac{1}{n_1 \gamma_1} Q \right), \quad (4)$$

$$Q_{ZC} = (n_1 \gamma_1 + \gamma_2) \left\{ 1 - \exp \left(-\frac{1}{n_1 \gamma_1} Q \right) \right\} \quad (5)$$

ここで、(2) ~ (5) の各式において

$$\begin{aligned} \beta_0 &= -\frac{1}{4} \left\{ p_3(4p_1 + 4p_2 + p_3 + 2) + \frac{p_2(4p_1 + p_2 - 2p_3 + 2)}{1 + r_2} + \frac{p_1(p_1 - 2p_2 - 2p_3 + 2)}{1 + r_2 + r_3} \right\}, \\ \beta_1 &= p_3(p_1 + p_2) + \frac{p_2(p_1 - p_3)}{1 + r_2} - \frac{p_1(p_2 + p_3)}{1 + r_2 + r_3}, \\ \beta_2 &= \frac{1}{4} \left\{ p_3(p_3 + 2) + \frac{p_2(p_2 + 2p_3 + 2)}{1 + r_2} + \frac{p_1(p_1 + 2p_2 + 2p_3 + 2)}{1 + r_2 + r_3} \right\}, \\ \gamma_1 &= \frac{P_3(P_3 + 2)}{4\beta_2}, \quad \gamma_2 = \frac{(P_3 + 2)\beta_0}{2\beta_2} = \frac{2\beta_0}{P_3} \gamma_1, \\ r_2 &= \frac{n_2}{n_1}, \quad r_3 = \frac{n_3}{n_1} \end{aligned}$$

であり、 $\chi_{P_3}^2(\alpha)$ は自由度 P_3 のカイ二乗分布の上側 100 α % 点である. (4) 及び (5) の変換統計量は、(2) で与えられる Q の正確な特性関数から得られるものであり、極限 ($n_1 \rightarrow \infty$) を考えたときに、 Q の分布関数は $O(n_1^{-1})$ となるのに対して、変換統計量の分布関数には $O(n_1^{-1})$ の項は現れず $O(n_1^{-2})$ となるため、 Q と比較して、より小さな n_1 でカイ二乗分布へ収束する.

最後に、いくつかのパラメータの下でモンテカルロ・シミュレーションを行い、検定統計量及び変換統計量のカイ二乗近似の精度などについて数値的評価を行った. また、数値例を用いることにより、本研究の結果を実データに適用できるかを確かめた.

参考文献

- [1] Fujikoshi, Y. (2000). Transformations with Improved Chi-Squared Approximations. *Journal of Multivariate Analysis*, **72**, 249-263.
- [2] Jinadasa, K. G., and Tracy, D.S. (1992). Maximum Likelihood Estimation for Multivariate Normal Distribution with Monotone Sample. *Communications in Statistics - Theory and Methods*, **21**, 41-50.
- [3] Yagi, A., Seo, T., and Hanusz, Z. (2019). Improved Simplified T^2 Test Statistics for a Mean Vector with Monotone Missing Data. *Communications in Statistics - Simulation and Computation*, **48**, 1606-1622.

Modal Interval Regression for Visualizing Data-frequent Areas

Sai YAO (Tohoku University)*1
Yuko ARAKI (Tohoku University)*2

January 26, 2024

1 Introduction

Modal interval [1, 2] is a statistic representing the spread of data, which can show the area where observations appear frequently and robust to distribution. Regression analysis is often used to analyze the continuous relationship between two random variables. A modal interval regression method for visualization of bivariate data is proposed in reference [2]. In this presentation, we improved the method in reference [2] by using kernel density estimation instead of bin segmentation. Numerical experiments show that, compare with the method in [2], proposed method improves the accuracy of the results.

2 Preliminaries

2.1 Modal Interval

For every $\alpha \in (0, 1)$, the $100\alpha\%$ modal interval [1, 2] is defined by

$$\text{MI}_\alpha := [m_\alpha^L, m_\alpha^R] := \underset{[x_L, x_R] \subseteq \text{cl}(x_{\inf}, x_{\sup})}{\text{argmin}} \quad x_R - x_L \quad \text{s.t.} \quad \int_{x_L}^{x_R} f_X(x) dx = \alpha.$$

2.2 Quantile Regression

Asymmetric absolute value function is defined by

$$\mathcal{J}_p(t) := \begin{cases} pt & \text{if } t \geq 0, \\ -(1-p)t & \text{if } t < 0, \end{cases}$$

with $p \in (0, 1)$. The quantile regression is formulated as the following optimization problem [3]

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \sum_{i=1}^N \mathcal{J}_p(y_i - r_{\boldsymbol{\theta}}(x_i)), \quad (1)$$

where (x_i, y_i) is observation, function $r_{\boldsymbol{\theta}}(x)$ is regression model, and vector $\boldsymbol{\theta}$ denotes adjustable parameters to be optimized.

2.3 Spline Smoothing

The spline function is a piecewise polynomial and possesses certain-times continuous differentiability [4]. Spline smoothing is to solve the following optimization problem

$$\underset{s \in \mathcal{S}_d^p(\sqcup_b)}{\text{minimize}} \quad \sum_{i=1}^N w_i \mathcal{L}(y_i - s(x_i)) + \lambda \int_I |s''(x)|^2 dx, \quad (2)$$

*1 e-mail: yaosai@dc.tohoku.ac.jp

*2 e-mail: yaraki@tohoku.ac.jp

This work was partly by JST SPRING, Grant Number JPMJSP2114.

Keywords: Modal Interval, Regression Analysis, Spline Smoothing, Data Visualization.

where $\mathcal{S}_d^\rho(\sqcup_b)$ is the set of all univariate spline functions of degree d at most and smoothness ρ with b polynomials, $w_i > 0$ is weight, $\mathcal{L} : \mathbb{R} \rightarrow [0, \infty)$ is loss function, $\lambda > 0$ is the regularization parameter to balance bias and variance to prevent overfitting.

2.4 Generalization of Nadaraya-Watson Estimator

The generalization of Nadaraya-Watson estimator can be used to estimate conditional modal interval [1, 2]. Conditional cumulative distribution function of Y given $X = x$ can be estimated by

$$\hat{F}_{Y|x}(y) := \frac{\sum_{i=1}^N \mathcal{K}_h(x - x_i) \mathcal{I}(y_i \leq y)}{\sum_{i=1}^N \mathcal{K}_h(x - x_i)}, \quad (3)$$

where $\mathcal{K}_h(\cdot)$ is a kernel function with bandwidth h , and $\mathcal{I}(y_i \leq y) = 1$ when $y_i \leq y$, otherwise $\mathcal{I}(y_i \leq y) = 0$. The $100\alpha\%$ conditional modal interval for given $X = x$ is estimated as

$$\underset{[y_{\text{low}}, y_{\text{up}}]}{\operatorname{argmin}} y_{\text{up}} - y_{\text{low}} \text{ s.t. } \hat{F}_{Y|x}(y_{\text{up}}) - \hat{F}_{Y|x}(y_{\text{low}}) \geq \alpha. \quad (4)$$

3 Spline Modal Interval Regression

Our proposed method is formulated as the following optimization problem

$$\begin{aligned} \underset{s, \bar{s} \in \mathcal{S}_d^\rho(\sqcup_b)}{\operatorname{minimize}} \quad & \sum_{i=1}^N w_i \mathcal{J}_{\hat{p}_{\text{low}}^{(i)}}(y_i - \underline{s}(x_i)) + \lambda \int_I |\underline{s}''(x)|^2 dx \\ & + \sum_{i=1}^N w_i \mathcal{J}_{\hat{p}_{\text{up}}^{(i)}}(y_i - \bar{s}(x_i)) + \lambda \int_I |\bar{s}''(x)|^2 dx \\ \text{subject to } & \forall x \ \underline{s}(x) \leq \bar{s}(x), \end{aligned} \quad (5)$$

where $\hat{p}_{\text{low}}^{(i)} = \hat{F}_{Y|x_i}(y_{\text{low}})$ and $\hat{p}_{\text{up}}^{(i)} = \hat{F}_{Y|x_i}(y_{\text{up}})$, which are estimated with the generalization of Nadaraya-Watson estimator in (4). $w_i = \sqrt[5]{\hat{f}(x_i)}$ and $\hat{f}(x_i)$ is estimated by kernel density estimation method. $\underline{s}(x)$ and $\bar{s}(x)$ are lower curve and upper curve of conditional modal interval, respectively. The condition $\forall x \ \underline{s}(x) \leq \bar{s}(x)$ is to ensure that the two curves will not cross each other.

4 Numerical Experiments

In numerical experiments, we estimate the 50% conditional modal interval of several simulation datasets with different distributions and sample sizes. Numerical experiments show that the results of proposed method in (5) have higher accuracy than the results of method in [2], since proposed method does not rely on bin segmentation. We also used proposed method to analyze the relationship between temperature and water vapor pressure in meteorological data [5]. Compared with general regression methods, the proposed method can not only visualize the relationship between two variables, but also visualize the changing trend of the spread of conditional distribution.

References

- [1] Jan G. De Gooijer and Ali Gannoun, Nonparametric conditional predictive regions for time series, *Computational Statistics & Data Analysis*, **33**, No.3, (2000), 259-275.
- [2] S. Yao, D. Kitahara, H. Kuroda and A. Hirabayashi, Modal Interval Regression Based on Spline Quantile Regression, *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, **E106.A**, No.2, (2023), 106-123.
- [3] R. Koenker and G. Bassett Jr., Regression quantiles, *Econometrica*, **46**, No.1, (1978), 33-50.
- [4] S. Wold, Spline functions in data analysis, *Technomet.*, **16**, No.1, (1974), 87-111.
- [5] Meteorological data of Beijing, China, from August 1 to 7, 2021, China Meteorological Data Service Centre, <http://data.cma.cn/en>.

Characterization of Asymptotic Distribution in Martingale Convergence of the Supercritical Branching Processes with Poissonian Offsprings

Ankhubayar Bat-Erdene¹ and Shuji Kawasaki²

¹ Graduate School of Science and Engineering, Iwate University

² Faculty of Science and Engineering, Iwate University

Abstract. A branching process (BP) $\{Z_n \mid n \in \mathbb{N}_0\}$ is a stochastic model of population increase by the proliferation of individuals. A normalized BP forms a martingale and converges, as $n \rightarrow \infty$, almost surely. The asymptotic distribution on $(0, \infty)$ of the martingale convergence is suggested to be used in various applications such as the potential theory of BP. Nevertheless, the explicit expression of the asymptotic distribution is unknown. We propose a characterization of the distribution based on the analysis of a functional equation for the Laplace transform (LT) of the distribution. Particularly, we present the Laguerre expansion of the asymptotic distribution. It turns out that, by numerical validation, our proposed model agrees well with the true distribution.

1. BACKGROUND AND PROBLEM

Let X be a random variable (r. v.) taking values in $\mathbb{N}_0 = \{0, 1, 2, \dots\}$, the set of nonnegative integers, and let $\mathbf{p} = \{p_k \mid k \in \mathbb{N}_0\}$ be its distribution. Then, a branching process (BP) $\{Z_n \mid n \in \mathbb{N}_0\}$ is defined by

$$Z_0 = 1, \quad Z_n = \sum_{i=1}^{Z_{n-1}} X_{n,i}, \quad n \in \mathbb{N},$$

where $\{X_{n,i} \mid i \in \mathbb{N}\}$ is an independently identically distributed sequence of r.v.s with $X_{n,i} \sim X$ for each $n \in \mathbb{N}_0$. That is, each of the i -th individuals in $(n-1)$ -th generation produces $X_{n,i}$ offspring according to the same distribution $\mathbf{p}$ independently of each other to form the n -th generation individuals. Because $Z_0 = 1$, $Z_1 \sim X$ especially. For a BP, the extinction and survival probabilities are, respectively, defined by

$$(1.1) \quad \eta = \mathbb{P}(\exists n : Z_n = 0) \quad \text{and} \quad \zeta = \mathbb{P}(Z_n > 0, \forall n \geq 0) = 1 - \eta.$$

If $\lambda = \mathbb{E}[Z_1] \leq 1$, the population dies out with extinction probability $\eta = 1$. If $\mu > 1$, the process is called supercritical. In this latter case, the extinction probability satisfies $\eta < 1$. Let $W_n = \mu^{-n} Z_n$, $n \in \mathbb{N}_0$. Then, $\{W_n\}$ is known [1, Theorem I.6.1] to form a martingale and converges to an r. v. W_∞ on $\mathbb{R}_+ = [0, \infty)$ almost surely as $n \rightarrow \infty$.

W_∞ is known to have a probability density function (p.d.f.) on $(0, \infty)$ [1, Corollary I.12.1], which we denote by $w : (0, \infty) \mapsto \mathbb{R}_+$, while W_∞ has a point mass at the origin, $P(W_\infty = 0) = \eta$ [1, Theorem I.6.2]. Thus, we may write $W_\infty \sim \eta\delta(x) + w(x)$ on $x \in \mathbb{R}_+$, where $\delta(x)$ denotes the Dirac delta [2, Theorem I.8.3].

However, not much is known about the p.d.f. $w(x)$ so far [3, p. 97], despite its importance in BP. Because the expression of $w(x)$ is unknown, further theory of BP has been developed with the unknown $w(x)$ as it is (see e. g., [1]). The objective of this study is to characterize the p.d.f. Particularly, we identify $w(x)$ in the form of Laguerre expansion.

2. CHARACTERIZATION AND MAIN RESULTS

We assume $\mathbf{p}$ is Poissonian and let $f : [0, 1] \mapsto [0, 1]$ be the probability generating function for $\mathbf{p}$. Then, $f(s) = e^{\lambda(s-1)}$, $s \in [0, 1]$. Furthermore, let $\varphi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be the LT of W_∞ : $\varphi(u) = \int_{(0,\infty)} e^{-ux} w(x) dx$. For φ and f , the following functional equation, called *Abel's equation*, holds [1]:

$$(2.1) \quad \varphi(u) = f\left(\varphi\left(\frac{u}{\lambda}\right)\right) = \exp\left[\lambda \varphi\left(\frac{u}{\lambda}\right)\right], \quad u \in \mathbb{R}_+.$$

Especially, taking $u = 0$ or ∞ in (2.1) corresponds in general to the two solutions of the equation $s = f(s)$: $s = 1 = \varphi_0(0)$ or $s = \eta = \varphi_0(\infty)$ [1, Corollary I.10.3], respectively. Thus, McLaurin's expansion of the LT $\varphi(u) = \sum_{n \in \mathbb{N}_0} \frac{\varphi^{(n)}(0)}{n!} u^n$. We consider McLaurin's coefficient $\varphi^{(n)}(0)$ and characterize the distribution $w(x)$.

The following are the main results of the study.

Proposition 1. $\varphi^{(n)}(0)$ has the following form:

$$(2.2) \quad \varphi^{(n)}(0) = -K_n(-\rho)^{n-1}, \quad \text{for } n \in \mathbb{N}_0,$$

where $\rho \triangleq \frac{\lambda}{\lambda - 1}$, and $K_n = K_n(\lambda)$ is a positive-valued function defined recursively by

$$(2.3) \quad K_1(\lambda) \equiv 1, \quad K_n(\lambda) = \frac{1}{\rho_{n-1}} \sum_{l=1}^{n-1} \binom{n-1}{l} \lambda^{-(n-1-l)} K_{n-l}(\lambda) K_l(\lambda), \quad \text{for } n \geq 2;$$

The function K_n is irreducible of ρ , i.e., K_n is not a multiple of ρ .

Theorem 2. The p.d.f. $w(x)$ has the Laguerre expansion

$$(2.4) \quad w(x) = e^{-x} \sum_{r \in \mathbb{N}_0} q_r L_r(x), \quad x \in \mathbb{R}_+$$

in $L^2(\mathbb{R}_+)$, for a real-valued sequence $\{q_n\} \in \ell^2(\mathbb{N}_0)$ given by

$$(2.5) \quad q_r = \sum_{m=0}^r \binom{r}{m} \frac{K_{m+1}}{m!} (-\rho)^m.$$

The constructed $w(x)$ is indeed well-defined in $L^2(\mathbb{R}_+)$ as can be seen from the following proposition. In the following, let $\mathbf{q}_n = [q_1, q_2, \dots, q_n]^\top$.

Proposition 3.

$$(2.6) \quad \sup_{n \in \mathbb{N}} \|\mathbf{q}_n\|_{\ell^2} < \infty.$$

REFERENCES

- [1] Athreya, K. B., and Ney, P. E., *Branching Processes*, Dover Publ., 2004.
- [2] Harris, T. E. *The Theory of Branching Processes*, Dover Publ., 1963.
- [3] Hofstad, R., *Random Graphs and Complex Networks*, Cambridge Univ Press, 2016.

Application of the Topological Data Analysis to Cosmology (位相的データ解析の宇宙論への応用)

竹内 努^{1,2}, 本武陽一³, 西澤 淳^{4,5,6}, 村上 広椰¹, 加納 龍生¹, 馬 海霞¹, 内田 舜也¹,
Suchetha COORAY⁷, 河野 海¹

1. 名古屋大学理学研究科素粒子宇宙物理学専攻, 2. 統計数理研究所統計的機械学習研究センター,
3. 一橋大学ソーシャル・データサイエンス研究科, 4. 岐阜聖徳学園大学 DX 推進センター 5. 名古屋大
学高等研究院, 6. 名古屋大学素粒子宇宙起源研究所, 7. 国立天文台科学研究部

銀河とは、星と星間物質(ガスと星間塵)、暗黒物質からなる巨大な天体である。銀河は現在観測可能な宇宙に数千億個存在しており、我々の目に見える波長(可視光線)での宇宙の姿を形作っている。しかし宇宙誕生当時、物質分布はほぼ一様であった。銀河は平均からわずかに密度が高い領域が重力によって成長し、合体成長を経て現在の姿へと進化してきたのである(図 1)。銀河進化は周囲の銀河の密度やガス密度など、銀河の置かれた環境にも強く依存する極めて複雑な過程である。そして銀河の形成と進化は 138 億年の宇宙進化の歴史の中でも非常に重要な現象のひとつであると考えられている。

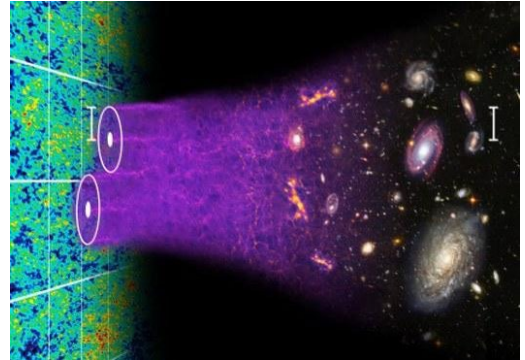


図 1: 一様な物質分布から銀河が形成され、進化する様子。

我々は 130 数億年にわたる銀河の形成・進化という複雑な物理現象を、これまでの宇宙物理学の方法とはまったく異なる、データ科学の最新手法である位相的データ解析(topological data analysis: TDA)に基づくアプローチによって新たな角度から解明することを試みている。本研究は特に、銀河形成以前の通常物質(宇宙物理学ではバリオンと呼ぶ)が光との相互作用によって生じる粗密波である「バリオン音響振動」(baryon acoustic oscillation: BAO)が銀河分布に残す痕跡について扱う。バリオンと光の相互作用が切れる宇宙の晴れ上がり(脱結合)時に物質が放射した熱放射は宇宙膨張によって波長が引き伸ばされ、現在は電波(ミリ波)の波長で宇宙マイクロ波背景放射(cosmic microwave background: CMB)として観測される。BAO は CMB のゆらぎに特徴的なスケールとして現れ、ゆらぎをフーリエ変換したパワースペクトルにおいて最も顕著なピークとして観測される(図 2)。このスケールは共動距離で観測した際に、どの赤方偏移においてもおよそ 148 Mpc ($1 \text{ Mpc} = 3.08 \times 10^{24} \text{ cm}$)の一定値をとる。この性質から、BAO シグナルの観測的スケールは宇宙論パラメータの推定の最も信頼

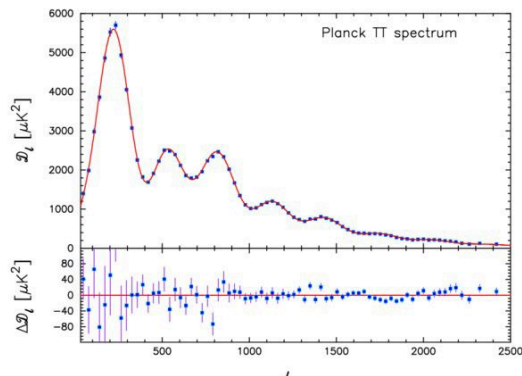


図 1: Planck 衛星が観測した CMB のパワースペクトル。最も顕著なピークが BAO である。

られ、現在は電波(ミリ波)の波長で宇宙マイクロ波背景放射(cosmic microwave background: CMB)として観測される。BAO は CMB のゆらぎに特徴的なスケールとして現れ、ゆらぎをフーリエ変換したパワースペクトルにおいて最も顕著なピークとして観測される(図 2)。このスケールは共動距離で観測した際に、どの赤方偏移においてもおよそ 148 Mpc ($1 \text{ Mpc} = 3.08 \times 10^{24} \text{ cm}$)の一定値をとる。この性質から、BAO シグナルの観測的スケールは宇宙論パラメータの推定の最も信頼

できる指標のひとつとなっている。

従来, BAO の検出には銀河分布の2点相関関数やパワースペクトルが用いられてきたが, BAO シグナルは銀河の重力的クラスタリングによって生じるシグナルに対して 1/100 程度と微弱であるため, 統計的に有意な結果を得るためには大規模かつサンプリングレートの高い銀河探査が必要であり, 現存する最大の銀河探査であるスローンデジタルスカイサーベイ (Sloan Digital Sky Survey: SDSS) によって初めてそのシグナルが検出された. この規模の探査は決して容易ではなく, 独立な, そしてより有効な検出方法が望まれている。

我々は, 銀河の空間分布がもつ位相幾何学的情報を抽出することで BAO シグナルの検出や宇宙大規模構造を解析することを目的として, パーシステントホモロジー(PH) の計算を行った. PH はトポロジカルデータ解析(TDA) の中心的方法論であり, ホモロジー群に対して関心のあるパラメータに関するフィルトレーションによって定義される. さらに, 穴の位置や形状などの宇宙論的に重要な情報を得ることができる点が特徴である. 本研究では, 共動距離(宇宙膨張の影響を取り除いた本質的な距離)をパラメータにとることにより, 空間に対する解像度を变化させた際に生じる構造を次元ごとに解析することができる。

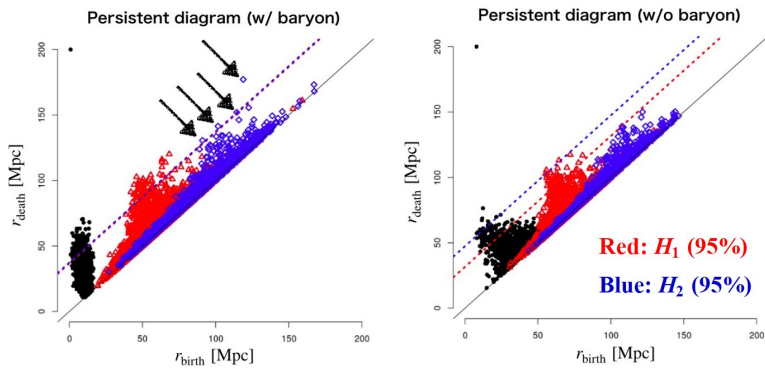


図 3: 宇宙論的シミュレーションデータから求めたパーシステント図. 左がバリオン有, 右がなしの場合を示す。

り, BAO の理論と整合的であった. さらに, SDSS eBOSS サーベイによって得られた実際のクエーサー観測データに対しても同様の解析を行い, H_2 に対して $\bar{r} = 146.6 \pm 2.0$ [Mpc] という推定値を得た. また本研究では PH の逆解析も行い, 得られた H_2 に対応する実際の宇宙の構造を可視化することに成功した(図 4). 銀河分布に対しては世界初の例である. これらの計算結果に H_1, H_0 についての考察を加味し, 銀河の空間分布に関して得られた結果を報告する。

我々はまず TDA の BAO 検出に対する有用性の検証のため, バリオン有り・無しの二通りの N 体シミュレーションデータに対して PH の計算を行った. その結果, バリオン有の場合にのみ有意な 2 次ホモロジー(H_2), すなわち中空のシェル構造を検出した(図 3). また, 検出された H_2 構造の半径の平均値は $\bar{r} = 150.16 \pm 8.46$ [Mpc] であ

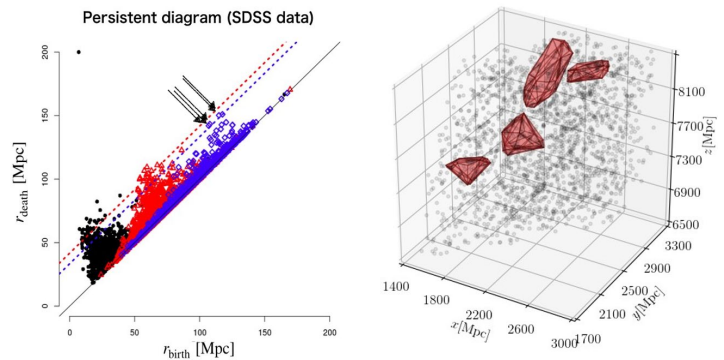


図 4: SDSS eBOSS データから求めた PH (左), およびその逆解析によって得られた銀河分布の対応構造。

代表的な接合関数の裾従属係数と裾次数*

東京都立大学 大学院経営学研究科 吉羽 要直†

金融機関のリスク管理においては、代表的な資産価格変動や信用状態を d 変量のリスクファクターとして、その d 変量の従属関係を正規接合関数や t 接合関数といった楕円接合関数で捉えることが多い。しかしながら、正規接合関数は変量間の裾従属性が弱く、 t 接合関数はリスクファクターの上下の変動に対して対称な従属関係しか表現できない。そのため、正規接合関数や t 接合関数に非対称性を入れたモデリングも発展してきている。本研究では、実務的に用いられることが多い代表的な接合関数の裾従属性について、 $d = 2$ 変量に焦点を当てて整理する。具体的には伝統的に用いられている裾従属係数とともに、弱い裾従属関係を捉える裾次数について、Joe (2014), 吉羽(2020, 2021), Yoshida, Koike and Kato (2023)の研究の成果に沿って整理する。

一般に、2変量確率ベクトル $\mathbf{X} = (X_1, X_2)^\top$ の同時分布関数 $F(x_1, x_2)$ は、各変量が連続確率変数であれば、各周辺分布関数 $F_1(x_1), F_2(x_2)$ と接合関数 $C(u_1, u_2)$ を用いて

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)) \quad (1)$$

と一意に表現できる。確率ベクトル $\mathbf{X}$ が相関 ρ の2変量標準正規分布に従う場合、内包される接合関数が相関 ρ の正規接合関数 $C_N(u_1, u_2; \rho)$ である。 $\mathbf{X}$ が相関 ρ 、自由度 ν の2変量 t 分布に従う場合は、相関 ρ 、自由度 ν の t 接合関数 $C_t(u_1, u_2; \rho, \nu)$ が内包される接合関数である。確率ベクトル $\mathbf{X}$ に非対称性を入れると、その同時分布に内包される非対称接合関数を定義できる。

一方、2変量の従属関係については、アルキメデス型の接合関数が用いられることもあり、代表的なものに1パラメータ θ により以下で表せる Gumbel, Clayton, Frank の各接合関数がある。

$$\text{Gumbel} \quad C_G(u_1, u_2; \theta) = \exp \left[-\{(-\ln u_1)^\theta + (-\ln u_2)^\theta\}^{1/\theta} \right], \quad \theta > 1 \quad (2)$$

$$\text{Clayton} \quad C_C(u_1, u_2; \theta) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{1/\theta}, \quad \theta > 0 \quad (3)$$

$$\text{Frank} \quad C_F(u_1, u_2; \theta) = -\frac{1}{\theta} \ln \left\{ 1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1} \right\}, \quad \theta \in \mathbb{R} \quad (4)$$

裾従属性は裾次数 κ_L あるいは κ_U が1の場合は裾従属係数 λ_L あるいは λ_U で従属性の強さを捉えられ、裾次数が(1,2)の範囲であれば、裾次数で従属性の強さを捉えられる。代表的な2変量接合関数について、下側・上側それぞれの裾従属係数 λ_L, λ_U あるいは裾次数 κ_L, κ_U は、Joe (2014)の Table 4.1 と Clayton (Mardia–Takahasi–Clayton–Joe) 接合関数に関する 4.6.1 節の議論より、自由度 ν の1変量 t 分布関数を $T_\nu(\cdot)$ として表1のようにまとめられる。

相関 $\tilde{\rho}$ の2変量標準正規確率ベクトル $\tilde{\mathbf{Z}} = (\tilde{Z}_1, \tilde{Z}_2)^\top$ とそれとは独立な標準正規確率変数 Z_0 、非対称性を表すパラメータ $\boldsymbol{\delta}^\top = (\delta_1, \delta_2) \in (-1, 1)^2$ を用いて、2変量確率ベクトル $\mathbf{Y} = (Y_1, Y_2)^\top$ を

$$Y_j = \delta_j |Z_0| + \sqrt{1 - \delta_j^2} \tilde{Z}_j, \quad j = 1, 2 \quad (5)$$

* 本研究は JSPS 科研費 JP21K01581 からの助成を受けている。

† E-mail: tyoshida@tmu.ac.jp

とすると、 Y_1 と Y_2 の相関は ρ となり、 $\mathbf{Y}$ は非対称性 δ の2変量非対称正規分布 $SN(\rho, \delta)$ に従い、内包される接合関数として非対称正規接合関数 $C_{SN}(u_1, u_2; \rho, \delta_1, \delta_2)$ が定義される。非対称正規接合関数は、正規接合関数と同様に、上下ともに裾従属係数はゼロで漸近独立であり、裾次数については、Yoshiba, Koike and Kato (2023)で以下のように整理されている。

$$\kappa_L(C_{SN}; \delta_1, \delta_2 \leq 0) = \kappa_U(C_{SN}; \delta_1, \delta_2 \geq 0) = \frac{2}{1 + \rho} \quad (6)$$

$$\kappa_L(C_{SN}; \delta_1, \delta_2 \geq 0) = \kappa_U(C_{SN}; \delta_1, \delta_2 \leq 0) = \frac{2}{1 + \tilde{\rho}} \quad (7)$$

表 1 代表的な2変量接合関数の裾従属係数と裾次数

接合関数	下側 λ_L/κ_L	上側 λ_U/κ_U
正規	$\kappa_L = \frac{2}{1 + \rho}$	$\kappa_U = \kappa_L = \frac{2}{1 + \rho}$
t	$\lambda_L = 2T_{v+1} \left(-\sqrt{\frac{(1-\rho)(v+1)}{(1+\rho)}} \right)$	$\lambda_U = \lambda_L$
Gumbel	$\kappa_L = 2^{1/\theta}$	$\lambda_U = 2 - 2^{1/\theta}$
Clayton	$\lambda_L = 2^{-1/\theta}$	$\kappa_U = 2$
Frank	$\kappa_L = 2$	$\kappa_U = \kappa_L = 2$

2変量確率ベクトル $\mathbf{Y} \sim SN(\rho, \delta)$ と独立にガンマ分布に従う確率変数 $V \sim G(v/2, v/2)$ を用いて、 $\mathbf{X} = \mathbf{Y}/\sqrt{V}$ で Azzalini and Capitanio (2003)の非対称 t 分布に従う d 変量確率ベクトル $\mathbf{X}$ が構成され、内包される接合関数としてAC非対称 t 接合関数 $C_{ACSt}(u_1, u_2; \rho, v, \delta_1, \delta_2)$ が定義される。吉羽 (2020, 2021)で示されているように、 $C_{ACSt}(u_1, u_2; \rho, v, \delta_1, \delta_2)$ は漸近従属で、裾従属係数 λ_L, λ_U は非対称性 δ_1, δ_2 によって異なる。

一方、相関 ρ の2変量標準正規確率ベクトル $\mathbf{Z}$ と $V \sim G(v/2, v/2)$ を用いて、非対称性パラメータを $\boldsymbol{\gamma} = (\gamma_1, \gamma_2) \in \mathbb{R}^2$ として、 $\mathbf{X} = \boldsymbol{\gamma}V^{-1} + \mathbf{Z}/\sqrt{V}$ により、一般化双曲型 (generalized hyperbolic) の非対称 t 分布に従う d 変量確率ベクトル $\mathbf{X}$ が構成され、内包される接合関数としてGH非対称 t 接合関数 $C_{GHSt}(u_1, u_2; \rho, v, \gamma_1, \gamma_2)$ が定義される。この裾従属係数は、 $\gamma_1 = \gamma_2 = \gamma$ での特殊ケースで $\gamma < 0$ ならば $\lambda_L = 1, \lambda_U = 0$ となり、 $\gamma > 0$ ならば $\lambda_L = 0, \lambda_U = 1$ で与えられる。裾従属係数がゼロで漸近独立の場合の裾次数の分析は、今後の課題である。

参考文献

- Azzalini, A. and Capitanio, A. (2003), “Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t -distribution,” *Journal of the Royal Statistical Society Series B*, **65**(2), 367–389.
- Joe, H. (2014), *Dependence Modeling with Copulas*, Boca Raton, Chapman & Hall/CRC.
- 吉羽 要直 (2020), 「非対称 t 接合関数の性質と統計的推定方法」『統計数理』**68**(1), 45–63.
- 吉羽 要直 (2021), 「極値での従属性および非対称性と信用ポートフォリオリスク」『日本統計学会誌』**51**(1), 157–178.
- Yoshiba, T., Koike, T. and Kato, S. (2023), “On a measure of tail asymmetry for the bivariate skew-normal copula,” *Symmetry*, **15**(7), 1410.

コピュラ・グラフィック推定量 – 理論と応用 –

江村剛志*

概要

生存時間解析の多くの手法は、打ち切り時間と生存時間が互いに独立であるという仮定（独立打ち切りの仮定）下で妥当である。一方、この仮定が成立しない「従属打ち切り」に起因するバイアスの問題が生存時間解析の文脈において活発に議論されている。例えば、従属打ち切り下における Kaplan-Meier 推定量のバイアスがよく議論される。本研究では、Kaplan-Meier 推定量に代わる推定量である、Copula-Graphic 推定量を紹介し、その理論と応用例を示す。

キーワード: Kaplan-Meier 推定量・コピュラ・Copula-Graphic 推定量・生存関数・2 標本検定・ノンパラメトリック検定・右打ち切り

1. はじめに

「従属打ち切り」とは、生存時間変数と打ち切り変数が互いに独立でない場合を指すが、通常の生存時間解析は、独立性の仮定を置いて行われる。この「独立打ち切り」の仮定が満たされないような状況は、生存時間解析の応用において散見される。これに関連する諸問題は生存時間データ解析の文脈において広く議論されてきた (Braekers and Veraverbeke(2005), Emura and Michimae(2017), Staplin *et al.* (2015), Xu *et al.* (2018), Moradian *et al.* (2019), Schneider *et al.* (2020, 2022), Deresa and Van Keilegom(2020a, b), Huang *et al.* (2021), Li *et al.* (2021), Lo and Wilke(2010), Deresa *et al.* (2022)). 独立打ち切りの仮定が満足されない場合、それを基にした Kaplan-Meier 推定量などは一致性を持たない（推定量が漸近バイアスを持つ）ため、解析結果はバイアスを持つ。

従属打ち切りによる漸近バイアスの研究結果がいくつか報告されている。従属打ち切り下で Rivest and Wells(2001)は Kaplan-Meier 推定量の漸近バイアスを解析的に導出した。また Emura and Chen(2016, 2018)は Cox 回帰法の漸近バイアスを解析的・数值的に調べた。Emura and Hsu(2020)は、2 標本問題における治療効果の漸近バイアスを導出した。道前(2023)は、競合リスクによる従属打ち切りが尤度推定量に与えるバイアスをシミュレーションで検証した。Emura *et al.* (2024)は因子実験計画における治療効果の推定バイアスをシミュレーションで検証した。上記の全ての研究結果では「コピュラ」(Nelsen 2006)で生存時間と打ち切り時間の従属構造をモデリングし、バイアスの構造を解析している。

本稿では、従属打ち切りの影響をコピュラで調節するために導出された Copula-Graphic 推定量を紹介する。この推定量は、生存時間と打ち切り時間の相関構造をコピュラで正しく特定することにより、バイアスの無い生存関数の推定を行うものである。

*統計数理研究所, Email:takeshiemura@gmail.com

る. さらに, それを応用した 2 標本検定法 (Emura and Hsu, 202)) と因子実験計画法 (Emura et al. 2024) を紹介する.

2. Copula-Graphic 推定量とその応用

本節では, 独立打ち切りのある生存データ解析の基本的な解説と, Copula-Graphic 推定量の定義を与える. またその 2 標本問題への応用を示す.

2.1 Copula-Graphic (CG) 推定量

以下で生存関数や打ち切りの基本的な定義を与え, Copula-Graphic 推定量を定義する.

2 標本問題を考える. 2 つの群のいずれか ($i = 1, 2$) に属する個体 ($j = 1, \dots, n_i$) の生存時間を T_{ij} で表し, その生存関数を $S_i(t) := P(T_{ij} > t)$ とする. n_i は群 i における個体数である. 合計 $n := n_1 + n_2$ 個の個体は全て独立であると仮定する. 本稿では T_{ij} がハザード関数を $\lambda_i(t) := f_i(t)/S_i(t)$ を持つとする.

打ち切り時間を U_{ij} , その生存関数を $G_i(u) := P(U_{ij} > u)$ とおく. 2 群の生存データを $(X_{ij}, \delta_{ij}); i = 1, 2; j = 1, \dots, n_i$ と定義し, ここで $X_{ij} := \min(T_{ij}, U_{ij})$, $\delta_{ij} := I\{T_{ij} \leq U_{ij}\}$, $I\{\cdot\}$ は指標関数とする.

また, $Y_{ij}(s) := I\{X_{ij} \geq s\}$, $\bar{Y}_i(s) := \sum_{j=1}^{n_i} Y_{ij}(s)$ を定義する.

従属打ち切り下で $S_i(\cdot)$ を推定することを考える. 以下では, 打ち切りの従属構造をコピュラでモデリングすることにより $S_i(\cdot)$ を一致推定することを考える.

まず T_{ij} と U_{ij} の同時分布が次のアルキメデスコピュラで特定されるとする:

$$P(T_{ij} > t, U_{ij} > u) = \phi_{i\theta_i}^{-1}[\phi_{i\theta_i}\{S_i(t)\} + \phi_{i\theta_i}\{G_i(u)\}], \quad i = 1, 2.$$

ここで $\phi_{i\theta_i}(\cdot)$ は生成素と呼ばれ θ_i , $i = 1, 2$, はパラメータである. 定義より, 生成素は連続な減少関数であり, $\phi_{i\theta_i}(0) = \infty$, $\phi_{i\theta_i}(1) = 0$ を満足する.

クレイトン型コピュラで生成素は $\phi_{i\theta_i}(t) = (t^{-\theta_i} - 1)/\theta_i$, $\theta_i > 0$ であり, このコピュラでの相関の強さを表すケンドール相関は $\tau_{\theta_i} = \theta_i/(\theta_i + 2)$ である. グンベル型コピュラで生成素は $\phi_{i\theta_i}(t) = \{-\log(t)\}^{\theta_i+1}$, $\theta_i \geq 0$ であり, ケンドール相関は $\tau_{\theta_i} = \theta_i/(\theta_i + 1)$ である.

Rivest and Wells(2001)は次式で Copula-Graphic 推定量を提案した:

$$\hat{S}_i^{CG}(t) := \phi_{i\theta_i}^{-1} \left[\sum_{j: X_{ij} \leq t, \delta_{ij}=1} \phi_{i\theta_i} \left(\frac{\bar{Y}_i(X_{ij}) - 1}{n_i} \right) - \phi_{i\theta_i} \left(\frac{\bar{Y}_i(X_{ij})}{n_i} \right) \right], \quad 0 \leq t \leq \max_j(X_{ij}).$$

生成素を $\phi_{i\theta_i}(t) = -\log(t)$ とおくと, アルキメデスコピュラは独立コピュラを生成し, $\hat{S}_i^{CG}(t) = \hat{S}_i(t) = \prod_{j: X_{ij} \leq t, \delta_{ij}=1} \{1 - 1/\bar{Y}_i(X_{ij})\}$ となる. よって, Copula-Graphic 推定量は Kaplan-Meier 推定量を一般化した推定量である.

クレイトン型コピュラの下での Copula-Graphic 推定量は

$$\hat{S}_i^{CG}(t) = \left[1 - \sum_{j: X_{ij} \leq t, \delta_{ij}=1} \left\{ \left(\frac{\bar{Y}_i(X_{ij}) - 1}{n_i} \right)^{-\theta_i} - \left(\frac{\bar{Y}_i(X_{ij})}{n_i} \right)^{-\theta_i} \right\} \right]^{-\frac{1}{\theta_i}},$$

グンベル型コピュラの下での Copula-Graphic 推定量は

$$\hat{S}_i^{CG}(t) = \exp \left[- \left(\sum_{j: X_{ij} \leq t, \delta_{ij}=1} \left[\left\{ -\log \left(\frac{\bar{Y}_i(X_{ij}) - 1}{n_i} \right) \right\}^{\theta_i+1} - \left\{ -\log \left(\frac{\bar{Y}_i(X_{ij})}{n_i} \right) \right\}^{\theta_i+1} \right] \right)^{\frac{1}{\theta_i+1}} \right]$$

で与えられる．これら推定量は R パッケージ *compound.Cox* (Emura *et al.* 2019a) の R 関数 `CG.Gumbel(.)`, `CG.Clayton(.)`, `CG.Frank(.)` で計算することが出来る．

この Copula-Graphic 推定量は，モデルが正しく特定されていれば生存関数を一致推定できる．これを応用し，2 標本検定などを行うことができる．また，3 群以上の場合や，2 つ以上の因子を持つ複雑な群構造を構成する分散分析も行うことができる．そのほか，生存木を構築したり，予後予測を行う際にも，Copula-Graphic 推定量は有用となる．詳細は，講演にて説明する．

参考文献

- Braekers, R. and Veraverbeke, N. (2005). A copula-graphic estimator for the conditional survival function under dependent censoring, *Canadian Journal of Statistics*, **33**, 429-447.
- Deresa, N. W. and Van Keilegom, I. (2020a). Flexible parametric model for survival data subject to dependent censoring, *Biometrical Journal*, **62**, 136-156.
- Deresa, N. W. and Van Keilegom, I. (2020b). A multivariate normal regression model for survival data subject to different types of dependent censoring, *Computational Statistics & Data Analysis*, **144**, 106879.
- Deresa, N. W., Van Keilegom, I. and Antonio, K. (2022). Copula-based inference for bivariate survival data with left truncation and dependent censoring, *Insurance: Mathematics and Economics*, **107**, 1-21.
- Emura, T., and Chen, Y.-H. (2016). Gene selection for survival data under dependent censoring, a copula-based approach, *Statistical Methods in Medical Research*, **25**, 2840-2857.
- Emura, T., and Chen, Y.-H. (2018). *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*, Springer.
- Emura, T., and Hsu, J.-H. (2020). Estimation of the Mann-Whitney effect in the two-sample problem under dependent censoring, *Computational Statistics and Data Analysis*, **150**, 106990.
- Emura, T., Matsui, S., and Chen, H.-Y. (2019a). compound. Cox: univariate feature selection and compound covariate for predicting survival, *Computer Methods and Programs in Biomedicine*, **168**, 21-37.

- Emura, T., and Michimae, H. (2017). A copula-based inference to piecewise exponential models under dependent censoring, with application to time to metamorphosis of salamander larvae, *Environmental and Ecological Statistics*, **24**, 151–173.
- Emura, T., Dizhaus M., Dobler, M., Murotani, K. (2024). Factorial survival analysis for treatment effects under dependent censoring, *Statistical Methods in Medical Research*, in press.
- Huang, X., Wang, W., and Emura, T. (2021). A copula-based Markov chain model for serially dependent event times with a dependent terminal event, *Japanese Journal of Statistics and Data Science*, **4**, 917-51.
- Li, D., Hu, X. J., and Wang, R. (2021). Evaluating association between two event times with observations subject to informative censoring, *Journal of the American Statistical Association*, doi.org/10.1080/01621459.2021.1990766.
- Lo, S. M., and Wilke, R. A. (2010). A copula model for dependent competing risks, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59**, 359-376.
- 道前洋史 (2023). 「競合リスクを伴う左側切断・右側打ち切りデータの解析～現状と今後の展望について～」『日本統計学会誌』, in press.
- Moradian, H., Larocque, D., and Bellavance, F. (2019). Survival forests for data with dependent censoring, *Statistical Methods in Medical Research*, **28**, 445-461.
- Nelsen, R. B. (2006). *An Introduction to Copulas*, Springer.
- Rivest, L.-P., and Wells, M. T. (2001). A martingale approach to the copula-graphic estimator for the survival function under dependent censoring, *Journal of Multivariate Analysis*, **79**, 138-155.
- Schneider, S., Demarqui, F. N., Colosimo, E. A., and Mayrink, V. D. (2020). An approach to model clustered survival data with dependent censoring, *Biometrical Journal*, **62**, 157-174.
- Schneider, S., Demarqui, F., and de Freitas Costa, E. (2022). Free-ranging dogs' lifetime estimated by an approach for long-term survival data with dependent censoring. *Environmental and Ecological Statistics*, **29**, 869–911.
- Staplin, N., Kimber, A., Collett, D., and Roderick, P. (2015). Dependent censoring in piecewise exponential survival models, *Statistical Methods in Medical Research*, **24**, 325-341.

医学研究における不完全データの感度解析

大阪大学大学院医学系研究科医学統計学教室 服部 聡

E-mail: hattoris@biostat.med.osaka-u.ac.jp

医学研究においては、治療法などの有用性を評価するために、多くの場合ヒトを対象にして研究が行われる。倫理面に配慮してデータを取得する必要があることから、意図通りにデータを取得する事は必ずしも容易ではなく、寧ろ不完全データを相手にせねばならないことの方が通常である。

複数の治療法を比較する研究において、被験者に治療法を無作為に割付ける無作為化研究は、一般的にもっとも信頼できる研究デザインとして広く受け入れられている。研究開始前に試験実施計画書を作成し、どのようなタイミングで評価項目の測定を行うかなど、詳細が事前に決定される。事前に周到に計画したとしても、副作用の発現や被験者からの申し出などの倫理的な理由から、試験途中で当該被験者の主要評価項目のフォローアップを中断せざるを得ない中途脱落の問題が生じえる。

無作為化研究は、組み入れられた被験者内での治療法の比較の妥当性、内的妥当性を極小の仮定で保証することが可能で、最も根拠レベルの高いゴールドスタンダードともいうべき研究デザインである。一方で、費用が高いなど実施上の困難が伴う。対象とする被験者が限定的であり、実際の医療で対象とする症例への有効性すなわち外的妥当性の担保には疑問が残る。このような背景から、電子カルテデータなどの既存データベースの解析により、効率的に無作為化研究を補足あるいは代替える気運が、リアルワールドエビデンスの名のもとに高まっている。このようなリアルワールドデータの解析は観察研究の解析に従うこととなるが、治療が無作為に割付けられていないことから、交絡による治療効果推定の偏りが問題となる。解析においては、交絡因子の影響を排除する必要があり、傾向スコア解析が頻繁に利用されている。

臨床試験における中途脱落に伴う欠測の問題、観察研究における交絡調整の問題などに対しては、傾向スコアによる逆確率重み付け (inverse probability weighting; IPW) 法を用いることが可能であり、実際の臨床研究での応用例も多い。通常、Missing At Random(MAR) あるいは強い意味での無視可能 (Strong Ignorable Treatment Assignment) 性のもとで、以下のステップで実行される。

- (1) 傾向スコアにロジスティック回帰等のパラメトリックモデルを想定する。
- (2) 未知パラメータを適当な推定方程式を解くことで推定する。多くの場合最尤法に基づくスコア方程式を解く。
- (3) 各被験者の傾向スコアを決定し、IPW 推定を行う。

傾向スコアのモデルが正しく特定されている場合には、IPW 推定量は平均因果効果の一致推定量を与えるが、傾向スコアを誤特定している場合や、SITA 条件が成立しない場合には、交絡の影響を調整できない。傾向スコアにより調整できるのは観測されている交絡因子に限定され、調整され損ねている未測定の影響因子の存在、残差交

絡の問題が、観察研究の解析では常に問題となる。SITA 条件を前提としない解析として、アウトカム依存傾向スコアが提案され、適当なパラメトリックモデルと推定方程式を用いることで、上述の (1)-(3) のステップで SITA 条件なしに IPW 推定を行うことができる (Miao et al. 2016) が、識別性の問題があり、十分に柔軟な解析ができない場合がある。未観測交絡に対する感度解析法が広く研究され、実際の臨床研究での適用例も見られるが、何らかの検証できない仮定に基づいているなど十分に使いやすいものとなっておらず、簡明かつ有効な方法の確立が望まれる。

本研究では、傾向スコアの IPW 推定に基づいて提案された、ある簡明な感度解析法の異議を検討する。アウトカム依存傾向スコアにおける (1)-(3) のステップにおいて、パラメトリックモデルの仮定 (1) を排除した上で、真の傾向スコアが取りえる範囲上での IPW 推定量の上限および下限を構成することで、平均因果効果の範囲を特定する方法を提案する。具体的には、(1) のモデルを要請しないため、(2) の推定のステップが必要がないものの、推定方程式を真の傾向スコアが満たすべき制約として利用して、線形計画法の簡単な適用で範囲を設定する。Tang et al. (2023) では、連続型あるいは二値データの場合の挙動を調べている。提案法のアイディアは簡明であることから、統一的に他の問題へ拡張できる可能性がある。生存時間アウトカムの場合への拡張は応用上重要である。メタアナリシスは複数の臨床試験の結果を併合して、より強固な根拠を得るための統計的方法である。公表バイアスの問題はメタアナリシスにおける不完全データの問題であり、特有の扱いがなされる。講演者らは、臨床試験登録を用いることで、公表バイアスの問題が通常的不完全データと同様のデータ構造を有する見なすことができることを指摘し (Huang et al., 2021)、IPW 法での調整法を提案した (Huang et al., 2022)。本研究での感度解析法は公表バイアスの最悪評価に用いることができる。時間が許せば、これらの進展中の研究についても触れる。

引用文献

- Tang C, Zhou Y, Huang A, Hattori S (2023). A simple sensitivity analysis method for unmeasured confounders via linear programming with estimating equation constraints. *arXiv:2312.04150*
- Miao W, Ding P, Geng Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *J Am Stat Assoc.* **111**, 1673–1683.
- Huang A, Komukai S, Friede T, Hattori S. (2021). Using clinical trial registries to inform Copas selection model for publication bias in meta-analysis. *Research Synthesis Methods* **12**, 658–673.
- Huang A, Morikawa K, Friede T, Hattori S. (2022). Adjusting for publication bias in meta-analysis via inverse probability weighting using clinical trial registries. *Biometrics* **79**, 2089–2102.

野菜の価格と気象との関連性について

石井 一夫(公立諏訪東京理科大学), 金打 聖菜 (公立諏訪東京理科大学)

【背景と目的】地球温暖化は年々進行し私達の生活を脅かすものとなっている。気温上昇や自然災害は、農産物の収穫量や品質に大きく影響し、食料安全保障上の懸念点となっている。野菜は、穀物などのように生産調整や価格調整が行われることが少なく、その価格に受容供給関係が直接に反映される事が多い。そこで、気候変動による経済影響を捉えるため野菜の価格と気象との関連性を統計手法により示すことを目的とした。

【方法】野菜の価格変動を捉えるため、株価の価格変動トレンドを分析するのによく用いられるローソク足チャートや移動平均線、MACD、RSIなどのテクニカル分析手法で用いられる指標を用いることを考えた。また、気象データ（気温、降水量など）や農産物の作況データとの経時的関連変化を、カイ二乗検定による p 値および $-\log p$ 値で評価した。

野菜の価格と生産量については、農林水産省の提供している青果物卸売市場調査（日別調査）[東京都]から、野菜の価格データおよび生産量データを入手した。気象データについては気象庁の提供している過去の気象データ・ダウンロードから各地の気象情報(日最高気温・日最低気温・日平均気温・降水量の日合計・日照時間)を入手した。分析対象期間としては、2010年10月1日～2023年4月10日までの日々のデータを使用した。データ分析を行った。なかでも、令和3年産野菜生産統計において、収穫量ランキングの2位との差が2倍以上のもの(日本における特定の野菜の生産の多くを1都道府県で担っているもの)に限定し、たまねぎ、キャベツ、レタス、トマトで分析を行った。

【結果】もっとも典型的な変動トレンドを示した北海道のタマネギの価格を用いた野菜の価格変動分析の結果を図1～6に示した。今回、検討したテクニカル分析の指標は、MACD(図2)、RSI(図3)、移動平均乖離率(図4)、ボリンジャーバンド(図5)、ストキャスティクス(図6)である。いずれの指標も価格変動に対応して、指標の変動を検出可能であった。



図1 ローソク足チャート



図2 ローソク足チャートと MACD



図3 ローソク足チャートとRSI



図4 ローソク足チャートと移動平均線乖離率

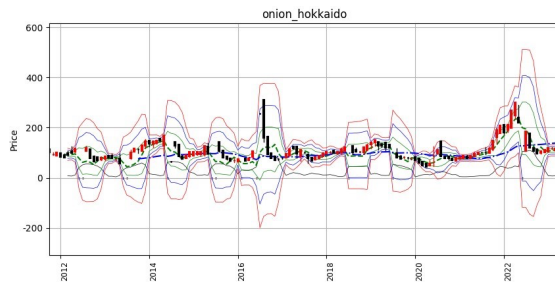


図5 ボリンジャーバンド

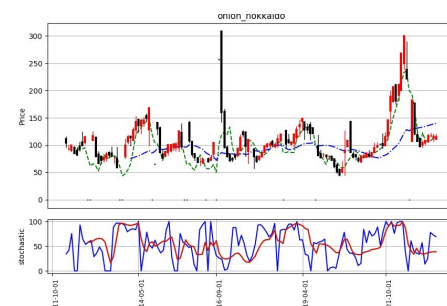


図6 ローソク足チャートとストキャスティクス

次に、野菜の価格変動（終値）と、野菜価格と、収穫量、降水量の関連性を評価した結果を図7～10に示した。各図は上から順に、価格のローソク足チャート、テクニカル検定結果(MACD/RSI[1年]), 検定対象(降水量/日照時間), カイ二乗検定(p 値), カイ二乗検定($-\log p$ の値)の結果である。

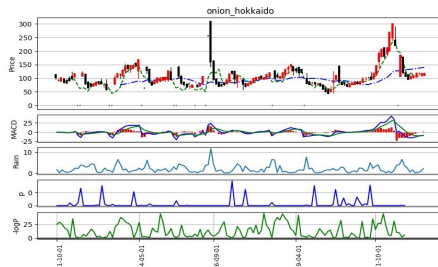


図7 MACD と降水量の関係



図8 MACD と日照時間の関係



図9 RSI(1年)と降水量の関係



図10 RSI(1年)と日照時間の関係

【考察】 このように、テクニカル分析指標である MACD, RSI, 移動平均線乖離率, ボリンジャーバンド, ストキャスティクスをもちいて野菜の価格変動を評価し、気候変動との関連について示すことが可能であった。

Sparse quantile regression via ℓ_0 -penalty

本田敏雄

一橋大学経済学研究科

本講演では以下の内容について報告した.

1. Best subset selection

まず本研究の動機となった, 情報量基準や ℓ_0 ペナルティーを用いて総当たりにより変数を選択する best subset selection 手法に関して, Bertsimas et al.(2016)[1], Hastie et al.(2020)[3], Hazimeh et al.(2023)[4] などの重要な文献について紹介した. [1] と [4] では混合整数計画法を用いている. [3] は, いくつかの手法について数値的な比較を行ったレビュー論文である. その他分位点回帰では, Chen and Lee(2023)[2] がある.

2. スパースな分位点回帰モデル, ℓ_0 回帰, 一致性

本研究では, (超) 高次元の線形モデル, 加法モデル, 変動係数モデルを統一的に扱っており, 特に線形モデル, 加法モデルについて報告した. τ 分位点線形モデルでは以下の通りである.

$$Y_i = \sum_{j=1}^p X_{ij}\beta_j^* + \epsilon_i \quad (1)$$

で, 非ゼロ要素の添え字集合 S_0 の推定量 $\widehat{S}$ は

$$Q_n(\widetilde{\beta}_S) \equiv R_n(\widetilde{\beta}_S) + \lambda \zeta_n |S| \frac{\log p_n}{n} \quad \text{かつ} \quad \widehat{S} \equiv \underset{S}{\operatorname{argmin}} Q_n(\widetilde{\beta}_S), \quad (2)$$

$$\widetilde{\beta}_S \equiv \underset{\beta_S}{\operatorname{argmin}} R_n(\beta_S) \quad \text{かつ} \quad R_n(\beta_S) \equiv \frac{1}{n} \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_{iS}^T \beta_S), \quad (3)$$

で定義される. ρ_τ は τ 分位点回帰のチェック関数, $p_n = p \vee n$, λ はチューニングパラメータ, ζ_n は適宜指定される無限大に発散する場合もある定数である. そして本研究では以下の二つの場合を扱った. $|S_0| (\ll p_n)$ は無限大に増加するとする.

1. $k_0 < |S_0| < k_1$ かつ $C|S_0| < k_0$ がわかっており, $\zeta_n = k_0$ とおく場合.
2. $|S_0|$ に関する情報はなく, $\zeta_n = 1$ とおく場合.

以上のそれぞれの場合について, ℓ_0 回帰モデル選択の一致性について報告した. 本講演で紹介した理論的な結果は, ある意味で Honda et al.(2019)[6] の定理 2 の真のモデルサイズが発散する場合への拡張と言え, 証明手法はその論文の証明手法に新

しいアイデアを加えた上で、より精密化したものである。詳細は Honda(2023)[5] を参照されたい。

3. 計算法

[4] と [2] の手法を合わせた、近似法による加法モデルと変動係数モデルの場合の、提案した手法の計算法のアイデアについて報告した。

参考文献

- [1] D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. *Ann. Statist.*, 44:1–32, 2016.
- [2] L.-Y. Chen and S. Lee. Sparse quantile regression. *Journal of Econometrics*, 235:2195–2217, 2023.
- [3] T. Hastie, R. Tibshirani, and R. Tibshirani. Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, 34:579–592, 2020.
- [4] H. Hazimeh, R. Mazumder, and P. Radchenko. Grouped variable selection with discrete optimization: Computational and statistical perspectives. *Ann. Statist.*, 51:1–32, 2023.
- [5] T. Honda. Sparse quantile regression via ℓ_0 -penalty. *Discussion Paper Series No. 2023-03, Graduate School of Economics, Hitotsubashi University*, 2023.
- [6] T. Honda, C.-K. Ing, and W.-Y. Wu. Adaptively weighted group lasso for semiparametric quantile regression models. *Bernoulli*, 25:3311–3338, 2019.

対数周辺尤度とその陰的正則化のスケーリング挙動

九州大学情報基盤研究開発センター

徳田 悟

古くは 17 世紀に発見されたケプラーの法則が象徴するように、単純な関数や方程式を用いて観測データを表す数理モデリングは様々な現象に対する理解を深めてきた。現象の本質を特徴付けるためにはできるだけパラメータの少ない単純なモデルが望ましい。こうした指針はオッカムの剃刀として知られる。統計学の文脈では、赤池情報量規準 (AIC) やベイズ情報量規準 (BIC) に象徴されるモデル選択の規準がオッカムの剃刀をサポートしている。特に BIC はベイズ統計における対数周辺尤度のラプラス近似として導出される統計量であり [1]、対数周辺尤度が有する陰的正則化と一致性を示している。しかし、この近似では観測データの質 (ノイズ分散) が無視できる極限を考えており、対数周辺尤度やその陰的正則化がノイズ分散の大きさにどう依存するかは非自明な問題である。

特異学習理論は対数周辺尤度の漸近挙動の解明に貢献してきた [2]。BIC が正則なモデルに対する対数周辺尤度の漸近挙動を表す一方、特異学習理論では非正則なモデルも含めた一般化が成される。同理論では対数周辺尤度の有する陰的正則化が実対数閾値 (RLCT) と呼ばれる双有理不変量によって定量化される。しかし、特異学習理論もまたノイズ分散が無視できる極限を考えており、前述の問題に対する回答を提示しない。

対数周辺尤度やその陰的正則化がノイズ分散に依存することを示すスケーリング挙動を明らかにする上で、以下の関連問題が挙げられる。

- RLCT は陰的正則化の極限值であり、観測可能な統計量でない
- 特異学習理論ではノイズ分散 $\beta^{-1} = O(1)$ の下でサンプルサイズの極限 $n \rightarrow \infty$ に着目
- 対数周辺尤度は $\beta = O(1)$ の下で一致性を示すが、 $n\beta = O(1)$ での挙動は不明

これらの問題に対し、我々は次のように貢献した [3] (図 1)。

- ✓ RLCT の観測可能な対応物 (ベイズ比熱) を導入
- ✓ $n\beta = O(1)$ の下での極限 $n \rightarrow \infty$ に着目し、 $n\beta$ に依存するスケーリング関数を導出
- ✓ スケーリング関数を元に対数周辺尤度の $n\beta$ への依存性や陰的正則化の寄与を実証

特に、3 つ目の貢献を通じて、ベイズ推定が観測データの質や量に応じた複数の「状態」を取り、状態によって異なるモデルを最適とみなす性質を発見した。観測の精細度 $n\beta$ が高い (データの量が多いまたは質が高い) 状態であるほど、より多くのパラメータを持つ複雑なモデルを最適とみなす性質である。こうしたスケーリング挙動は BIC では説明がつかず、導出したスケーリング関数によって初めて明らかになったものである。

以上の知見は研究者の洞察に頼ってきたモデルの単純化をデータに根ざして客観化し、様々な現象の理解に役立つことが期待される。モデルの妥当性を実証するために「どの程度の量や質のデータが必要か？」に答えるものでもあり、計測の効率化にも示唆を与える。

所得不平等度の分解の推定

千葉大学 川久保 友超
名古屋市立大学 各務 和彦

1 一般化エントロピーとその分解可能性

x_i ($i = 1, \dots, N$) を、母集団における i 番目の世帯の所得とし、添字集合 $I = \{1, \dots, N\}$ によって有限母集団を表すとする。ここで、 I の所得不平等の尺度として、以下で定義される一般化エントロピー尺度 (generalized entropy measures, GE) を考える。

$$GE_\theta = \frac{1}{\theta(\theta-1)} \frac{1}{N} \sum_{i \in I} \left\{ \left(\frac{x_i}{\mu} \right)^\theta - 1 \right\}, \quad \theta \neq 0, 1. \quad (1)$$

ただし、 $\mu = N^{-1} \sum_{i=1}^N x_i$ は母集団 I における平均所得を表す。

次に、母集団 I (全国) を互いに排反な部分母集団 (地域) $I_1, \dots, I_J$ に分解し、 I_j の GE を考える。ここで、 $N_j = \#(I_j)$ を地域 I_j のサイズ (世帯数)、 $\lambda_j = N_j/N$ を I_j の世帯数シェア、 $\mu_j = N_j^{-1} \sum_{i \in I_j} x_i$ を I_j の平均所得とする。このとき、地域 I_j の GE は、

$$GE_{\theta,j} = \frac{1}{\theta(\theta-1)} \frac{1}{N_j} \sum_{i \in I_j} \left\{ \left(\frac{x_i}{\mu_j} \right)^\theta - 1 \right\}, \quad \theta \neq 0, 1.$$

と定義される。

GE には、以下の分解可能性という望ましい性質がある。

$$GE_\theta = \underbrace{\sum_{j=1}^J \lambda_j^{1-\theta} s_j^\theta GE_{\theta,j}}_{\text{地域内不平等}} + \underbrace{\frac{1}{\theta(\theta-1)} \left\{ \sum_{j=1}^J \lambda_j \left(\frac{\mu_j}{\mu} \right)^\theta - 1 \right\}}_{\text{地域間不平等}}.$$

ただし、 $s_j = \lambda_j \mu_j / \mu$ は I_j の所得シェアを表す。分解の第一項は地域内不平等 (within-region inequality) を表し、各地域の GE の加重平均と理解できる。一方で第二項は地域間不平等 (between-region inequality) を表し、「各地域の全世帯の所得が、その地域の平均所得 μ_j である場合の母集団の GE」と理解できる。

2 GE の分解の推定

2.1 グループデータにもとづいた GE の推定

公的統計における所得データの多くは、ある標本調査によって収集された所得の情報が、グループデータの形式で公開されている。 $S \subseteq I$ は、ある標本調査において標本として選ばれた世帯 i をあつめた添字集合とする。ここで、予め決められた境界値 $0 = c_0 < c_1 < \dots < c_G = +\infty$ に対して、それぞれの所得階級内の標本における世帯数

$$y_g = \sum_{i \in S} I(c_{g-1} \leq x_i < c_g), \quad (g = 1, \dots, G),$$

がカウントされ、度数分布 $\mathbf{y} = (y_1, \dots, y_G)^\top$ が公表されているとする。

グループ数 $G = 10$ 程度のグループデータ $\mathbf{y}$ にもとづいて GE を推定するために、パラメトリックな所得分布を仮定する。 $F(\cdot | \omega)$ および $f(\cdot | \omega)$ を、パラメータ ω を持つ確率分布の累積分布関数および確率密度関数とする。このとき、グループデータ $\mathbf{y}$ にもとづいた尤度は、

$$L(\omega) \propto \prod_{g=1}^G \{F(c_g | \omega) - F(c_{g-1} | \omega)\}^{y_g}.$$

という多項分布の尤度となる。密度関数 $f(\cdot | \omega)$ を持つ確率分布の GE は、(1) 式の有限母集団の GE の定義における有限母集団平均（算術平均）をとる操作を、確率分布の平均（期待値）におきかえた以下の式で定義される。

$$\text{GE}_\theta(\omega) = \frac{1}{\theta(\theta-1)} \left\{ \int_0^\infty \left(\frac{x}{\mu}\right)^\theta f(x | \omega) dx - 1 \right\},$$

所得分布として当てはまりの良いとされる多くのパラメトリックな確率分布は、 $\text{GE}_\theta(\omega)$ 、 $T_0(\omega)$ および $T_1(\omega)$ が ω の陽な関数として書ける。よって、尤度 $L(\omega)$ にもとづいて ω が推定できれば、GE も容易に推定できる。

2.2 GE の分解の推定

次に、全国の GE と各地域の GE を同時に推定する問題を考える。各地域の標本添字集合を $\mathcal{S}_j = \mathcal{S} \cap I_j$ と定義し、各地域の世帯所得の度数分布 $\mathbf{y}_{R,j} = (y_{j1}, \dots, y_{jG})^\top$ が観測されているとする。ただし、

$$y_{jg} = \sum_{i \in \mathcal{S}_j} I(c_{g-1} \leq x_i < c_g), \quad (g = 1, \dots, G).$$

である。 $\mathbf{y}$ と $\mathbf{y}_j$ ($j = 1, \dots, J$) が観測されているもとで、以下の推定手法を提案する。

まず、全国のパラメトリック分布 $f(\cdot | \omega)$ と事前分布 ω を仮定して全国の GE を推定し、その推定値を $\widehat{\text{GE}}_\theta$ とする。サンプルサイズとグループ数 G に応じて適切なパラメトリック分布を選択することで、全国の GE を効率的に推定することができる。次に、各地域 I_j のパラメトリック分布 $f_j(\cdot | \omega_j)$ と事前分布 $\pi_j(\omega_j)$ を仮定し、 I_j の平均所得推定値 $\hat{\mu}_j$ を求める。また、 $\hat{\mu} = \sum_{j=1}^J \lambda_j \hat{\mu}_j$ として全国の平均所得推定値 $\hat{\mu}$ を求める。なお、世帯数シェア λ_j は既知とする。 $\hat{\mu}_j$ と $\hat{\mu}$ にもとづいて、地域間不平等を

$$\hat{B} = \frac{1}{\theta(\theta-1)} \left\{ \sum_{j=1}^J \lambda_j \left(\frac{\hat{\mu}_j}{\hat{\mu}}\right)^\theta - 1 \right\}. \quad (2)$$

と推定する。最後に、分解の等式を制約として、各地域の GE を制約付きベイズ推定量によって推定する。制約付きベイズ推定量は、以下の事後リスクの制約付き最小化問題

$$\min_{d_1, \dots, d_J} \sum_{j=1}^J \phi_j E \left[\{d_j - \text{GE}_{\theta,j}(\omega_j)\}^2 | \mathbf{y}_R \right] \quad \text{s.t.} \quad \widehat{\text{GE}}_\theta = \sum_{j=1}^J w_j d_j + \hat{B}, \quad (3)$$

の解であり、

$$\widehat{\text{GE}}_{\theta,j}^{\text{CB}} = \widehat{\text{GE}}_{\theta,j}^{\text{B}} + \frac{r_j}{q} \left(\widehat{\text{GE}}_\theta - \hat{B} - \widehat{\text{GE}}_\theta^w \right), \quad j = 1, \dots, J,$$

と解ける。

3 日本の所得データを用いた実証分析

住宅・土地統計調査の所得データを用いて、全国の GE を都道府県間不平等と都道府県内不平等に分解し、さらに各都道府県の GE を市区町村間不平等と市区町村内不平等に分解する実証分析を行った。

Bayesian analysis of stochastic conditional duration models with intraday and intra-deferred future seasonalities in high-frequency commodity market

Makoto Nakakita¹, Tomoki Toyabe², and Teruo Nakatsuma²

¹AIP, RIKEN, Japan

²Faculty of Economics, Keio University, Japan

1 Introduction

We propose an extension of the stochastic conditional duration (SCD) model to capture intraday seasonality patterns in stock trading intervals and an improved Markov chain Monte Carlo (MCMC) method. The trading time intervals of financial data are known to have very slow-decaying autocorrelations, and the Ancillarity-Sufficiency Interweaving Strategy (ASIS) overcomes the disadvantage of instability in estimation by incorporating a centered parameterization of the model. The estimation results using gold futures tick data from JPX show that trading time intervals for each forward month of the contract month has a trend of a certain pattern and that the use of ASIS leads to stable estimation results.

2 Stochastic Conditional Duration Model

Considering the time of the i th transaction τ_i and the duration $y_i = \tau_{i+1} - \tau_i$ as a product of a latent valuable α_i and a positive random variable ε_i , the SCD model with intraday seasonality is written as:

$$y_i = \exp(x_i' \beta + \alpha_i) \varepsilon_i, \quad \varepsilon_i > 0, \quad i \in \{1, \dots, N\}, \quad (1)$$

$$\alpha_i = \phi \alpha_{i-1} + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma^2), \quad i \in \{2, \dots, N\},$$

$$\alpha_1 \sim \mathcal{N}(0, \sigma^2 / (1 - \phi^2)), \quad |\phi| < 1, \quad (2)$$

$$x_i = (x_i^B; x_i^L) \quad \beta = (\beta^B; \beta^L).$$

Note that the transaction time is normalized to $[0, 1]$ (i.e., the market starts at time 0 and ends at time 1) and both ε_i and η_i are mutually and serially independent in this model. As for the distribution of the random variable ε_i , the weibull distribution and the gamma distribution have been used in the literature. Following Toyabe and Nakatsuma (2022), we use $\text{Gamma}(\gamma, \gamma)$ instead of the commonly used

Gamma($\gamma, 1$) for computational stability. Thus the probability density function (pdf) of ε_i and y_i are

$$p(\varepsilon_i|\gamma) = \begin{cases} \gamma \varepsilon_i^{\gamma-1} \exp(-\varepsilon_i^\gamma), & \text{(Weibull distribution),} \\ \frac{\gamma^\gamma}{\Gamma(\gamma)} \varepsilon_i^{\gamma-1} \exp(-\gamma \varepsilon_i), & \text{(gamma distribution),} \end{cases} \quad (3)$$

and

$$p(y_i|\alpha_i, \beta, \gamma) = \begin{cases} \gamma y_i^{\gamma-1} \exp(-\gamma(x'_i\beta + \alpha_i) - y_i^\gamma \exp(-\gamma(x'_i\beta + \alpha_i))), & \text{(Weibull distribution),} \\ \frac{\gamma^\gamma}{\Gamma(\gamma)} y_i^{\gamma-1} \exp(-\gamma(x'_i\beta + \alpha_i + y_i \exp(-x'_i\beta - \alpha_i))), & \text{(gamma distribution).} \end{cases} \quad (4)$$

To capture the structure of the intraday seasonality of trading time intervals, this model considers that $x_i^{B'}\beta^B$ is a Bernstein polynomial

$$x_i^{B'}\beta^B = \sum_{k=0}^N \beta_k^B x_{k,i} = \sum_{k=0}^N \beta_k^B b_{k,m}(\tau_{i+1}) \quad (5)$$

$$b_{k,m}(v) = \binom{m}{k} v^k (1-v)^{m-k}, \quad k = 0, \dots, m, \quad v \in [0, 1]. \quad (6)$$

Note that the τ_{i+1} , not the time interval y_i , is in the equation here. This allows us to estimate the pattern of intraday seasonality with time as an argument.

By the same method as intraday seasonality, we added the order book information as

$$x_i^{L'} = (d_{i1}, d_{i2}, d_{i3}, d_{i4}, d_{i5}) \quad (7)$$

where d_{i1} is the price return, d_{i2} is the absolute value of price return, d_{i3} is the trade volume, $d_{i4} = d_{i1} * d_{i3}$ and $d_{i5} = d_{i2} * d_{i3}$.

3 Empirical results

Setting of hyperparameters is as follows.

$$(a_\phi, b_\phi) = (3, 2), \quad (a_\tau, b_\tau) = (3, 10), \quad (\gamma_a, \gamma_b) = (3, 10)$$

In the proposed model, it is not feasible to estimate the degree of the Bernstein polynomial using MCMC. We use the widely applicable information criterion (WAIC) proposed by Watanabe (2013) as a criterion for model selection. The detailed results are omitted here for reasons of space limitation.

4 Conclusions

From the curve depicting the pattern and seasonality of trading time intervals over a two-month period estimated from the proposed model, it can be observed that, consistently, trading time intervals exhibit an inverted U-shaped form. This represents an approach that captures the continuous behavior of the most actively traded futures, previously considered merely for the sake of removing strong autocorrelation. It aligns with the results of previous research Toyabe and Nakatsuma (2022), showing an inverted U-shaped intraday seasonality in stock trading.

微小ノイズをもつ空間 2 次元線形放物型確率偏微分方程式 モデルの係数パラメータの推定

大阪大学大学院基礎工学研究科 渡名喜庸蔵

1 モデルと設定

$D = (0, 1)^2$ 上の線形放物型確率偏微分方程式

$$dX_t = (\theta_2 \Delta + (\theta_1, \eta_1) \cdot \nabla + \theta_0) X_t dt + \epsilon dW_t, \quad t \in [0, 1]$$

を考える. ただし, 初期条件 $X_0 \in L^2(D)$, デイリクレ境界条件 $X_t(y, z) = 0, (t, y, z) \in [0, 1] \times \partial D$ を満たすとする. ここで $\{W_t\}_{t \geq 0}$ は D 上のソボレフ空間上の Q -ウィーナー過程とし, $(\theta_0, \theta_1, \eta_1, \theta_2) \in \mathbb{R}^3 \times (0, \infty)$ を未知パラメータ, パラメータ空間 $\Theta \subset \mathbb{R}^3 \times (0, \infty)$ はコンパクト凸集合で, $(\theta_0^*, \theta_1^*, \eta_1^*, \theta_2^*) \in \text{Int}(\Theta)$ を真値とする. また $\epsilon \in (0, 1]$ を既知の微小攪乱パラメータとする.

$-A_\theta = \theta_2 \Delta + (\theta_1, \eta_1) \cdot \nabla + \theta_0$ とおく. $l_1, l_2 \in \mathbb{N}, (y, z) \in D$ に対して

$$\lambda_{l_1, l_2} = \pi^2 \theta_2 (l_1^2 + l_2^2) + \frac{\theta_1^2 + \eta_1^2}{4\theta_2} - \theta_0, \quad e_{l_1, l_2}(y, z) = 2 \sin(\pi l_1 y) \sin(\pi l_2 z) e^{-\frac{\theta_1}{2\theta_2} y} e^{-\frac{\eta_1}{2\theta_2} z},$$

$u, v \in L^2(D)$ に対して $\langle u, v \rangle = \int_0^1 \int_0^1 u(y, z) v(y, z) e^{\frac{\theta_1}{\theta_2} y} e^{\frac{\eta_1}{\theta_2} z} dy dz$, $\|u\| = \sqrt{\langle u, u \rangle}$ とおく. $\alpha \in (0, 1)$ は既知で, $\mu_{l_1, l_2} = \pi^2 (l_1^2 + l_2^2) + \mu_0$ とする. ただし $\mu_0 \in (-2\pi^2, \infty)$ は未知パラメータでパラメータ空間は $(-2\pi^2, \infty)$ のコンパクト凸集合で, 真値 μ_0^* はその内部に属するとする. Q -ウィーナー過程を次で定める.

$$W_t = \sum_{l_1, l_2 \geq 1} \mu_{l_1, l_2}^{-\alpha/2} w_{l_1, l_2}(t) e_{l_1, l_2}, \quad t \geq 0.$$

ただし $\{w_{l_1, l_2}(\cdot)\}_{l_1, l_2 \in \mathbb{N}}$ は独立な 1 次元標準ウィーナー過程である.

観測データとして高頻度時空間データ $\mathbb{X}_{N, M} = \{X_{t_i}(y_j, z_k)\}_{0 \leq i \leq N, 0 \leq j \leq M_1, 0 \leq k \leq M_2}$ を考える. ただし $t_i = i\Delta = i/N$, $y_j = j/M_1$, $z_k = k/M_2$, $M = M_1 M_2$ とする. また, 初期値 $X_0 \in L^2(D)$ は $\langle X_0, e_{1,1} \rangle \neq 0$, $\|A_\theta X_0\| < \infty$ を満たし, $\lambda_{1,1}^* = 2\pi^2 \theta_2^* + \frac{(\theta_1^*)^2 + (\eta_1^*)^2}{4\theta_2^*} - \theta_0^* > 0$ とする.

2 $\theta_1, \eta_1, \theta_2$ の推定

$b \in (0, 1/2)$ を固定し, $b \leq \tilde{y}_0 < \tilde{y}_1 < \dots < \tilde{y}_{m_1} \leq 1 - b$, $b \leq \tilde{z}_0 < \tilde{z}_1 < \dots < \tilde{z}_{m_2} \leq 1 - b$ を満たす観測データ $\mathbb{X}_{N, M}$ の空間間引きデータ $\mathbb{X}_{N, m}^{(1)} = \{X_{t_i}(\tilde{y}_j, \tilde{z}_k)\}_{0 \leq i \leq N, 0 \leq j \leq m_1, 0 \leq k \leq m_2}$ を考える.

$\Delta_i X(y, z) = X_{t_i}(y, z) - X_{t_{i-1}}(y, z)$ に対して

$$Z_N(y, z) = \frac{1}{N\Delta^\alpha} \sum_{i=1}^N \Delta X_i^2(y, z), \quad f_\alpha(y, z; \theta_1, \eta_1, \theta_2) = \frac{\Gamma(1-\alpha)}{4\pi\alpha\theta_2^{1-\alpha}} e^{-\frac{\theta_1}{\theta_2} y} e^{-\frac{\eta_1}{\theta_2} z}$$

($\Gamma(\cdot)$ はガンマ関数) とおき, $\mathbb{X}_{N, m}^{(1)}$ を用いて $(\theta_1, \eta_1, \theta_2)$ の最小コントラスト推定量を次で定める.

$$(\hat{\theta}_1, \hat{\eta}_1, \hat{\theta}_2) = \underset{\theta_1, \eta_1, \theta_2}{\operatorname{arginf}} \sum_{k=1}^{m_2} \sum_{j=1}^{m_1} \left\{ \epsilon^{-2} Z_N(\tilde{y}_j, \tilde{z}_k) - f_\alpha(\tilde{y}_j, \tilde{z}_k; \theta_1, \eta_1, \theta_2) \right\}^2.$$

$\mathcal{R}_{N,\epsilon} = N^{1/2 \wedge (1-\alpha)} \wedge \epsilon N^{5/4-\alpha} \wedge \frac{\epsilon^2 N^{2-\alpha}}{\log N}$ とすると $N, m_1, m_2 \rightarrow \infty, \epsilon \rightarrow 0$ のとき正則条件の下,

$$\mathcal{R}_{N,\epsilon}(\hat{\theta}_1 - \theta_1^*, \hat{\eta}_1 - \eta_1^*, \hat{\theta}_2 - \theta_2^*) = \mathcal{O}_{\mathbf{P}}(1)$$

が成り立つ.

3 θ_0 の推定

$n \leq N$ とする. $\tilde{t}_i = \lfloor \frac{N}{n} \rfloor \frac{i}{N}$, $i = 0, \dots, n$ に対して観測データ $\mathbb{X}_{N,M}$ の時間間引きデータ $\mathbb{X}_{n,M}^{(2)} = \{X_{\tilde{t}_i}(y_j, z_k)\}_{0 \leq i \leq n, 0 \leq j \leq M_1, 0 \leq k \leq M_2}$ を考える.

$x_{1,1}(t) := \langle X_t, e_{1,1} \rangle$ は以下を満たす 1 次元オルンシュタイン・ウーレンベック過程となる.

$$dx_{1,1}(t) = -\lambda_{1,1}x_{1,1}(t)dt + \epsilon\mu_{1,1}^{-\alpha/2}dw_{1,1}(t), \quad x_{1,1}(0) = \langle X_0, e_{1,1} \rangle.$$

$\mathbb{X}_{n,M}^{(2)}$ と上で構成した推定量 $\hat{\theta}_1, \hat{\eta}_1, \hat{\theta}_2$ を用いた $x_{1,1}(\cdot)$ の近似過程

$$\hat{x}_{l_1, l_2}(\tilde{t}_i) = \frac{2}{M} \sum_{k=1}^{M_2} \sum_{j=1}^{M_1} X_{\tilde{t}_i}(y_j, z_k) \sin(\pi l_1 y_j) \sin(\pi l_2 z_k) e^{\frac{\hat{\theta}_1}{2\hat{\theta}_2} y_j} e^{\frac{\hat{\eta}_1}{2\hat{\theta}_2} z_k}$$

による $(\lambda_{1,1}, \mu_{1,1})$ の最小コントラスト推定量

$$(\hat{\lambda}_{1,1}, \hat{\mu}_{1,1}) = \operatorname{arginf}_{\lambda, \mu} \left\{ \sum_{i=1}^n \frac{(\hat{x}_{1,1}(\tilde{t}_i) - e^{-\lambda/n} \hat{x}_{1,1}(\tilde{t}_{i-1}))^2}{\frac{\epsilon^2(1-e^{-2\lambda/n})}{2\mu^\alpha \lambda}} + n \log \frac{1 - e^{-2\lambda/n}}{2\mu^\alpha \lambda/n} \right\}.$$

を考え, θ_0, μ_0 の推定量を

$$\hat{\theta}_0 = -\hat{\lambda}_{1,1} + \frac{\hat{\theta}_1^2 + \hat{\eta}_1^2}{4\hat{\theta}_2} + 2\pi^2 \hat{\theta}_2, \quad \hat{\mu}_0 = \hat{\mu}_{1,1} - 2\pi^2$$

で定める. $G_1(\lambda, \mu) = \frac{1-e^{-2\lambda}}{2\lambda} \mu^\alpha x_{1,1}(0)^2$, $G_2(\mu) = \frac{\alpha^2}{2\mu^2}$ とすると $n \rightarrow \infty, \epsilon \rightarrow 0$ のとき正則条件の下,

$$(\epsilon^{-1}(\hat{\theta}_0 - \theta_0^*), \sqrt{n}(\hat{\mu}_0 - \mu_0^*)) \xrightarrow{d} \mathcal{N}(0, \operatorname{diag}\{G_1(\lambda_{1,1}^*, \mu_{1,1}^*), G_2(\mu_{1,1}^*)\}^{-1})$$

が成り立つ.

4 数値シミュレーション

高頻度時空間データ $\mathbb{X}_{N,M}$, $N = 10^3$, $M_1 = M_2 = 200$ による大規模数値シミュレーションを実行し, 推定量 $\hat{\theta}_1, \hat{\eta}_1, \hat{\theta}_2, \hat{\theta}_0, \hat{\mu}_0$ の漸近的挙動を検証する.

参考文献

- M. Bibinger and M. Trabs. (2020). Volatility estimation for stochastic PDEs using high-frequency observations. *Stochastic Processes and their Applications*, 130(5), 3005–3052.
- Y. Kaino and M. Uchida. (2021). Adaptive estimator for a parabolic linear SPDE with a small noise. *Japanese Journal of Statistics and Data Science*, 4, 513–541.
- Y. Tonaki, Y. Kaino, and M. Uchida. (2023). Parameter estimation for linear parabolic SPDEs in two space dimensions based on high frequency data. *Scandinavian Journal of Statistics*, 50(4), 1568–1589.
- Y. Tonaki, Y. Kaino, and M. Uchida. (2023). Parameter estimation for a linear parabolic SPDE in two space dimensions with a small noise. *Statistical Inference for Stochastic Processes*. <https://doi.org/10.1007/s11203-023-09301-2>.
- M. Sørensen and M. Uchida. (2003). Small-diffusion asymptotics for discretely sampled stochastic differential equations. *Bernoulli*, 9(6), 1051–1069.

VAR models for surface time series

fPCA analysis of nighttime population in Tokyo
during Jan. 2016-Dec. 2023

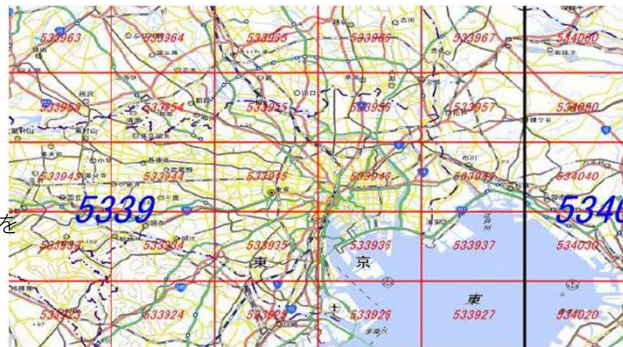
Yasumasa Matsuda and Daisuke Kurisu
Tohoku University and University of Tokyo

1

Data(東京圏夜間人口の分析)

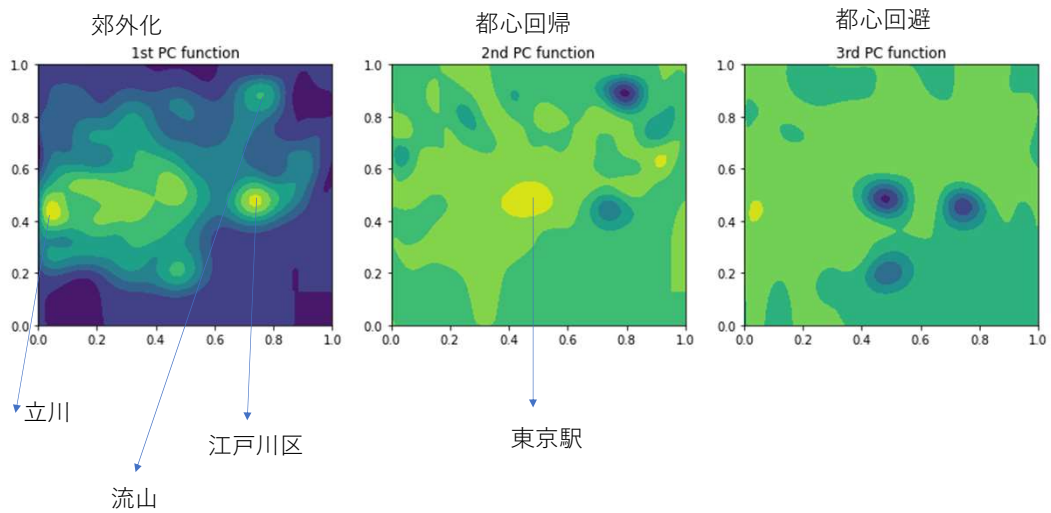
- 東京圏2400メッシュ(1kmx1km)における午前4時の人口
- 2016/1-2018/12の月平均を収集
- 96か月 x 2400メッシュの時空間データ

右図の1メッシュを
10x10に分けたメッシュを
用いる



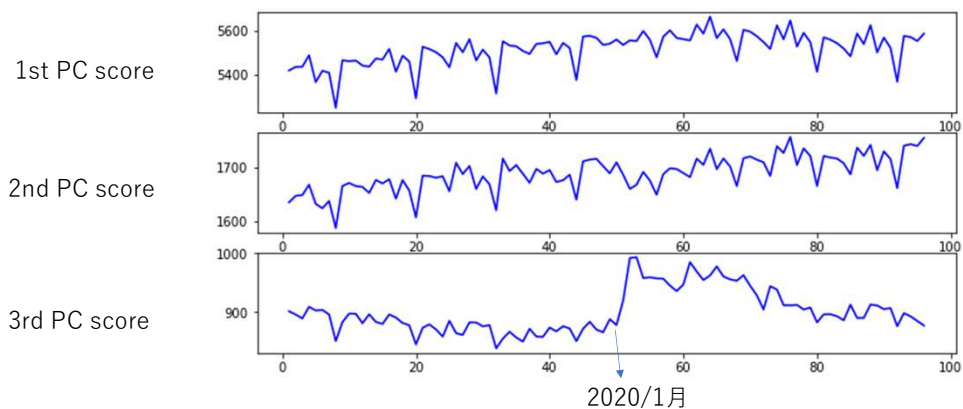
2

PC function for 対数差分系列



3

主成分スコア for 原系列



コロナによる影響がみられるのは、第3主成分の都心回避成分。現在(2023/12)では元の水準に回復
郊外化、都心回帰成分はコロナの前後でもほぼ変わりなく、都心回帰成分が増加傾向にある

4