2023 年度科学研究費シンポジウム

データサイエンスにおける
統計的理論・方法論の新展開

- 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：2023 年 11 月 13 日（月）～ 11 月 14 日（火）
- 会場：九州大学 伊都キャンパス ウエスト 1 号館 IMI オーディトリアムと Zoom のハイブリッド
- 開催責任者：川野 秀一（九州大学）
- 内容・目的：計測技術の高度な発展に加え AI 技術の爆発的な進展により、現在さまざまな分野において大規模かつ多種多様なデータが取得されています。これらのデータから有益な知識や情報を抽出するためには、データサイエンス分野、特に統計科学分野の理論および方法論の発展が必要不可欠です。本シンポジウムでは、統計科学や機械学習における理論や方法論に関する研究をはじめ、さまざまな分野におけるデータサイエンスの事例研究や実応用研究など、幅広いテーマを扱います。最新の研究動向や各分野における問題点を共有することにより、データサイエンス研究の更なる発展を目的としています。

講演プログラム

11 月 13 日（月）

10:00 ~ 10:05　オープニング

10:05 ~ 10:35 対面　武冨 奈菜美（久留米大学），張 元宗（目白大学），今野 良彦（大阪公立大学），森 美穂子（久留米大学），江村 剛志（統計数理研究所）

メタ分析における正規母平均の Pretest 推定量に基づく Pivot 信頼区間

10:35 ~ 11:05 zoom　張 南（広島修道大学）

Measuring Global Flow of Funds and Financial Network

11:05 ~ 11:35 対面　大石 峰暉（東北大学）

線形回帰モデルのカテゴリ変数に対するカテゴリ選択と変数選択

11:35 ~ 13:00　昼食

13:00 ~ 13:25 対面　堀川 慧斗（広島大学），永井 勇（中京大学），門田 麗（広島大学），栁原 宏和（広島大学）

GMANOVA モデルの 3 相データへの拡張

13:25 ~ 13:50 対面　柴山 佐内（広島大学），栁原 宏和（広島大学）

線形予測値で表される多変量回帰モデルにおける修正 Cp 規準

14:50~ 14:15 対面　岩重 文也（広島大学），橋本 真太郎（広島大学）

有限混合モデルの混合事前分布を用いた確率的ブロックモデルのベイズ推測

14:15 ~ 14:40 対面　桃木 光輝（鹿児島大学），吉田 拓真（鹿児島大学）

Asymptotic properties for small area estimation in extreme value theory

14:40 ~ 15:00　休憩

15:00 ~ 15:30 対面　得丸 久文

情報理論におけるエントロピー概念の誤り訂正　S/N(信号対雑音)比とエントロピーを考える

15:30 ~ 16:00 対面　島谷 健一郎（統計数理研究所）

動物の移動軌跡モデルと presence and pseudo-absence data

16:00 ~ 16:30 zoom　深澤 圭太（国立環境研究所）

空間標識再捕獲法による個体密度と景観連結性の推定

16:30 ~ 17:00 zoom　加納 寛子（山形大学）

シミュレーションによるオンライン教育に対する効果分析

17:00 ~ 17:20　休憩

17:20 ~ 17:45 対面　桝田 陸（東京大学），入江 薫（東京大学）
逐次的ベイズ予測統合
17:45 ~ 18:10 対面　外城 稔雄（星光 PMC），鈴木 譲（大阪大学）
LiNGAM-MMI を用いた料理レシピの因果探索
18:10 ~ 18:35 対面　田坂 理英子（大阪大学），新村 亮介（大阪大学），鈴木 譲（大阪大学）
Fused Lasso に対する PSI を用いたモデル選択の新提案

11 月 14 日（火）
10:00 ~ 10:25 対面　松本 美幸（千葉大学），内藤 貫太（千葉大学）
Further analysis on LASSO-type estimator in nonparametric regression
10:25 ~ 10:50 対面　小杉 拓生（千葉大学），内藤 貫太（千葉大学）
Semiparametric regression with localized Bregman divergence
10:50 ~ 11:15 対面　浦崎 航（東京理科大学），中川 智之（明星大学），田畑 耕治（東京理科大学）
二元分割表における divergence 型尺度を用いた対称性の可視化
11:15~ 11:40 対面　田中 俊太郎（滋賀大学，日本総合研究所），松井 秀俊（滋賀大学）
複雑な構造を持つ高次元データに対する変数スクリーニング

11:40 ~ 13:00　昼食

13:00 ~ 13:30 zoom　貝野 友祐（神戸大学）
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推定
13:30 ~ 14:00 対面　入江 薫（東京大学），羽村 靖之（京都大学），菅澤 翔之助（慶應義塾大学）
外れ値と過剰なゼロを含む計数データの事後ロバスト分析
14:00 ~ 14:30 対面　柳本 武美（統計数理研究所）
Jeffreys prior と MLE に共通した役割とその含意
14:30 ~ 15:00 対面　作村 建紀（法政大学），柳本 武美（統計数理研究所）
共役ガンマ事前分布を用いたワイブル分布のベイズ推定

15:00 ~ 15:20　休憩

15:20 ~ 15:50 対面　増田 弘毅（東京大学），江口 翔一（大阪工業大学）

Robustifying Gaussian quasi-likelihood inference for volatility

15:50 ~ 16:20 対面　小池 祐太（東京大学）

Estimation of the number of relevant factors from high-frequency data

16:20 ~ 16:50 対面　内藤 貫太（千葉大学）

Simultaneous Confidence Region of an Embedded One-Dimensional Curve in Multi-Dimensional Space

16:50 ~ 16:55　クロージング

メタ分析における正規母平均の Pretest 推定量に基づく Pivot 信頼区間

武冨　奈菜美[1] ・渡辺　元宗（張　元宗）[2] ・今野 良彦[3] ・森　美穂子[4] ・江村　剛志[5]

1　はじめに

メタ分析は公表された複数の研究結果を統合する統計的手法の一つである．メタ分析では各研究が共通平均をもつと仮定し，推定する場合が多いが，平均に順序がある場合など共通平均が存在しないシナリオでは，共通平均が情報を持たない場合がある．近年，Pretest 推定量（Judge and Bock (1978)）を用いて，メタ分析で与えられた個々の研究の推定値を平均二乗誤差の観点から改良した推定法が開発された（Taketomi *et al.* (2021)）．Pretest 推定量に基づく方法は，与えられた研究の平均ベクトルがスパースになっているシナリオにおいて，得られている推定値を改良する．これまで Pretest 推定量の決定論的性質やバイアス，平均二乗誤差について調べられているが，推定量の累積分布関数や信頼区間に関する議論は見当たらない．本研究では，Pretest 推定量の標本分布を解析的に導出し，Pretest 推定量に基づく信頼区間を構築し，信頼区間の被覆確率を検証する．なお，本研究の詳細は Taketomi *et al.*(2023) で報告されている．

2　Pretest 推定量に基づく pivot 信頼区間

正規母集団から得られた G 個の独立な標本（Study）を考える．各 Study から平均 μ_i の正規分布に従うアウトカム Y_i とその標準偏差 σ_i が得られている $(i = 1, \cdots, G)$ というメタ分析の設定 (Fleiss(1993)) を考える．i 番目の study のアウトカムを Y_i で表し，$Y_i \sim N\left(\mu_i, \sigma_i^2\right)$ とする．ただし，$\mu_i \in (-\infty, \infty)$ は未知，$\sigma_i > 0$ は既知とする．このとき，μ_i の Pretest 推定量を次式で定義する．

$$\delta_i^{\mathrm{PT}} = Y_i I\left(\left|\frac{Y_i}{\sigma_i}\right| > z_{\alpha/2}\right).$$

ただし，$I(\cdot)$ は指標関数, z_α は標準正規分布 $N(0,1)$ の上側 $100(1-\alpha)\,\%$ 点, $0 < \alpha < 1$ である．δ_i^{PT} の分布関数の pivot（Casella and Berger (2002)）により，μ_i の pivot 信頼区間を次で定義する．

定義 2.1 (Pivot 信頼区間)．δ_i^{PT} に基づく μ_i の $100(1-\gamma)\,\%$ pivot 信頼区間を次で定義する．

$$\left[L_i^{\mathrm{PT}}, U_i^{\mathrm{PT}}\right] = \begin{cases} \left[\delta_i^{\mathrm{PT}} - z_{\gamma_2}\sigma_i, \delta_i^{\mathrm{PT}} + z_{\gamma_1}\sigma_i\right] & \text{if} \quad \delta_i^{\mathrm{PT}} < -z_{\alpha/2}\sigma_i, \\ \left[-\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma_i, \left(z_{\alpha/2} + z_{\gamma_1}\right)\sigma_i\right] & \text{if} \quad \delta_i^{\mathrm{PT}} = 0, \\ \left[\delta_i^{\mathrm{PT}} - z_{\gamma_2}\sigma_i, \delta_i^{\mathrm{PT}} + z_{\gamma_1}\sigma_i\right] & \text{if} \quad \delta_i^{\mathrm{PT}} > z_{\alpha/2}\sigma_i. \end{cases}$$

ただし，$\gamma_1 > 0$ と $\gamma_2 > 0$ は $\gamma_1 + \gamma_2 = 1$ を満たす定数である．特に，$\gamma_1 = \gamma_2 = \gamma/2$ のとき，

[1]久留米大学 大学院医学研究科, e-mail: nnmamikrn@gmail.com
[2]目白大学 社会学部, e-mail: chogenso@gmail.com
[3]大阪公立大学 大学院理学研究科, e-mail: konno@omu.ac.jp
[4]久留米大学 大学院医学研究科, e-mail: mkano@kurume-u.ac.jp
[5]統計数理研究所, 久留米大学 バイオ統計センター. e-mail: takeshiemura@gmail.com

$100\,(1-\gamma)\,\%$ pivot 信頼区間を次式で定義する．

$$\left[L_i^{\mathrm{PT}}, U_i^{\mathrm{PT}}\right] = \begin{cases} \left[\delta_i^{\mathrm{PT}} - z_{\gamma/2}\sigma_i, \delta_i^{\mathrm{PT}} + z_{\gamma/2}\sigma_i\right] & \text{if} \quad \delta_i^{\mathrm{PT}} < -z_{\alpha/2}\sigma_i, \\ \left[-\left(z_{\alpha/2} + z_{\gamma/2}\right)\sigma_i, \left(z_{\alpha/2} + z_{\gamma/2}\right)\sigma_i\right] & \text{if} \quad \delta_i^{\mathrm{PT}} = 0, \\ \left[\delta_i^{\mathrm{PT}} - z_{\gamma/2}\sigma_i, \delta_i^{\mathrm{PT}} + z_{\gamma/2}\sigma_i\right] & \text{if} \quad \delta_i^{\mathrm{PT}} > z_{\alpha/2}\sigma_i. \end{cases}$$

定義 2.1 の pivot 信頼区間は，任意の μ_i に対する被覆確率の下限が名目値 $1-\gamma$ という意味で妥当な信頼区間といえる．

3　実データへの適用

R パッケージ：meta.shrinkage（Taketomi *et al.* (2022)）を用いて，提案法を久留米大学医学部の解剖学実習でのアレルギー反応のデータセットに対して適用した．このデータは解剖学実習中にホルムアルデヒドに曝露された医学部生を対象とした，5 年間のデータである $(i = 1, \cdots, 5)$．$\gamma_1 = \gamma_2 = 0.025$ とし，男女間で目のアレルギー症状に差があるかを目の症状スコアの男女差（MD）とその標準誤差（SE）を用いて各年の MD の δ_i^{PT} と pivot 信頼区間を算出したところ，2015, 2016, 2018, 2022 年のデータでは $\delta_i^{\mathrm{PT}} = 0$ であったが，2019 年のデータでは $\delta_4^{\mathrm{PT}} = Y_4 = -0.75$ となり，有意水準 5%で男女差の目の症状スコアに有意差が見られた．

参考文献

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. 2nd edition. Duxbury Press.

Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, **2**(2), 121–145. https://doi.org/10.1177/096228029300200202.

Judge, G. G. and Bock, M. E. (1978). *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics*. North Holland, New York.

Taketomi, N., Konno, Y., Chang, Y. T. and Emura, T. (2021). A meta-analysis for simultaneously estimating individual means with shrinkage, isotonic regression and pretests. *Axioms*, **10**, 267. https://doi.org/10.3390/axioms10040267.

Taketomi, N., Michimae, H. Chang, Y. T. and Emura, T. (2022). meta.shrinkage: An R package for meta-analyses for simultaneously estimating individual means. *Algorithms*, **15**, 26. https://doi.org/10.3390/a15010026.

Taketomi, N., Chang, Y. T., Konno, Y., Mori, M. and Emura, T. (2023). Confidence interval for normal means in meta-analysis based on a pretest estimator. *Japanese Journal of Statistics and Data Science*, in press.

Measuring Global Flow of Funds and Financial Network: Shock Dynamics and Propagation

Nan Zhang
Hiroshima Shudo University

This study seeks to construct the Global Flow of Funds (GFF) matrix model to measure global financial stability. We use GFF data and the sectoral account data establish the sectoral from-whom-to-whom financial stock matrix (SFSM). The SFSM focuses on counterparty national and cross-border exposures of the sectors between China and the United States to construct country-specific financial networks and connect each country-level network based on cross-border exposures. We use financial network analysis to run an empirical analysis of SFSM to study the local propagation dynamics of quantity shocks in investment and financing. To that aim, we propose a decomposition of shocks into n-order effects on the basis of an "inverse of Leontief" representation of the w-t-w matrices. And we further propose an eigenvector decomposition of the effects to provide an analytical description of the propagation process. This reveals the deep connection between the propagation role of financial instruments/countries and their centrality in the w-t-w network.

This study enhances the above research studies by improving upon the GFF statistical framework, integrating data sources, improving the compilation methods, and especially identifying the interconnections between the sectors' national tables based on the W-t-W model. This enhancement is achieved by combining the GFFM and the financial inflow–outflow table based on the sector data.

China (CN), Japan (JP), the United States (US), and the United Kingdom (UK) (collectively, the G-4) are the four largest economies in the world. Although their economic systems, market maturities, and political systems are different and face many political challenges, a GFF analysis can grasp the basic structure, mutual dependence, financial exposure risk, and homogeneity and heterogeneity of the external flow of funds of these four economies. This will undoubtedly be useful in understanding global financial stability. Therefore, on the theoretical basis of improving upon GFF statistics and developing application methods, this study also focuses on the setting of counterparty country sectors in CN, JP, the US, and the UK. This study explores new theoretical methods and proposes practical countermeasures to prevent financial crises.

This paper is arranged as follows. Section 2 improves upon the GFF statistical framework and reduces statistical discrepancies, discusses the integration and consistency of data sources. On this basis, section 3 discusses the methodology for preparing counterparty country sector tables, create a theoretical analysis model. Section 4 estimate bilateral exposures by international SFSM to show shock dynamics in each sector of G-4, analytical representation of shock propagation in sectors, including the financial network, to show the financial position and cross-border exposures of the countries in GFF and shock dynamics between the Sectors. The contributions of this paper have the following three main points.

First, this is the first paper to compare national financial exposures across sectors of G-4 economies by the GFF analysis framework. We use FA, CDIS, CPIS, and BIS's data to estimate bilateral financial exposures between G-4 economies. Furthermore, it connects national financial networks to each other via cross-border exposures (see Figure 1), which are calculated by merging information from the FA, CDIS, CPIS, and BIS datasets.

Second, it based on the framework of W-t-W to illustrate the use of a Leontief representation for W-t-W financial matrices. According to the relationship and certain regularities in the w-t-w algebraic structure, the dynamics are decomposed into contributions by the eigenvectors of the diffusion matrix, the matrix containing the ratios of financing by counterpart sector per unit of investment by sector

measured in stock terms. The dynamics of the direct and indirect effects emerge as linear combinations of declining trajectories associated with the eigenvalues of the matrix, weighted on the components of the shocks in the base of eigenvectors. This paper focuses on cross-border interconnections of portfolios investment to measure the external shocks dynamics and the propagation of the shock between China and the US. We have illustrated on the basis of China, Japan, and the US, and the UK example, the use of a Leontief representation for w-t-w financial matrices. In particular to study the propagation dynamics of quantity shocks in investment and financing by making use of a Leontief power series expansion.

Third, as an empirical analysis, we can find an interesting fact which was by calculated. That is, although we assumed that the external investment of the United States is reduced by 1 unit, and China also reduced external financing by one unit, while other sectors of G-4 remain unchanged. However, with the large share in the international portfolio market and the close relationship with multi-country in portfolio trading, the U.S. has strong resilient to shocks, and the observed cumulative limit effect value is still the highest. Under this assumption, we do not see the effect of increased financial investment between China, Japan, and the UK, even if the scale of external portfolio investment in the United States declines. Another interesting result is that when we transpose the C matrix and observe the impact on the United States, Japan, and China from the financing (liability) side of the securities under the same assumed conditions, we get a completely different result. That is, once the w-t-w structure is considered, as done via the inverse of Leontief, the effects of the shock are more complicated than just a reduction and an increase in the investment by countries offsetting each other.

Regarding our future work, we need to focus on the following issues:

- The proposed framework can be used to decompose effects caused by quantity shocks of any nature. Such as central bank quantitative easing affecting the volume of assets held by the relevant sectors, shock affecting the distribution of stock value, and price shocks.
- Future research should tackle the lack of temporal dimension in the description of the propagation process.
- Beyond the use to characterise propagation effects, the Leontief representation can be used to extract other relevant information from the w-t-w data.
- Improve Financial Accounts and GFF statistics, and develop application methods, including network theory and data science

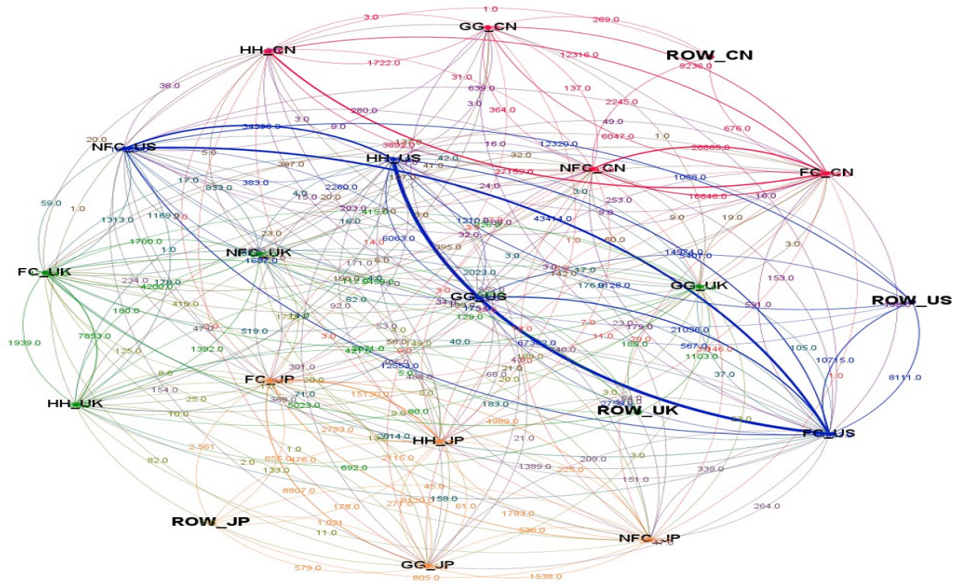


Figure 1 Assets and liabilities network in the sectors of G-4 at the end-2021

線形回帰モデルのカテゴリ変数に対するカテゴリ選択と変数選択

大石 峰暉
東北大学 データ駆動科学・AI 教育研究センター

m 個のカテゴリ変数を説明変数に持つ線形回帰モデルを考える. $\boldsymbol{y}$ を n 次元目的変数ベクトル, $\boldsymbol{X}_j$ ($j \in \{1,\ldots,m\}$) を p_j カテゴリを持つ第 j 説明変数を表す $n \times p_j$ カテゴリ変数行列とする. つまり, $\boldsymbol{X}_j$ の各成分は 1 か 0 で, $\boldsymbol{X}_j \mathbf{1}_{p_j} = \mathbf{1}_n$ を満たす. ここで $\mathbf{1}_n$ は全ての成分が 1 である n 次元ベクトルである. このとき, 以下の線形回帰モデルを考える.

$$\boldsymbol{y} = \sum_{j=1}^{m} \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}. \tag{1}$$

ここで, $\boldsymbol{\beta}_j = (\beta_{j1},\ldots,\beta_{jp_j})'$ は第 j 説明変数に対する p_j 次元回帰係数ベクトル, $\boldsymbol{\varepsilon}$ は $\mathrm{E}[\boldsymbol{\varepsilon}] = \mathbf{0}_n$, $\mathrm{Cov}[\boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{I}_n$ である n 次元誤差変数ベクトルであり, $\mathbf{0}_n$ はすべての成分が 0 である n 次元ベクトルである. 今回扱うカテゴリ変数としては, 2 種類の変数を想定する. 1 つは個体の属性などを表す質的変数, もう 1 つは量的変数を分割することにより質的変数に変換された変数である. いずれにおいてもカテゴリ数はモデルのパラメータ数に対応するため, カテゴリが細かすぎると予測精度の低下, 粗すぎるとモデルの当てはまりの低下という問題が生じる. よって, カテゴリの最適化は重要な問題である.

大石・柳原 (2023) は上述のカテゴリ選択に加え, 説明変数の変数選択を行うための手法として, 以下で説明する方法 I (Ohishi *et al.*, 2021) と方法 II (大石ら, 2021) を数値的に比較した. 方法 I は一般化 fused Lasso を用いてカテゴリ選択をする手法で, 以下の罰則付き残差平方和の最小化に基づいて推定する.

$$\sum_{j=1}^{m} \|\boldsymbol{y} - \boldsymbol{X}_j \boldsymbol{\beta}_j\|^2 + \lambda \sum_{j=1}^{m} \sum_{k=1}^{p_j} \sum_{\ell \in D_{jk}} w_{j,k\ell} |\beta_{jk} - \beta_{j\ell}|.$$

ここで, λ は非負のチューニングパラメータ, $D_{jk} \subseteq \{1,\ldots,p_j\} \backslash \{k\}$ は第 j 説明変数の第 k カテゴリの隣接関係を表す添え字集合, $w_{j,k\ell} > 0$ は第 j 説明変数の第 k, ℓ カテゴリ間の重みである. また, $\ell \in D_{jk} \Leftrightarrow k \in D_{j\ell}$, $w_{j,k\ell} = w_{j,\ell k}$ を満たす. これは, まずカテゴリを細かく分割し, それらの隣接関係に基づいて等しいカテゴリを結合するというものである. 方法 I では明示的な変数選択は行われないが, すべてのカテゴリが結合されればその変数は意味を持たないので, この意味で変数選択が可能である. 方法 II は明示的な変数選択を行うために方法 I にグループ Lasso を加えた手法で, 以下の罰則付き残差平方和の最小化に基づいて推定する.

$$\sum_{j=1}^{m} \|\boldsymbol{y} - \boldsymbol{X}_j \boldsymbol{\beta}_j\|^2 + \lambda_1 \sum_{j=1}^{m} \sum_{k=1}^{p_j} \sum_{\ell \in D_{jk}} w_{j,k\ell} |\beta_{jk} - \beta_{j\ell}| + \lambda_2 \sum_{j=1}^{m} w_j \|\boldsymbol{\beta}_j\|.$$

ここで, λ_1, λ_2 は非負のチューニングパラメータ, $w_j > 0$ は第 j 説明変数に対する重みである. 大石・柳原 (2023) は数値比較により, グループ Lasso が変数選択の性能と当てはめ値の MSE を向上させるが, 一方でカテゴリ選択の性能を低下させてしまうという結果を得た. よって, 明示的な変数選択は必要であるが, グループ Lasso による縮小推定がパラメータの過剰な縮小を引き起こし, カテゴリ選択に悪い影響を与えているのではないかと考えられる. そこで本研究では, 縮小推定をしない変数選択法として Kick-One-Out 法 (KOO 法: Zhao *et al.*, 1986; Bai *et al.*, 2018) に着目する.

c_0 をすべての説明変数を用いたモデルの下でのモデル選択規準の値, c_j を第 j 説明変数を抜いたモデルの下でのモデル選択規準の値とする. このとき, KOO 法で選ばれる説明変数は以下のように定義される.

$$\{j \in \{1,\ldots,m\} \mid c_j > c_0\}. \tag{2}$$

提案法 I として, 方法 I と KOO 法を組み合わせた以下の手法を提案する.

Step 1. すべての説明変数を用いて方法 I を実行し, c_0 を定義する.

Step 2. 第 j 説明変数を除いた $(m-1)$ 個の説明変数を用いて方法 I を実行し, c_j を定義する. この操作を $j = 1, \ldots, m$ に対して行う.

Step 3. Step 1, Step 2 で定義した $c_0, c_1, \ldots, c_m$ を用い, (2) 式に基づいて変数選択を行う.

Step 4. Step 3 で選ばれた説明変数を用いて方法 I を実行する.

提案法 I では縮小推定をすることなく変数選択を行うため, 方法 I のカテゴリ選択の性能が維持されることが期待できる. 一方で, 方法 I を $(m+2)$ 回繰り返す必要がある. 方法 I はチューニングパラメータの選択を含む繰り返し計算が必要であるため, これを繰り返すことで計算コストが高くなってしまう. そこで提案法 I を改良した以下の提案法 II を検討する.

Step 1. すべての説明変数を用いて方法 I を実行する.

Step 2. Step 1 で得られたカテゴリ選択の結果でカテゴリを固定し, モデル (1) を書き換える.

Step 3. Step 2 で新たに得られたモデルに対して通常の KOO 法を実行し, 変数選択を行う.

提案法 II では方法 I のカテゴリ選択の性能が維持され, さらに方法 I の実行は 1 回のみであるため, 高速な計算が可能となる.

本発表では, 上述の 2 つの提案法の性能を数値的に評価した結果を報告した. まずカテゴリ選択の性能について, 提案法 I が方法 II を上回るという期待通りの結果が得られた. しかしながら方法 I には及ばず, 明示的な変数選択が方法 I のカテゴリ選択の性能を低下させるという方法 II と同様の結果となった. 変数選択の性能は提案法 I が優れているという結果が得られた. 一方で提案法 II の変数選択の性能は方法 I と同等の結果となり, カテゴリ選択後の KOO 法が機能していないという結果となった. それでも提案法 II は当てはめ値の MSE が最も良く, 計算時間も方法 I とほどんど変わらないため, 最終的には提案法 II が最も良いという結果となった.

引用文献

[1] 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.

[2] Ohishi, M., Okamura, K., Itoh, Y. & Yanagihara, H. (2021). Optimizations for categorizations of explanatory variables in linear regression via generalized fused Lasso. I. Czarnowski, R. J. Howlett & L. C. Jain, eds, Intelligent Decision Technologies. Springer Singapore. Singapore. 457–467.

[3] 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.

[4] 大石峰暉・岡村健介・伊藤嘉道・栁原宏和 (2021). Generalized fused Lasso による説明変数のカテゴリの最適化. 2021 年度統計関連学会連合大会講演報告集, 206.

[5] 大石峰暉・栁原宏和 (2023). 線形回帰モデルにおけるカテゴリ変数の選択. 2023 年度統計関連学会連合大会講演報告集, 77.

GMANOVA モデルの 3 相データへの拡張 (報告書)

[1] 堀川 慧斗, [2] 永井 勇, [1] 門田 麗, [1] 柳原 宏和
[1] 広島大学大学院 先進理工系科学研究科, [2] 中京大学 教養教育研究院

本講演では [3] で提案された一般化多変量分散分析 (GMANOVA) モデルを拡張した Three-mode GMANOVA モデルを提案し, そのモデルにおける未知の行列の推定法や推定値を得るためのアルゴリズムを構築した.

従来の GMANOVA モデルは, 一つの項目に対して収集された経時測定データの分析にも適したモデルであり, 分析に用いた経時測定データに潜む時間による変化 (経時変化) を捉えることができる. ここで, 経時測定データは各個体に対して複数時点で計測することで得られるデータである. 例えば, 各学生の毎年の身長を測定して得られるデータが経時測定データである. 一方で, 一つの項目に対する経時測定データではなく, 複数の項目に対する経時測定データを収集する場合もある. 講演で例示したように, 各学生の毎年の身長だけでなく体重なども測定した場合は, 複数の項目に対する経時測定データが得られる. このように複数の項目に対して収集した経時測定データは, 個体・時点・項目の 3 次元構造を持つデータとみなすことができ, このような構造を持つデータを本講演では 3 相データ (three-mode data) と呼んだ. このような 3 相データの分析法として代表的なものは, データの要約を行う三相主成分分析 [2] がある. このようなデータの経時変化を捉えるという観点から, 本講演では GMANOVA モデルの拡張を考えた.

上述した例のように複数の項目に対して収集された経時測定データが得られたとき, 各項目の経時測定データに対して個別に GMANOVA モデルを用いて分析することができる. しかしながら, 各項目に対して個別に GMANOVA モデルを適用した場合, 項目の異なる経時測定データ間の関連などを分析結果から上手く解釈できないという問題がある. この問題を解決するために, 項目に関する情報を従来の GMANOVA モデルに組み込んだモデルとして Three-mode GMANOVA モデルを本講演では提案した. 本講演で提案した Three-mode GMANOVA モデルは次のように表されるモデルである,

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1' \\ \vdots \\ \boldsymbol{y}_n' \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1'\boldsymbol{\Theta}(\boldsymbol{C} \otimes \boldsymbol{X})' + \boldsymbol{\varepsilon}_1' \\ \vdots \\ \boldsymbol{a}_n'\boldsymbol{\Theta}(\boldsymbol{C} \otimes \boldsymbol{X})' + \boldsymbol{\varepsilon}_n' \end{pmatrix} = \boldsymbol{A}\boldsymbol{\Theta}(\boldsymbol{C} \otimes \boldsymbol{X})' + \boldsymbol{\mathcal{E}},$$

ここで, $\boldsymbol{Y}$ は $n \times mp$ 行列であり, $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m)$ という構造を持ち, $\boldsymbol{Y}_j$ $(j = 1, \ldots, m)$ は $n \times p$ 行列で j 番目の項目について収集した経時測定データからなる行列である. つまり, $\boldsymbol{Y}$ は n 個の個体に対し m 個の項目について, 各個体や各項目で p 回収集した経時測定データを横結合した行列とみなすことができる. また, $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)'$ は各個体の特徴を表す測定時点に無関係な値からなる $\mathrm{rank}(\boldsymbol{A}) = k$ $(< n)$ と仮定した $n \times k$ 個体間計画行列, $\boldsymbol{\Theta}$ は $k \times lq$ 未知パラメータ行列, $\boldsymbol{C}$ は項目間の関連性を表す $\mathrm{rank}(\boldsymbol{C}) = l$ $(< m)$ と仮定した $m \times l$ 項目間計画行列, $\boldsymbol{X}$ は時間トレンドを表現する関数からなる $\mathrm{rank}(\boldsymbol{X}) = q$ $(< p)$ と仮定した $p \times q$ 時点間計画行列である. 本講演で考えた上述のモデルは, $\boldsymbol{C}$ をモデルに組み込むことで項目間の経時測定データの関連性を捉える形のモデルとなっている (例えば [2] 参照). また, $\boldsymbol{X}$ には様々な基底関数を適用することができ, 例えば測定時点からなる多項式基底やフーリエ基底を経時変化に当てはめることができることを紹介した. 最後に, $\boldsymbol{\mathcal{E}} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)'$ は $n \times mp$ 誤差行列であり, 本講演では $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{mp}(\boldsymbol{0}_{mp}, \boldsymbol{\Omega})$ と仮定した. ここで $\boldsymbol{\Omega}$ は $mp \times mp$ 未知正定値行列である.

このモデルにおける未知の分散共分散行列 $\boldsymbol{\Omega}$ に含まれる要素数は, 項目数 m あるいは時点数 p の増加に伴って増加してしまうという問題があることを本講演では指摘した. そこで本講演では, この要素数の増加を抑えるために $\boldsymbol{\Omega}$ に [1] にあるようなクロネッカー構造を組み込み, $\boldsymbol{\Omega} = \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$ と仮定した. ここで, $\boldsymbol{\Psi}$ と $\boldsymbol{\Sigma}$

はそれぞれ $m \times m$, $p \times p$ 未知正定値行列である．また, $\mathbf{\Psi}$ と $\mathbf{\Sigma}$ の識別性の為, $\mathbf{\Psi}$ の $(1,1)$ 要素 $\mathbf{\Psi}_{1,1}$ は 1 とした．この仮定により, 要素数の増加が抑えられることを例示した．

しかしながら, このクロネッカー構造を組み込んだことなどにより, 未知パラメータ行列の推定量を陽な形で導出することができないという問題が発生してしまう．そこで, 本講演では, いくつかの未知の行列を固定した下での残りの未知行列の推定量を提案した．さらに, 提案した推定量を得るために, 本講演では反復計算を行う推定アルゴリズムの提案を行った．本講演では特に, 項目間計画行列が未知の場合に, 経時測定データなどから推定する手法を提案し, アルゴリズムを構築した．

本講演で構築したアルゴリズムを用いて, 実データの解析を行った．その結果, 大多数の個体についての経時変動が推定できることが分かった．しかし, いくつか問題も見つかったため, 今後は見つかった問題を解決するための手法を探っていこうと考えている．

参考文献

[1] Guggenberger, P., Kleibergen, F., Mavroeidis, S.: A test for Kronecker product structure covariance matrix. Jornal of Econometrics **233**, 88–112 (2023). doi:10.1016/j.jeconom.2022.01.005

[2] Kroonenberg, P. M., de Leeuw, J.: Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika **45**, 69–97 (1980). doi:10.1007/BF02293599

[3] Potthoff, R. F., Roy, S. N.: A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika **51**, 313–326 (1964). doi:10.2307/2334137

線形予測値で表される多変量回帰モデルにおける修正C_p規準

広島大学　大学院先進理工系科学研究科　柴山 佐内
広島大学　大学院先進理工系科学研究科　栁原 宏和

本研究では, 多変量線形回帰や多変量リッジ回帰, 多変量地理的加重回帰など, 予測値がハット行列で与えられる多変量回帰モデルにおける選択問題を考える. モデル選択規準として, 標準化予測二乗誤差に基づくリスク関数の不偏推定量として与えられる修正 C_p 規準を提案する. モデルの良し悪しを測る指標には様々なものがあるが, 標準化予測二乗誤差もその一つである. 修正 C_p 規準は, 多変量回帰モデルにおいて提案され (Fujikoshi and Satoh, 1997), 多変量リッジ回帰への拡張 (Yanagihara and Satoh, 2010) や一般化リッジ回帰への拡張 (栁原ら, 2009) などが行われている. 近年では, Yanagihara *et al.* (2023) により一般的なモデルを想定した修正 C_p 規準の拡張が行われた. 今回は前述した上記 3 つを含み, Yanagihara *et al.* (2023) には含まれないクラスでの修正 C_p 規準の拡張を試みた.

今, 標本数を n として, p 次元目的変数 $\boldsymbol{y}_i\,(i=1,\ldots,n)$ と切片項を含めた $k+1$ 次元説明変数 $\boldsymbol{x}_i\,(i=1,\ldots,n)$ が与えられているとする. このとき,

$$\boldsymbol{y}_i=\boldsymbol{\mu}_i(\boldsymbol{x}_i)+\boldsymbol{\Sigma}^{1/2}\boldsymbol{\varepsilon}_i\quad(i=1,\ldots,n),\quad \boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_n\sim i.i.d.\,F,$$

で与えられるモデルを考える. ただし, F は平均 $\boldsymbol{0}_p$, 分散 $\boldsymbol{I}_p$ の分布とする. $\boldsymbol{\Sigma}$ は p 次元正定値行列, $\boldsymbol{\mu}_i(\cdot)$ は $\mathbb{R}^{k+1}\to\mathbb{R}^p$ の滑らかな関数とする. また, モデルの真の構造として,

$$\boldsymbol{y}_i=\boldsymbol{\gamma}_i^*+\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\varepsilon}_i\quad(i=1,\ldots,n),\quad \boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_n\sim i.i.d.\,F_*,$$

を仮定する. F_* は平均 $\boldsymbol{0}_p$, 分散 $\boldsymbol{I}_p$ をもつ Orthant Symmetry (Efron, 1969) と呼ばれる対称分布である. $\boldsymbol{\gamma}_i^*$ は真の平均構造, $\boldsymbol{\Sigma}_*$ は p 次元正定値行列である. また, 目的変数行列 $\boldsymbol{Y}=(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)'$ の予測値と, 分散の不偏推定量が,

$$\hat{\boldsymbol{Y}}=\boldsymbol{H}\boldsymbol{Y},\quad \boldsymbol{S}=\frac{1}{\mathrm{tr}(\boldsymbol{M})}\boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y},$$

で与えられているとする. ただし, $\boldsymbol{H}$ は n 次元正方行列, $\boldsymbol{M}$ は n 次元半正定値行列である.

今, 標準化二乗距離を $d(\boldsymbol{A},\boldsymbol{B})=\mathrm{tr}\{(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{A}-\boldsymbol{B})'\}$, その推定量である推定標準化二乗距離を $\hat{d}(\boldsymbol{A},\boldsymbol{B})=\mathrm{tr}\{(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{S}^{-1}(\boldsymbol{A}-\boldsymbol{B})'\}$ で定める. また, $\boldsymbol{\Gamma}_*=(\boldsymbol{\gamma}_1^*,\ldots,\boldsymbol{\gamma}_n^*)'$, $\boldsymbol{\mathcal{E}}=(\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_n)'$ とおき,

$$L_C(\boldsymbol{A})=\mathrm{E}\left[d(\boldsymbol{Y},\boldsymbol{A})\right]=np+d(\boldsymbol{\eta},\boldsymbol{A}),$$

とする. 修正 C_p 規準は, 次の標準化予測二乗誤差に基づくリスク関数の不偏推定量として与えられる.

$$R_C=\mathrm{E}\left[L_C(\hat{\boldsymbol{Y}})\right]=\theta_*+\mathrm{tr}(\boldsymbol{H}'\boldsymbol{H})+np.$$

ただし, $\theta_*=d(\boldsymbol{\Gamma}_*,\boldsymbol{H}\boldsymbol{\Gamma}_*)$ である. このリスク関数 R_p の不偏推定量を求めるには, 未知パラメータ θ_* の推定が必要となる. そこで, 次の期待値を利用して, θ_* の不偏推定量を構成することを考える.

$$\mathrm{E}\left[d(\boldsymbol{Y},\hat{\boldsymbol{Y}})\right]=\mathrm{E}\left[\hat{d}(\boldsymbol{\Gamma}_*,\boldsymbol{H}\boldsymbol{\Gamma}_*)\right]+\mathrm{E}\left[\hat{d}(\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_*^{1/2},\boldsymbol{H}\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_*^{1/2})\right].$$

まず, 第 1 項については, $\boldsymbol{\Theta}_*=\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Gamma}_*'(\boldsymbol{I}_n-\boldsymbol{H})'(\boldsymbol{I}_n-\boldsymbol{H})\boldsymbol{\Gamma}_*\boldsymbol{\Sigma}_*^{-1/2}$ とおけば, $\mathrm{tr}(\boldsymbol{\Theta}_*)=\theta_*$ であり,

$$\mathrm{E}\left[\hat{d}(\boldsymbol{\Gamma}_*,\boldsymbol{H}\boldsymbol{\Gamma}_*)\right]=m\mathrm{tr}\left\{\mathrm{E}\left[\left(\boldsymbol{\mathcal{E}}'\boldsymbol{M}\boldsymbol{\mathcal{E}}\right)^{-1}\right]\boldsymbol{\Theta}_*\right\}=mc_1\theta_*,$$

を得る. ただし, 2 つ目の等号は次の補題より得られる.

補題 (第 1 項における期待値)**.** $\boldsymbol{\mathcal{E}}'\boldsymbol{M}\boldsymbol{\mathcal{E}} = \begin{pmatrix} z_{11} & \boldsymbol{z}_{21}' \\ \boldsymbol{z}_{21} & \boldsymbol{Z}_{22} \end{pmatrix}$ とおき, $z_{11\cdot 2} = z_{11} - \boldsymbol{z}_{21}'\boldsymbol{Z}_{22}^{-1}\boldsymbol{z}_{21}$ とする. このとき, $c_1 = \mathrm{E}\left[1/z_{11\cdot 2}\right]$ として, 以下が成り立つ.

$$\mathrm{E}\left[\left(\boldsymbol{\mathcal{E}}'\boldsymbol{M}\boldsymbol{\mathcal{E}}\right)^{-1}\right] = c_1\boldsymbol{I}_p.$$

また, 第 2 項については, $\boldsymbol{C}_2 = \mathrm{E}[\boldsymbol{\mathcal{E}}\left(\boldsymbol{\mathcal{E}}'\boldsymbol{M}\boldsymbol{\mathcal{E}}\right)^{-1}\boldsymbol{\mathcal{E}}']$ とおくことで, 次のように整理できる.

$$\mathrm{E}\left[\hat{d}(\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_*^{1/2}, \boldsymbol{H}\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_*^{1/2})\right] = m\mathrm{tr}\left\{\boldsymbol{C}_2(\boldsymbol{I}_n - \boldsymbol{H})'(\boldsymbol{I}_n - \boldsymbol{H})\right\}.$$

実際には, c_1, $\boldsymbol{C}_2$ の期待値は,

$$c_1 = \lim_{\alpha\to\infty}\frac{1}{\alpha}\sum_{\ell=1}^{\alpha}\frac{1}{p}\mathrm{tr}\left\{(\boldsymbol{\mathcal{E}}_{(\ell)}'\boldsymbol{M}\boldsymbol{\mathcal{E}}_{(\ell)})^{-1}\right\}, \quad \boldsymbol{C}_2 = \lim_{\alpha\to\infty}\frac{1}{\alpha}\sum_{\ell=1}^{\alpha}\boldsymbol{\mathcal{E}}_{(\ell)}\left(\boldsymbol{\mathcal{E}}_{(\ell)}'\boldsymbol{M}\boldsymbol{\mathcal{E}}_{(\ell)}\right)^{-1}\boldsymbol{\mathcal{E}}_{(\ell)}',$$

の形を用いて, モンテカルロ積分より求められる. ただし, $\boldsymbol{\mathcal{E}}_{(\ell)}$ は ℓ 回目の繰り返しによる誤差行列である. 以上より,

$$\mathrm{E}\left[\hat{d}(\boldsymbol{Y}, \hat{\boldsymbol{Y}})\right] = mc_1\theta_* + m\mathrm{tr}\left\{\boldsymbol{C}_2(\boldsymbol{I}_n - \boldsymbol{H})'(\boldsymbol{I}_n - \boldsymbol{H})\right\},$$

が導かれる. したがって, θ_* の不偏推定量を,

$$\hat{\theta}_* = \frac{1}{mc_1}\hat{d}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) - \frac{1}{c_1}\mathrm{tr}\left\{\boldsymbol{C}_2(\boldsymbol{I}_n - \boldsymbol{H})'(\boldsymbol{I}_n - \boldsymbol{H})\right\},$$

とおき, 修正 C_p 規準を次で定義することで, 今までの議論より以下の定理が成り立つ.

定義 (修正 C_p 規準)**.**

$$MC_p = \hat{\theta}_* + p\mathrm{tr}(\boldsymbol{H}'\boldsymbol{H}) + np.$$

定理. MC_p はリスク関数 R_p の不偏推定量である. つまり,

$$\mathrm{E}[MC_p] = R_p.$$

ただし, $\hat{\theta}_*$ の計算では, 期待値が具体的に求まらないため, すでに述べたようにモンテカルロ積分を利用する. そのため, 誤差項の従う具体的な分布をあらかじめ知っておくことが必要となる. また, 誤差項に正規分布を仮定した場合や $\boldsymbol{M}$ にべき等性をもたせた特別な仮定を置いた場合では, よりシンプルな形で $\hat{\theta}_*$ を表すことができる.

参考文献

[1] Efron, B. (1969). Student's t-test under symmetry conditions. *J. Amer. Stat. Assoc.*, **64**, 1278–1302.

[2] Fujikoshi, Y. & Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707–716.

[3] 栁原 宏和, 永井 勇, 佐藤 健一 (2009). 多変量一般化リッジ回帰におけるリッジパラメータ最適化のためのバイアス補正 C_p 規準. 応用統計学, **38**, 151–172.

[4] Yanagihara, H. & Satoh, K. (2010). An unbiased C_p criterion for multivariate ridge regression. *J. Multivariate Anal.*, **101**, 1226–1238.

[5] Yanagihara, H. , Nagai, I. , Fukui, K. & Hijikawa, Y. (2023). Modified C_p criterion in widely applicable models. *Smart Innov. Syst. Tec.*, **352**, 173–182.

有限混合モデルの混合事前分布を用いた確率的ブロックモデルのベイズ推測

岩重文也，橋本真太郎
広島大学大学院先進理工系科学研究科

ネットワーク解析においては，背後に確率的なブロック構造を仮定する確率的ブロックモデル (stochastic block model, 略して SBM) がよく用いられる．ネットワークデータとしては例えば，SNS のフォローワーの関係や，論文の引用・被引用などがある．ここでは特に，ベイズ統計学の枠組みにおいて確率的ブロックモデルを用いたネットワークデータのコミュニティ検出について考える．単純無向グラフにおける，ノード数 n，コミュニティ数 k の SBM として以下のものを考える:

$$A_{ij}|z,Q,k \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{ij}), \quad \theta_{ij} = Q_{z_i z_j}, \quad 1 \le i < j \le n.$$

ただし，$z_i = \{1,\dots,k\}$ はコミュニティ割り当て，$Q = (Q_{rs}) \in [0,1]^{k\times k}$ は対称行列で各成分は 2 つのコミュニティから任意にノードを取ってきたときに辺が引かれる確率を表す．ただし，隣接行列 $A = (A_{ij}) \in \{0,1\}^{n\times n}$ は対称行列で自己ループは許容しない，つまり，$A_{ij} = A_{ji}$, $A_{ii} = 0$ である．ネットワークデータのクラスタリング問題はコミュニティ検出問題とも呼ばれ，特に SBM を中心に理論と方法論が発展してきている．しかし，コミュニティ数の選択問題は非常に難しい問題である．近年，コミュニティ数自体も不確実なものであると捉え，事前分布を仮定して推測を行う方法が発展してきている．これにより，コミュニティ数をデータから推定したり，その不確実性の評価が可能になる．一つの方法として，コミュニティ数に事前分布を仮定し，事後分布からサンプリングを行う方法があるが，リバーシブルジャンプ MCMC に代表される非効率かつ複雑な MH アルゴリズムを使用しなくてはならずスケーラブルではない．ノンパラメトリックベイズの枠組みにおいて，クラスタリングに対する事前分布として中華料理店過程 (Chinese restaurant process, CRP) がよく用いられる．この場合は，効率的なギブスサンプラーを構成することが可能であるが，クラスタ数に関する事後分布の一致性が成り立たないことがあることが知られている．この問題を解決するために，Miller and Harrison (2017) は次の有限混合モデルの混合 (mixture of finite mixture model, MFM) を用いた事前分布を提案し，この事前分布のもとでのクラスタ数の事後分布の一致性を示した．

$$k \sim p(\cdot), \quad (\pi_1,\dots,\pi_k)|k \sim \text{Dir}(k;\gamma,\dots,\gamma), \quad z_i|k,\pi \sim \sum_{h=1}^{k} \pi_h \delta_h, \quad i = 1,\dots,n.$$

ただし，k はクラスタ数，$p(\cdot)$ は $\{1,2,\dots\}$ 上の確率関数，各 z_i はクラスタ割り当て，$\gamma > 0$，各 δ_h はディラック測度である．Geng et al. (2019) では，MFM 事前分布を SBM に基づくコミュニ

ティ検出の問題に適用し，いくつかの限定的な仮定の下で SBM におけるコミュニティ数の事後分布の一致性を示した．しかしながら，Geng et al. (2019) の方法は，ネットワークの辺に重みがある場合や自己ループを持つようなネットワークには適用できない．例えば，Train bombing data（列車事故におけるテロリストのネットワーク, `http://konect.cc/networks/moreno_train/`）では，辺の重みがテロリストによる関係性の深さを表している．

本研究では Geng et al. (2019) のモデルを拡張し，事後分布の計算のための効率的なマルコフ連鎖モンテカルロ法を構成する．具体的には，Geng et al. (2019) のモデルの尤度関数をポアソン分布に変えることで多重辺と自己ループを許したモデルを考える．つまり，

$$
\begin{aligned}
&A_{ij}|z, Q, w, k \overset{\text{ind}}{\sim} \text{Poisson}(\theta_{ij}), \quad \theta_{ij} = Q_{z_i z_j}, \quad 1 \leq i < j \leq n, \\
&A_{ii}/2|z, Q, w, k \overset{\text{ind}}{\sim} \text{Poisson}(\theta_{ii}/2), \quad i = 1, \dots n.
\end{aligned}
$$

自己ループは 2 倍して捉えられるので，実際の隣接行列の対角成分は 2 の倍数となることを考慮している．Geng et al. (2019) では，共役性の観点から対称行列 Q にはベータ分布を仮定しているが，提案モデルでは尤度にポアソン分布を仮定するので Q には共役なガンマ分布を仮定する．このモデルに対して，本研究では Geng et al. (2019) と同様，コミュニティ数に関して MFM 事前分布を用いることを考える．事後分布の計算においては，Geng et al. (2019) と同様に k, π について周辺化することで z の分布に着目する．この正当性は，k, π に関して周辺化した際に解析的に計算可能であること，z は交換可能であること，そして Dirichlet process mixtures と同様の Pólya urn scheme を構成することに依る．z の distinct value の事後分布が漸近的にコミュニティ数の事後分布に一致することが Miller and Harrison (2017) により示されているため，近似的にはコミュニティ数の事後分布を用いた不確実性の評価を行うことが可能になることもこのアプローチの重要な点である．また，SBM では同一のコミュニティに属するノード同士が関係を持つかどうかは確率的に等しいという欠点がある．ノード間の異質性を考慮した次数修正確率的ブロックモデル（degree-corrected stochastic block model, DCSBM） がある．ポアソン尤度を考えることで次数修正パラメータとモデルの識別可能性のための条件を加えれば DCSBM への拡張も容易である (e.g. Herlau and Mørup, 2014; 新田, 2018)．しかしながら，MFM 事前分布を用いたときに，Geng et al. (2019) で示されたような漸近的な性質が成り立つかどうかは未知であり，検討中である．当日は，いくつかの他の手法との数値比較と簡単な実データ解析を報告する予定である．

参考文献

Geng, J., A. Bhattacharya, and D. Pati (2019). Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association 114*(526), 893–905.

Herlau, M. N. S. and M. Mørup (2014). Infinite-degree-corrected stochastic block model. *Physical Review E 90*(032819).

Miller, J. W. and M. T. Harrison (2017). Mixture models with a prior on the number of components. *Journal of the American Statistical Association 113*(521), 340–356.

新田 (2018). 次数修正確率ブロックモデルにおけるベイズ推定, 東京大学大学院情報理工学系研究科数理情報学専攻修士論文 要旨.

Asymptotic properties for small area estimation in extreme value theory

鹿児島大学　桃木 光輝　　鹿児島大学　吉田 拓真

本研究の背景

極値統計学では，稀な事象を解析するための様々な統計的モデリング手法が提案されてきた．本研究の主な関心は，複数のエリアで収集されたデータから，各エリアのリスクを精度良く予測することにあった．たとえば，図1は，日本中の代表的な気象観測所（エリア）の位置を示した地図である．この場合，大きな被害をもたらす規模の豪雨が発生するリスクを解析したい．ところが，各エリアの利用可能なデータはそれほど多くない．図2の上段パネルは，京都の日降水量の時系列プロットを示している（`https://www.data.jma.go.jp/gmd/risk/obsdl/index.php`）．このとき，極値理論に基づくモデリングでは，図2の下段パネルのように，年最大値や閾値超過に加工された大きな値のデータだけが使用される．それゆえ，各エリアのデータ数は少なくなる傾向にあり，データをエリアごとに隔てて解析してしまうと，得られる推定値は分散が大きいために信頼性に欠けてしまうという問題があった．

図 1: 代表的な気象観測所の位置

本研究の概要

先行研究では，クラスタリングを通して，同質な複数のエリアのデータを1つのエリアにプールすることで，リスクの予測精度を安定化させる方法が研究されてきた（de Carvalho et al. 2023 ; Dupuis et al. 2023）．本研究では，それらとは異なるアプローチとして，小地域推定と呼ばれる分野の混合効果モデルを極値統計理論に組込んだ．

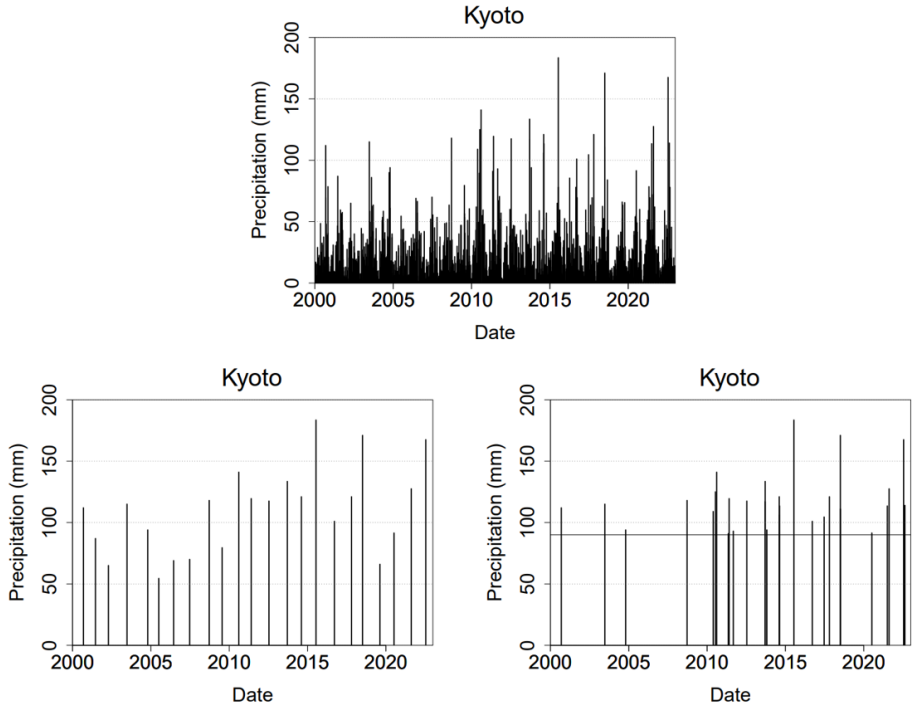


図 2: 京都の降水量の時系列プロット（上：日降水量，左下：年最大値，右下：閾値超過）

混合効果モデルの特徴は，エリアごとの推定値を取得するのに，その他の全てのエリアの情報を利用するところにある．このような特徴は,「Borrowing of strength」と呼ばれる．実際，小地域推定において，混合効果モデルは，サンプルサイズが小さいエリアの推定値を改善するための道具として重宝されてきた（Sugasawa and Kubokawa 2020）．先に述べたように，極値解析に利用できるデータは限られる．それゆえ，混合効果モデルは，リスクの予測精度の底上げに貢献できると考えた．

講演では，混合効果モデルを組込んだ極値理論の性質とそのパフォーマンスについて報告した．

参考文献

[1] de Carvalho, M., Huser, R., Rubio, R. (2023). “Similarity-based clustering for patterns of extreme values”. *Stat* **12**(1), e560.

[2] Dupuis, D. J., Engelke, S., Trapin, L. (2023). “Modeling panels of extremes”. *The Annals of Applied Statistics* **17**(1), 498-517.

[3] Sugasawa, S., Kubokawa, T. (2020). “Small area estimation with mixed models: a review”. *Japanese Journal of Statistics and Data Science* **3**(2), 693-720.

情報理論におけるエントロピー概念の誤り訂正

S/N(信号対雑音)比とエントロピーを考える

得丸久文 Tokumaru Kumon

1　デジタルは自動的に複雑進化する

デジタルとは、①　離散・有限信号の一次元配列によって情報を伝達すると、②　自律的に自己増殖し、複雑進化するシステムである。生命システムは、RNA と DNA という 4 種の核酸をデジタル信号として、原核生物から真核生物、脊椎動物、哺乳類へと複雑進化した。ヒトの言語は、哺乳類の鳴き声が、周波数離散成分「音素」と時間離散成分「モーラ(拍)」を獲得してデジタル化し、文字（消えない音節）と bit（対話する音節）の発明によって跳躍的に進化した。

2　生命の跳躍進化を生み出す情報源の低雑音環境

原核生物➔真核生物➔脊椎動物➔哺乳類という跳躍進化は、生物学も進化学も説明できていない。情報理論とエントロピーを熱力学的に理解することで、そのメカニズムが姿を現す。

地球上に小惑星や巨大隕石が衝突していた時期に、生命体は生命情報過程を環境雑音から保護するための突然変異を生み出す。細胞膜、核膜、脳室・脳脊髄液、子宮が、それぞれ原核生物、真核生物、脊椎動物、哺乳類を生み出した。この保護装置は、生命情報過程が成り立たないほどの雑音を打ち消す能力をもつので、S/N 比で+80dB 規模の保護環境だと考える。

厳しい生息環境が、元の静かな環境に復元したとき、保護装置は+80dB を余剰能力としてもつことになり、それを利用して、今まで考えたこともなかったような複雑進化を生み出す。(表 1)

元の状態	時期/ 新種	雑音を低減する物理進化	低雑音下で洗練された論理進化
生命以前	38 億年前 / 生命誕生 原核生物	細胞膜	核酸 (DNA と RNA: GC, AU), 転写 (DNA➔RNA), 翻訳 (RNA➔アミノ酸)
原核生物	20 億年前 / 真核生物	核膜で保護された核	DNA 二重らせん, 制限酵素, 転写後修飾, オルガネラとの共生 (葉緑体, ミトコンドリア)
真核生物	5 億 4000 万年前 脊椎動物	多細胞生物, 中枢神経（脳室、脳脊髄液）	感覚運動メカニズム 記号反射
有袋類	6600 万年前 哺乳類	子宮、胎盤、授乳、大きな脳、温かい血	社会性、好奇心、学習、音声コミュニケーション

表 1　原核生物から哺乳類までの生命体の物理・論理進化 (著者作成)

3　知能の跳躍進化を生み出す情報源の低雑音環境

ヒトの言語は、周波数離散成分である音素、時間離散成分であるモーラ(拍)をもつ音節を獲得して、デジタル進化が始まった。音節は音素であるので、順列によって無限の造語が可能である。音節のあと、文字と bit が発明されたが、それぞれ低雑音環境で論理進化を生み出すことで跳躍的な知能進化が生まれた。

文法を、「主として単音節の付加・変化によって意味の修飾や接続を指示することで、習得すると無意識に変調・復調ができる。」と定義する。ヒトは無意識の復調のために、母語を片耳で聴き取り、脳幹聴覚神経核の方向定位能力で文法語の音韻ベクトルを解析している。片耳聴覚するので、周囲の外敵からの防御が手薄になる。文法を使うためには、安全な開墾地で生活する必要がある。ブラジルの熱帯雨林に暮らすピダハンは大人も母語を両耳聴覚しており、文法がない。

概念は、「その言葉と結びつくすべての記憶の集合を意味とする」。条件反射や日常語は、言葉と意味が 1 対 1 で結びつく。経験と思考を重ねて、言葉と意味が 1 対全の関係になるものが概念である。概念を獲得するためには、徹底的に思考して、例外がないことを確かめる必要がある。例外がないことが、1 対全が成り立つということと等価であるからだ。概念は僧院や修道院で生まれたが、それを獲得し正しく使うためには、世俗から隔離された低雑音環境が必要だからだ。

前方誤り訂正とは、「情報発信者にコンタクトすることなく、情報の誤りを訂正する」技術である。これは、①著者本人の真正な言葉であることを確かめ、②著者がどう考えてその結論を出したかを追体験することで確かめられる。責任のない言葉は誤り訂正できない。思考過程の明らかでない知識は、誤りかどうかを判断しようがない。くり返し読み、吟味し、他の分野科学とも総合して、どこにも曖昧さや矛盾がないことを確かめると、前方誤り訂正ができる。それを実施するためには、超低雑音環境に身を置くことが必要であろう。

元の状態	時期/ 新種	環境・信号の物理進化	低雑音下で洗練された論理進化
哺乳類	300 万年前 / 初期人類	ロックシェルター(岩陰) 直立二足歩行	道具の使用 調理に火の使用(30 万年前)
原人	13 万年前 / 身体的現代人	洞窟居住(晩成化) 1400cc の大きな脳	祖父母の家族参加(真社会性化) 仲間を識別する鳴き声の離散化
真社会性 動物	7 万年前 現生人類	トバ火山灰の冬期 喉頭降下して音節の獲得 開墾地に住む	音素共有(クリック子音) 文法処理（母語の片耳聴覚）
口誦文化	5000 年前 文明人	文字(消えない音節)獲得 修道院・僧院	学校教育 科学的概念(群は例外がない)
文明	60 年前 ホモサピエンス	Bit(対話する音節)とインターネット検索エンジン	人類共有知、前方誤り訂正 (ネットワークの矛盾をなくす)

表 2　哺乳類からホモサピエンスまでの生命体の物理・論理進化（著者作成）

動物の移動軌跡モデルと presence and pseudo-absence data

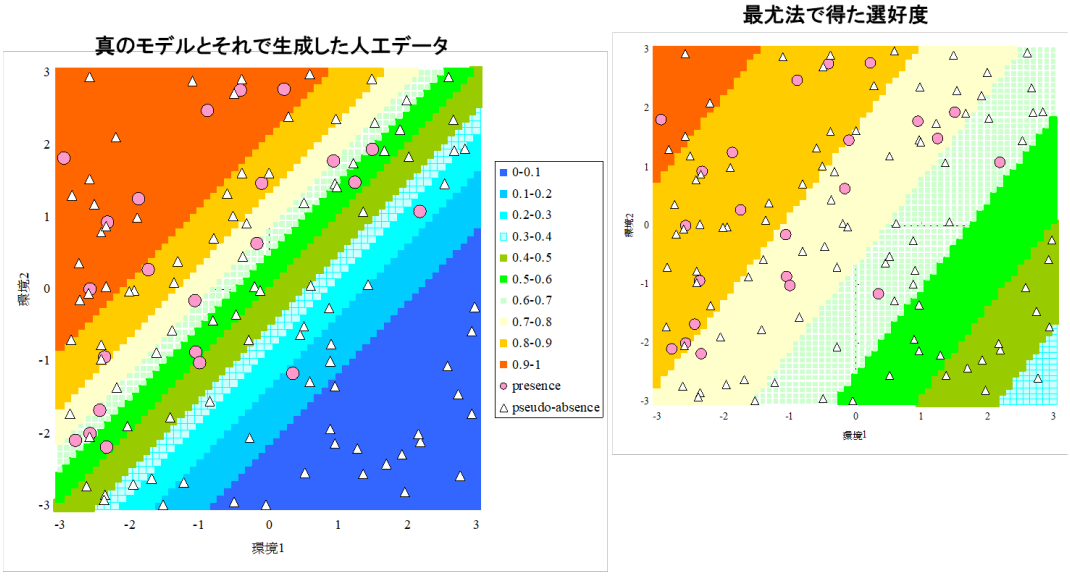
島谷健一郎・統計数理研究所

野生動物に GPS (global positioning system) ロガーを装着する動物学研究は、2000 年代半ばころから盛んになり、現在では広く普及している。GPS 軌跡は動物がどこを訪れたかを明示する。また、ある動き（移動）の次にどう動いた（移動した）かも示すし、ある領域から別の領域までどういうルートで移動したかも見て取れる。

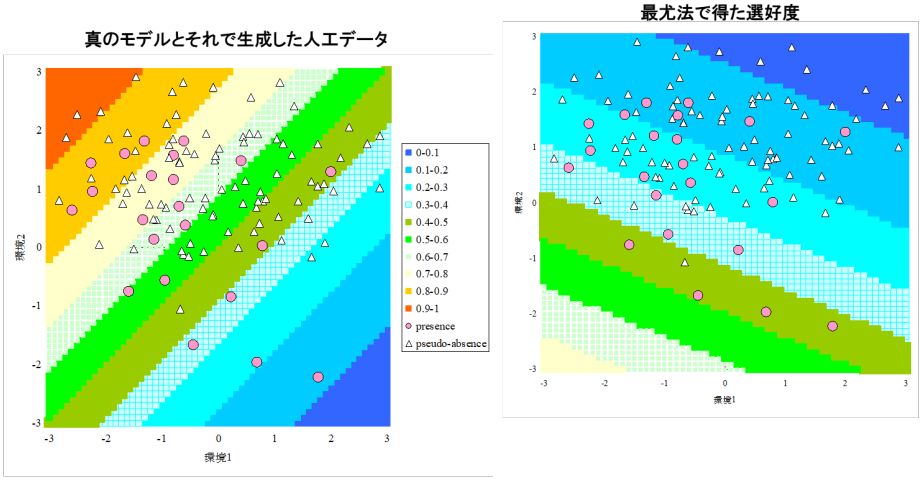
ところで、特定の場所を訪れたという情報には、訪れなかった場所という情報も含んでいる。　つまり、GPS データには、ある地点を訪れた・訪れなかった、次はどう移動した・移動しなかった、どのルートで移動した・どれは使われなかった、のように、presence-absence (0-1) データの情報を含有する。したがって、地点やルートの環境情報などを入手すれば、ロジスティック回帰などで、どういう環境の地点を訪れ、どういう所は避け、次は環境条件がどういう地点へ移動しどういう方向は避け、どういう環境条件下のルートをより好んで使うか等々を定量的に評価できるはずである。

GPS 測位点から presence データは得られるが、absence データは、測位地点やルート以外すべてだから連続無限個ある。2000 年代に入り、この問題は統計学でも議論の俎上に上るようになった。無限にある（証拠のない）absence 候補から、どのように有限個の absence を生成するか（pseudo-absence と呼ばれる）。特に、ルート選択では path を random walk で生成してそこに環境情報を付ける。その結果を見て初めて、環境変量空間で pseudo-absence の分布と presence の分布を比べられる。

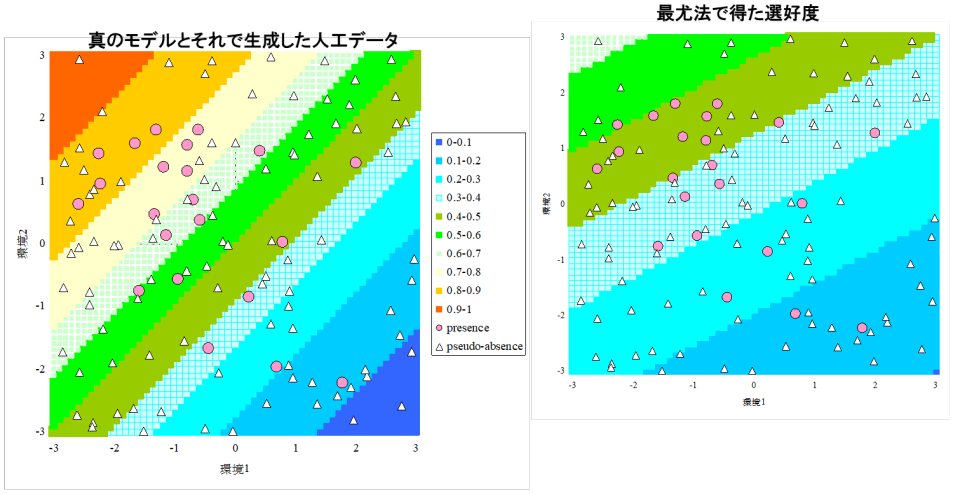
下の図の左のように pseudo-absence が環境変数空間で一様ランダムに分布し、presence が傾向を持って分布していたら、右のように動物の環境選好度は適切に推定できる。



しかし、次の図の左のように pseudo-absence が presence を覆っていない分布だと、右のように選好度を推定できない。



そしてこれと同じ presence に対し、pseudo-absence を以下の左ように生成していれば、右のように選好度を推定できる。



Pseudo-absence 生成法は、かなり厄介な問題のようである。現状では、自然に思える random walk アルゴリズムで path を生成し、pseudo-absence の分布が presence の分布を覆っていたらひとまず妥当な結果を期待できると考え、研究事例を積み重ねていく路線が有効と考えられる。

空間標識再捕獲法による個体密度と景観連結性の推定

深澤圭太（国立環境研究所）

空間標識再捕獲モデルは空間的に配置された複数の検出器によって得られた標識再捕獲データから個体密度を推定するための統計モデルであり、生態学や野生動物管理の評価において幅広く用いられている。本研究では、景観構造によって異方的なホームレンジを形成するメカニズムを組み込んだ空間標識再捕獲モデル（移流拡散標識再捕獲モデル, ADCR）を開発し、個体密度に加えて場所間の個体の透過性を推定することが可能となった。ADCR の詳細は、プレプリント(Fukasawa, 2023, https://www.biorxiv.org/content/10.1101/2023.03.01.530712v2)にてすでに公開している。

空間標識再捕獲モデルの検出モデルは、動物個体が検出器（個体識別可能な DNA サンプラーやカメラトラップ等）に遭遇する確率的プロセスをホームレンジ行動と関連付けて記述している。個体のホームレンジは空間内の個体の利用頻度に相当する確率分布（利用分布）として記述される。一般的に、利用分布は個体が引き寄せられる活動中心 $\boldsymbol{\mu} = (\mu_x, \mu_y)$を潜在変数として含む。例えば、既存の空間標識再捕獲モデルは利用分布が定常的な二変量正規分布 $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$であることを仮定している。本研究では不均一な透過性景観の影響を受けて変化するより柔軟な利用分布を導入するために、個体が検出器に遭遇する頻度と利用分布、$p^*(\mathbf{s} \mid \boldsymbol{\mu})$との間に明示的なリンクを記述する検出モデルを採用した。ここで、$p^*(\mathbf{s} \mid \boldsymbol{\mu})$は活動中心 $\boldsymbol{\mu}$ を持つ個体の位置 $\mathbf{s} = (x, y)$での確率密度を定義している。

空間標識再捕獲調査は空間に配置された複数の検出器により行われる。検出器では、その場所に滞在した個体それぞれの検出回数または検出の有無が記録される。調査は複数の時間区間(オケージョン)で繰り返される場合、調査期間全体を単一のオケージョンとして扱う場合の両方がある。検出器と動物行動が相互に独立な場合、k 番目のオケージョンに i 番目の個体が j 番目の検出器に遭遇する頻度 λ_{ijk} は次のように記述される。

$$\lambda_{ijk} = \exp(g_0) p^*(\mathbf{s}_j \mid \boldsymbol{\mu}_i) \qquad \text{(eqn. 1)}$$

ここで、g_0 は対数検出効率のパラメータ、$\boldsymbol{\mu}_i$ は i 番目の個体の活動中心、$\mathbf{s}_j$ は j 番目の検出器の位置である。検出イベントが相互に独立で、検出器が個体ごとの検出回数を記録する場合、検出回数 Y_{ijk} はポアソン分布 $\text{Poisson}(Y_{ijk} \mid \lambda_{ijk})$に従う。これは従来の空間標識再捕獲モデルと同様に、さまざまな検出器タイプ（例えば滞在・非滞在の 2 値観測等）および不均一な観測努力量に容易に拡張することが可能である。

ADCR の利用分布 $p^*(\mathbf{s} \mid \boldsymbol{\mu})$は、異質な透過性景観において定常的なホームレンジを形成する微視的な移動モデルから導出される。Ornstein-Uhlenbeck 過程は、活動中心へのドリフトとランダムウォークに従う動物のホームレンジ行動を記述する基本的な移動モデルである。透過性はランダムウォークのステップ長に対応している。Ornstein-Uhlenbeck 過程の下では、個体位置の確率分布は定常分布に収束する。本研究では、透過性が景観共変量（パッチとマトリクス等）に依存する Ornstein-Uhlenbeck 過程を考え、その下で生じる定常分布を利用分布とした。

利用分布を導出するために、ドリフトとランダムウォークによって駆動される空間

内の確率密度の流れに着目したオイラー型の定式化を行った。Ornstein-Uhlenbeck 過程の下では、個体位置の確率分布は以下の Fokker-Planck 方程式に従う。

$$\frac{\partial p}{\partial t} = -c\left((\mu_x - x)\frac{\partial p}{\partial x} + (\mu_y - y)\frac{\partial p}{\partial y}\right) + D(x,y)\left(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2}\right) \quad \text{(eqn.2)}$$

ここで、$D(x, y)$ (> 0)は空間的に変化する拡散係数であり、c (> 0) はドリフトの強度である。右辺の第 1 項は活動中心へのドリフトに対応する移流項であり、第 2 項はランダムウォークに対応する拡散項である。空間内で変化する拡散係数は透過性（またはサーキット理論におけるコンダクタンス）と解釈される。景観共変量が透過性に与える影響をモデル化するために、本研究では対数線形モデル $\ln(D(x, y)) = \beta_0 + \boldsymbol{\beta}\mathbf{z}(x, y)$を適用した。ここで β_0 は切片、$\mathbf{z}(x, y)$は位置(x, y)での景観共変量のベクトルであり、$\boldsymbol{\beta}$ は透過性が景観共変量に依存する程度をコントロールする係数である。ドリフトの強度 c は正値を取るため、$\ln(c)$をパラメータ α_0 として表す。

本研究では、拡散係数に影響を与える景観共変量がラスターデータセットとして利用可能な一般的な状況を想定し、このシステムにおける定常確率密度分布 $\mathbf{p^*(s \mid \mu)}$の数値解を有限差分法により得た。モデルに含まれるパラメータは、活動中心を積分消去した周辺尤度最大化法により求めることが可能で、最尤推定の毎ステップで定常確率密度分布を計算して尤度計算している。

移動シミュレーションにより生成した疑似データに対して本手法を適用したところ、個体密度やホームレンジ形状をバイアスなく推定できることが分かった。また、推定値はデータ生成プロセスと利用可能な景観データの解像度の不一致に対して非常に頑健であった。本手法は、不均一な景観における個体・個体群レベルの現象の統合的な理解に有用であると考えられる。

シミュレーションによるオンライン教育に対する効果分析

山形大学　学術研究院　加納　寛子

研究の背景

＜オンライン授業と学生生活＞

近年，とりわけ Covid-19 パンデミックを機に，オンライン教育（デジタル機器を使用した遠隔教育・オンライン授業と同義）が急速に拡大している．大学生協が実施した「第 56 回学生生活実態調査」によれば，Covid-19 パンデミックの影響により，オンライン授業が常態化した 2020 年の学生のうち特に 1 年生は，「学生生活が充実している」と感じる 1 年生は 83 年以降最低値 56.5％にとどまったことが報告されている．同項目について，2015 年および 2019 年の調査では，すべての学年において，85％以上であったことから，30 ポイント以上充実度が低下していた．また，同調査によれば，学生生活が充実していない要因として，1 年生の 7 割が「授業が充実していない」と回答していた．

また，同「第 58 回学生生活実態調査」によれば，授業形態が対面授業に戻りつつある 2021 年 2022 年の大学 1 年生の大学生の充実度は回復してきている．さらに，対面授業の割合が高い学生ほど，学生生活が充実していると回答している傾向が報告されていた．

さらに，2021 年 10 月 10 日の朝日新聞では「在宅のオンライン授業に孤独感」というタイトルで，対面授業を望む一人の学生の声が掲載された．だが，Twitter 上のつぶやきを分析したところ，一定数の対面を望む声とともに，オンライン授業を望む声も見られた（加納，2022）．

＜デジタルと伝統教育＞

デジタルと伝統教育の比較については，紙&鉛筆とタブレット PC の比較をしている研究では，操作性について，紙&鉛筆が 4.58 であるのに対し，タブレット PC が 3.18 であり，紙&鉛筆の方が，タブレット PC 利用よりも操作性が高い，という結果が報告されている（竹野，藤田，2015）．一方で，学習意欲については，タブレット PC を活用する方が伝統的教育よりも，自分の課題達成に対する意識や，学習意欲が向上することが報告されている（南，長谷川，2016）．

研究の目的

オンライン教育の形態が学習者に与える影響は多岐に渡り，これらの手法やツールが教育に与える影響については，しばしば経験則に依存した評価が行われがちである．竹野，藤田（2015）の調査では，紙&鉛筆の方が，タブレット PC 利用よりも操作性が高い，という結果が報告されていたが，筆者の感覚では，紙&鉛筆よりもタブレット PC の方が，様々な点で利便性が高いと感じている．実際，紙の手帳を使用している人をあまり見かけなくなり，ビジネスマンの多くは，デジタルツールを用いて，スケジュール管理やメモなどをしている場合が多い．

実データに基づいた効果検証は重要であるが，母集団の資質や嗜好に影響を受けがちである．フィードバックを目的とする場合，実データによる評価は妥当であるが，同程度の偏差値の学校であれば，文系クラスと理系クラスでは，数学のテストの平均値が文系クラスよりも理系クラスの方が高くなるであろうことは，実測する前から容易に推測がつく．一方で，いじめの経験について質問紙調査をした場合，たまたま一つの学校で，理系クラスの方が文系クラスよりも経験が高い結果が得られたとしても，多様なバックグラウンドが影響するため，すべての学校にあてはまるとは結論づけられない．

また，デジタル機器を活用し始めた時期は家庭環境に依存しており，全く活用していない状態から活

用の効果を長期にわたり実測することは非常に困難である. そこで, 本研究の主な目的は, 伝統的な教育とオンライン教育が学習者の学習レベルに与える影響を, シミュレーションによって定量的に評価することである. 特に, オンライン教育のプラス効果とマイナス効果が時間の経過とともにどのように変化するのかを解析する.

方法

研究の手続きは, Python プログラムを用いたシミュレーションによって行う. シミュレーションは 10 週間の期間で設定し, 毎週の学習効果を測定した. 100 名の学習者を仮定し, その学習レベルを 0 から 100 の範囲でランダムに初期化した. 伝統的な教育とオンライン教育の二つのケースについてシミュレーションを実施し, それぞれのケースで学習レベルの平均値を算出した.

パソコンやタブレットを使用したからといって, すぐに結果に現れるわけではなく, 徐々に影響が出ると推測する, オンライン教育においては, 「指数関数的なプラス効果」を用いる. また, デジタル機器を使用し始めた直後は, 様々なアクシデントや戸惑いがあり, 使用しない場合に比べて負担となると予測されるが, 使用していくうちに慣れてきて負担が軽減されて行くであろうと仮定し, 「徐々に減少するマイナス効果」を加味した. そのため, プラス効果は, 週数に対して指数関数的に増加すると仮定し, マイナス効果については, 週数に対して減少すると仮定した.

結果

シミュレーションの結果 (参考資料 3), 伝統的な教育とオンライン教育の間で明確な違いが観測された. 具体的には, オンライン教育の方が初めは徐々に学習レベルが向上し, デジタル操作が軽減されていく後半になり急激に向上する傾向が予測された. 一方で, 伝統的な教育では学習レベルの向上が緩やかだが, 安定していた. これにより, オンライン教育は, 即時的な効果が認められない場合も, 長期的には有用である可能性が高いと考えられる.

考察

シミュレーションによって得られた結果より, オンライン教育は, 即時的な効果が認められない場合も, 長期的には有用である可能性が高いことが推察された. そのため, 活用直後の効果に一喜一憂することなく, デジタル教育を成功させるためには, 導入段階で情報リテラシーを計画的に指導していくカリキュラムを立てるなど, 総合的な教育プログラムの設計や評価計画の必要性が示唆された.

しかし, 本研究はあくまでシミュレーションに基づいているため, 実データによる効果については検証が必要である. 今後は, 本研究のシミュレーションモデルを基にした実証的な研究が求められる.

参考文献

大学生協 (2021) , 第 56 回学生生活実態調査
大学生協 (2023) , 第 58 回学生生活実態調査
濱田竜郎, (声) 在宅のオンライン授業に孤独感, 朝日新聞, 2021 年 10 月 10 日
加納寛子 (2022) , 新型コロナウイルス (COVID-19) 流行時におけるインターネット上の大学や遠隔授業に対する人々の言説分析, 日本情報教育学会, 4(1) 1-11.
竹野 英敏, 藤田 和幸 (2015) , 脳波からみたﾃﾞｼﾞﾀﾙﾍﾟﾝ, ﾀﾌﾞﾚｯﾄ PC 及び紙と鉛筆を用いた記述回答時のストレス比較, 日本科学教育学会研究会研究報告, 30(8), 53-58.

報告集原稿：講演タイトル『逐次的ベイズ予測統合』

桝田 陸, 入江 薫

November 16, 2023

Masuda and Irie 2023 に基づいた今回の発表「逐次的ベイズ予測統合」(2023/11/13, 17:20) では、まずベイズ予測統合 (McAlinn and West 2019) について述べたのち、SMC:Sequential Monte Carlo について簡潔に説明し、実データにおける計算時間と近似精度の評価を見たのちに、LDF:Loss Discounting Framework(Bernaciak and Griffin 2022) への応用を述べた。LDF の枠組みにベイズ予測統合を組み合わせることにより、より柔軟かつ性能の高い予測統合を実行することができ、そのためには SMC による計算負荷の軽減が必要であることを述べた。McAlinn and West 2019 で利用される Dynamic Linear Model 型のベイズ予測統合では、解析的な計算を利用することによって SMC の効率化 (Rao-Blackwellization) が可能であり、これにより通常の MCMC と比較して遥かに短い時間で計算を実行することが可能となった。

以下、会場でいただいた質問とその回答についてまとめる。

LDF を使うとコロナ禍のデータの変動が補足できるのか

コロナ禍でのデータの変動が補足できているというよりは、不確実性の評価が適切にできるようになっている。その点から、予測分布の良さを測る指標である Log Predictive Density Ratio が他のモデルよりも高くなっている。

MCMC 介入が多くなるようなデータは何かあるか

予測が強く外れるような場合に MCMC 介入が入ることになるが、本質的に予測が困難なデータであればそもそも不確実性の高い予測分布が得られるためそこまで大外しするということはないはずであり、そのようなデータはすぐには思いつかない、という回答を会場では返した。今まで安定的な変動をしていたのに、急激な変化が唐突に起こるような場合に MCMC 介入が発生するということになるが、このようなデータとしては fused lasso のような手法が適切になるようなデータが考えられるか。

データがどういう動き方をすると SMC の近似が悪くなるのか

前項とかぶるが、今まで安定的な変動をしていたのに、急激な変化が唐突に起こるような場合に MCMC 介入が発生するということになる。

References

McAlinn, Kenichiro and Mike West (2019). “Dynamic Bayesian predictive synthesis in time series forecasting”. In: *Journal of econometrics* 210.1, pp. 155–169.

Bernaciak, Dawid and Jim E Griffin (2022). “A loss discounting framework for model averaging and selection in time series models”. In: *arXiv preprint arXiv:2201.12045*.

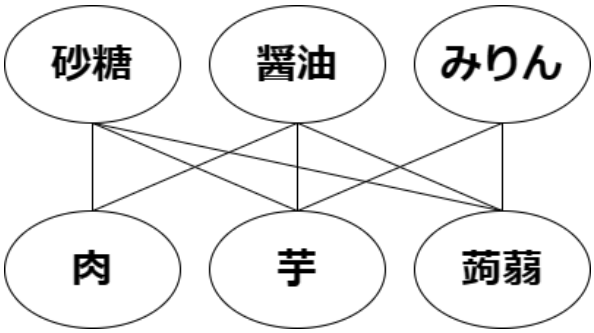
Masuda, Riku and Kaoru Irie (2023). “Sequential Bayesian Predictive Synthesis”. In: *arXiv preprint arXiv:2308.15910*.

LiNGAM-MMI を用いた料理レシピの因果探索

外城稔雄(星光 PMC)、鈴木譲(大阪大)

料理レシピ

料理のレシピは食材と調味料の複雑な因果関係によって成立しており、より美味しくより簡単になるようにレシピの最適化がなされたものと考えることができる。料理の因果の一例として肉じゃがの例を示す。調味料は食材に影響を与えているはずであるが、どの食材に影響を与えているかのか？そしてどの順序が最適なのか？因果探索によってこの因果関係を明らかにするのが本研究の目的である。



LiNGAM-MMI

料理のレシピは食材間の複雑な交絡があると推測される。LiNGAM は交絡が無い事を前提としているが、近年になって交絡の存在を許容する LiNGAM の一般化の枠組みおよびその最適解を求める方法が提案されている[1]。オリジナルの LiNGAM では、真のモデルの交絡が 0(上記の一方の値が 0)であることが仮定されているが、この手法にはその制約はない。一般に変数がm個あって、ある順序を仮定した場合の雑音が$e_1,\ldots,e_m$である場合には、交絡の大きさを下式で定量化できる。交絡の大きさが最小になる順序から因果の序列を得ることが出来る。

$$\mathbb{E}\left[\log\frac{P(e_1,\ldots,e_m)}{P(e_1)\cdots P(e_m)}\right]$$

LiNGAM は３変数の以上の順序を決める場合全ての変数のペアについて独立性を検定していくが、LiNGAM-MMI は３変数間について独立性の評価を行う。3 変数の相互情報量 I を用いて、順序グラフの最短経路を求めることで交絡の最小化を行う。

$$I(e_1,e_2,e_3)=I(e_1,\{e_2,e_3\})+I(e_2,e_3)$$

データセットへの適用戦略

インターネットや書籍で人々に共有だれている料理レシピはより少ない手順でより良い結果（=美味しい）が得られるように、多くの人々によって手順を最適化されているものと考える。この手順の最適化は、交絡の最小化によって成されているという仮説を立てた。 この仮説が正しいと仮定すると、交絡の最小化を行えば料理の最適な手順が得られるはずである。

本研究では、国立情報学研究所の IDR データセット提供サービスによりクックパッド株式会社様から提供を受けた「クックパッドデータセット」を利用した[2]。料理の分量を濃度に変換し確率として取り扱っている。 前述した通り料理レシピは交絡には多くの交絡が含まれると考えられるので、LiNGAM-MMI を用いて因果探索を行った。「クックパッドデータセット」には多くの料理レシピが含ま

れる。 今回を代表的な和食のレシピである肉じゃがのレシピを使用した。 同じレシピ（肉じゃが）の中で「時短レシピ」と「手順を重視したレシピ」に分類を行った。 この二つのデータセットに対して LiNGAM-MMI を適用した時に分量から手順（即ち因果の順序）を推測できるか検証した。

結果

LiNGAM-MMI によって得られた因果の序列を以下に示す。

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
逐次投入	砂糖	だしの素	いも	酒	蒟蒻	人参	めんつゆ	塩	たまねぎ	油	肉	水	だし汁	醤油	みりん
一括投入	砂糖	醤油	酒	油	たまねぎ	人参	めんつゆ	肉	だし汁	みりん	蒟蒻	いも	塩	だしの素	水

この中から調味料の順番のみを抽出した結果を以下に示す。

	逐次投入	一括投入
1	砂糖	砂糖
2	酒	醤油
3	塩	酒
4	水	みりん
5	醤油	塩
6	みりん	水

和食のノウハウの順序

1. 酒
2. さとう
3. しお
4. す（酢）
5. せうゆ（醤油）
6. みそ
7. みりん

逐次投入の因果の順序

1. さとう
2. 酒
3. しお
4. せうゆ（醤油）
5. みりん

和食には調味料に投入順序について一般的に知られた調味料のさしすせそと呼ばれるノウハウがある。酒は臭み取りが目的なので、一番最初が良く、砂糖は塩に比べると分子の大きさが大きいので浸透が遅いため早めににいれるべきであり、醤油と味噌は加熱しすぎると風味が飛ぶので順番は最後の方、みりんは料理に光沢を付与するのが目的なので一番最後に入れるべきと言われている。実際にレシピを確認すると調味料を逐次投入するレシピの多くはこのノウハウを取り入れていた。

PairWise 法を用いて順序の一致率を評価した結果をこちらに示す。 LiNGAM-MMI によって得られた調味料の順番は、逐次投入レシピはノウハウとの一致率が高く、一括投入のレシピはあまり一致していないことが分かった。 交絡を許容する LiNGAM-MMI を用いると料理のノウハウを可視化出来る事が分かった。

比較	順序一致率
ノウハウ vs 逐次投入	80%
ノウハウ vs 一括投入	50%

まとめ

料理レシピのような交絡の多いデータセットに対しても LiNGAM-MMI が有用であることが分かった。交絡を許容する LiNGAM-MMI を用いれば、ノウハウのような交絡が多い事象を可視化できる可能性がある。

Fused Lasso に対する PSI を用いたモデル選択の新提案

大阪大学大学院基礎工学研究科
○田坂理英子 新村亮介 鈴木譲

2023 年 11 月 13 日

1 はじめに

スパース推定の一種で，Generalized Lasso と呼ばれる方法群が知られている．目的関数

$$\underset{\mu\in\mu R^n}{\operatorname{argmin}} \frac{1}{2}\|y-\mu\|_2^2+\lambda\|D\mu\|_1, \tag{1}$$

を最小にする (ただし $\lambda>0$)．この行列 $D\in\mathbb{R}^{m\times p}, m<p$ によって，l_1 ノルム項の制約が決定される．仮説検定など従来の統計的推測において，仮説やモデルはあらかじめ与えられるものと仮定されている．そのため，データからモデルを推定し，再度同一のデータを用いて統計的推測を行う際，従来の検定ではバイアスが生じる．

一方，Post-Selective Inference (PSI) という，パラメータに関する検定をモデル選択に基づき行う方法が近年盛んに研究されている (Tibshirani *et al*, 2016; Lee *et al*, 2016)．本稿では PSI の中で特に Tibshirani *et al.*(2016) で提案された Forward Stepwise に対する TG (Truncated Gaussian) Test という検定方法に着目する．TG Test は線形回帰の場合における Forward Stepwise に対してのみ行われてきたが，本研究ではその結果を拡張して，Fused Lasso(Tibshirani *et al.*, 2004) に適用する．Test はステップ k においてモデルに追加された変数 j_k に対する β_{j_k} について 0 か否かを検定するため，Fused Lasso に対するモデル選択が可能になる．したがって，検定の立場からモデルを決定することが可能になることは，Fused Lasso におけるモデル選択および PSI の双方に対する貢献となると考えた．

1.1 表記及び仮定

本稿内での表記においてサンプルサイズは $n\geq 1$ ，変数の数は $p\geq 1$ と示す．観測データ $y\in\mathbb{R}^n$ は，未知の定数 $\mu\in\mathbb{R}^n$ に，平均がゼロで既知の分散 σ^2 を持つ正規誤差 $\epsilon\in\mathbb{R}^n$ を加えることによって得られるものとする．

$$y=\mu+\epsilon,\quad \epsilon\sim N(0,\sigma^2 I) \tag{2}$$

2 Fused Lasso

Fused Lasso (Tibshirani *et al.*, 2004) を考える。観測値 $y\in\mathbb{R}^n$ と正の定数 $\lambda>0$ が与えられた場合、Fused Lasso は次式を最小化する $\mu\in\mathbb{R}^n$ を見つける．

$$\underset{\mu\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}\|y-\mu\|_2^2+\lambda\|D\mu\|_1.$$

ただし $D\in\mathbb{R}^{(n-1)\times n}$ は次のような行列である．

$$D=\begin{pmatrix} -1 & 1 & 0 & \dots & 0 & 0\\ 0 & -1 & 1 & \dots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix}. \tag{3}$$

3 FS for fused lasso

先行研究と同様に fused lasso での条件付け $\hat{\mathcal{M}}(y)=\hat{\mathcal{M}}(y^{obs})$ が $\{Ay\leq b\}$ の形をしており，検定統計量が正規分布に従う限り検定を構成することができる (Lee *et al.*(2016))．Fused Lasso のための TG 検定を実現するために，Forward Stepwise を考える必要がある．Tibshirani and Taylor(2011) を用いて変形し, 線形回帰の形と同様の式を導くことで以下のような多面体を得ることが出来た。詳細は予稿を参照のこと．$j\notin\mathcal{M}_k$ に対して，

$$\left(\frac{X_{1,j}^\top(I-P_2)_j^\top P_{k-1}^\perp}{\|X_{1,j}^\top(I-P_2)_j^\top P_{k-1}^\perp\|_2}-s_k\frac{X_{1,j_k}^\top(I-P_2)_{j_k}^\top P_{k-1}^\perp}{\|X_{1,j_k}^\top(I-P_2)_{j_k}^\top P_{k-1}^\perp\|_2}\right)(I-P_2)y\leq 0 \tag{4}$$

であるから，多面体が構成できる．この手順を第 m ステップまで行い，不等式の両辺の行を $k=1,\dots,m$ で足し合わせると $2pm-m^2-m$ より，$Ay\leq b$ なる $A\in\mathbb{R}^{(2pm-m^2-m)}, b\in\mathbb{R}^{2pm-m^2-m}$ が得られる．

検定統計量についても, $\tilde{\eta} := (I - P_2)X_1(X_1^\top (I - P_2)^\top (I - P_2)X_1)^{-1}e_j$ としたとき，帰無仮説 $H_0 : \tilde{\eta}^\top \tilde{\mu} = 0$($\tilde{\mu}$ は $(I - P_2)y$ の母平均) とする．このとき，Forward Stepwise によるモデル選択を行った際の多面体は 2.5 節より (4) から構成される．したがってこの下で，Polyhedral Lemma より帰無仮説 $H_0 : \tilde{\eta}^\top \tilde{\mu} = 0$ の下で

$$F_{\mu}^{\mathrm{FL}}(x) := F_{\tilde{\eta}^\top \mu, \tilde{\eta}^\top \tilde{\eta}\sigma^2}^{[\nu^-(z), \nu^+(z)]}(\tilde{\eta}^\top (I - P_2)y | Ay \leq b) \sim U[0, 1] \tag{5}$$

が成り立つ．したがって，有意水準 α の下で

$$\frac{\alpha}{2} \leq F_{\theta_{1,j}}(\tilde{\eta}^\top (I - P_2)y^{\mathrm{obs}}) \leq 1 - \frac{\alpha}{2}. \tag{6}$$

Trend Filtering の場合をはじめとして，ほかの D が行フルランクの場合においては，変数を追加するだけでなく除外するステップが必要となるため，同様の式変形は可能なものの Forward Stepwise をそのまま適用することができない．この点については詳細を当日報告する．

4 数値実験

本研究の提案方法について検出力を見るため Tibshirani *et al.* (2016) と同様の数値実験を行った．この先行研究では，β_1, β_2 の真値がそれぞれ 5 と-5 であり、他がすべて 0 というデータを発生させていた．(**??**) で，Fused Lasso の隣接項に差があることが，線形 lasso の値が非ゼロであることと同値となるように変形を行ったことから，$y_1, \ldots, y_{20}, y_{21}, \ldots, y_{40}, y_{41}, \ldots, y_{60}$ の真値 $\mu_1, \ldots, \mu_{20}, \mu_{21}, \ldots, \mu_{40}, \mu_{41}, \ldots, \mu_{60}$，について，それぞれ差が 5 となるようにした．また，データの形状が凸凹状の時と、一様に増加していく時で検出力に差が出る可能性を考え、2 通りのデータを発生させることとした．詳細な設定と結果は予稿を参照のこと．

真の変化点の数は 2 つであるため，この数値実験ではステップ 1 および 2 では帰無仮説を棄却し，ステップ 3 では p 値は $U[0, 1]$ の帰無分布に従うことが望ましい．実際，ステップ 1，ステップ 2 で棄却し，ステップ 3 ではおおむね帰無仮説に従っているのが見て取れた．

4.1 まとめ

本研究では，まず，Generalized Lasso の目的関数を [2] により Lasso と同様の形に変形することで，Fused Lasso について，線形回帰の場合と同様に Forward Stepwise アルゴリズムで解くことができることを示した．次に，モデル選択の条件付けを $\{Ay \leq b\}$ で書き表し，検定統計量を構成した．

一方，D が行 Full Rank である場合においても、Trend filtering のように解が追加されるのみの場合には適用が難しい．また，D が行 Full Rank であるというのは一般に成り立たないため，すべての Generalized Lasso で同様に TG Test は構成できないという問題がある．したがって，今後は双対アルゴリズムをもとに線形多面体を構成し，検定を行うことを検討していきたい．

参考文献

D.R.Cox. A note on data-splitting for the evaluation of significance levels. Biometrika, 62:441 – 444, 1975. doi: 10.1093/biomet/62.2.441.

W. Fithian, D. L. Sun, and J. Taylor. Optimal inference after model selection. Arxiv, 2014. doi: arXiv:1410.2597v4.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. The Annals of Statistics, 44(3):907 – 927, 2016. doi: 10.1214/15-AOS1371.

R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B, 58(1):267 – 288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society Series B, 67(1):91 – 108, 2004. doi: 10.1111/j.1467-9868.2005.00490.x.

R. J. Tibshirani, J. E. Taylor, and R. Lockhart. Exact post selection inference for sequential regression procedures. Journal of the American Statistical Association, 111:600 – 620, 2016. doi: 10.1080/01621459.2015.1108848.

R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. The Annals of Statistics, 39(3):1335 – 1371, 2011. doi: 10.1214/11-AOS878.

Further analysis on LASSO-type estimator in nonparametric regression

千葉大・融合理工学府　松本 美幸
千葉大・理学研究院　内藤 貫太

設定：目的変数 $Y_i \in \mathbb{R}$ と説明変数 $\boldsymbol{X}_i \in \mathbb{R}^d$ に対してノンパラメトリック回帰モデル

$$Y_i = f(\boldsymbol{X}_i) + e_i, i = 1, \ldots, n \tag{1}$$

を用いて，ある推定点 $\boldsymbol{x} \in \mathbb{R}^d$ における $f(\boldsymbol{x})$ を推定したい．このために，局所線形回帰における局所最小 2 乗法の流れから，(1) を推定点 $\boldsymbol{x}$ における線形回帰モデル

$$\boldsymbol{Z}_{\boldsymbol{x}} = A_{\boldsymbol{x}}\boldsymbol{\theta}^*(\boldsymbol{x}) + \boldsymbol{\varepsilon}_{\boldsymbol{x}}$$

に落とし込む，ここで，

$$\boldsymbol{Z}_{\boldsymbol{x}} = n^{-1/2}W_x^{1/2}\boldsymbol{Y},\ A_{\boldsymbol{x}} = n^{-1/2}W_x^{1/2}\begin{pmatrix} 1 & (\boldsymbol{X}_1 - \boldsymbol{x})^T/h \\ \vdots & \vdots \\ 1 & (\boldsymbol{X}_n - \boldsymbol{x})^T/h \end{pmatrix},\ \boldsymbol{\varepsilon}_{\boldsymbol{x}} = \boldsymbol{Z}_{\boldsymbol{x}} - A_x\boldsymbol{\theta}^*(\boldsymbol{x}),$$

$W_{\boldsymbol{x}} = \mathrm{diag}\left\{K_h\left(\boldsymbol{X}_1 - \boldsymbol{x}\right), \ldots, K_h\left(\boldsymbol{X}_n - \boldsymbol{x}\right)\right\}, \boldsymbol{Y} = \left(Y_1 \cdots Y_n\right)^T$ とする．また，真のパラメータを $\boldsymbol{\theta}^*(\boldsymbol{x}) = \left(\theta_0^*(\boldsymbol{x})\ \theta_1^*(\boldsymbol{x})\ \cdots\ \theta_d^*(\boldsymbol{x})\right)^T = \left(f(\boldsymbol{x})\ h\partial_1 f(\boldsymbol{x})\ \cdots\ h\partial_d f(\boldsymbol{x})\right)^T$ とし，∂_i は第 i 成分での偏微分を表す．$K_h(\boldsymbol{x}) = K\left(\boldsymbol{x}/h\right)/h^d$ であり，K は原点を中心に対称な d 次元密度である核関数，$h > 0$ はバンド幅である．

問題：推定点 $\boldsymbol{x}$ における $\theta_0^*(\boldsymbol{x}) = f(\boldsymbol{x})$ の推定と変数選択を実行する際，従来の LASSO(Bertin and Lecué (2008)) による

$$\widetilde{\boldsymbol{\theta}}(\boldsymbol{x}) = \underset{\boldsymbol{\theta}\in\mathbb{R}^{d+1}}{\arg\min}\left\{\left\|\boldsymbol{Z}_{\boldsymbol{x}} - A_{\boldsymbol{x}}\boldsymbol{\theta}\right\|_2^2 + 2\lambda\left\|\boldsymbol{\theta}\right\|_1\right\} \tag{2}$$

を用いると，$\boldsymbol{\theta} = \left(\theta_0\ \theta_1\ \cdots\ \theta_d\right)^T$ において $f(\boldsymbol{x})$ に対応する切片項 θ_0 が正則化項に入っていることから，関数推定の精度が悪くなる可能性があった．この問題を解決するため，Matsushima and Naito (2022) では

$$\widehat{\boldsymbol{\theta}}(\boldsymbol{x}) = \underset{\boldsymbol{\theta}\in\mathbb{R}^{d+1}}{\arg\min}\left\{\left\|\boldsymbol{Z}_{\boldsymbol{x}} - A_{\boldsymbol{x}}\boldsymbol{\theta}\right\|_2^2 + 2\lambda\left\|\boldsymbol{\theta}_{-0}\right\|_1\right\} \tag{3}$$

を提案している．ここで，$\boldsymbol{a} = \left(a_0\ a_1\ \cdots\ a_d\right)^T$ について $\boldsymbol{a}_{-0} = \left(a_1\ \cdots\ a_d\right)^T$ と表すことにする．本講演では (2) の $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}(\boldsymbol{x})$ および (3) の $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ に関し，いくつかの観点から，これらの理論的比較を行った．

理想的推定量：整数 $d^* \leq d$，関数 $g : \mathbb{R}^{d^*} \to \mathbb{R}$，濃度 d^* の部分集合 $J = \{i_1, \ldots, i_{d^*}\} \subset \{1, \ldots, d\}$ が存在して，$(x_1, \ldots, x_d) \in \mathbb{R}^d$ に対して

$$f\left(x_1, \ldots, x_d\right) = g\left(x_{i_1}, \ldots, x_{i_{d^*}}\right)$$

を満たすことを仮定する．d^* を用いて，

$$\boldsymbol{\theta}^*(\boldsymbol{x}) = \begin{pmatrix}\theta_0^*(\boldsymbol{x}) & \boldsymbol{\theta}_{(1)}^*(\boldsymbol{x})^T & \boldsymbol{\theta}_{(2)}^*(\boldsymbol{x})^T\end{pmatrix}^T,\ \theta_0^*(\boldsymbol{x}) \in \mathbb{R},\ \boldsymbol{\theta}_{(1)}^*(\boldsymbol{x}) \in \mathbb{R}^{d^*},\ \boldsymbol{\theta}_{(2)}^*(\boldsymbol{x}) \in \mathbb{R}^{d-d^*},$$
$$A_{\boldsymbol{x}} = \begin{pmatrix}\boldsymbol{A}_{\boldsymbol{x}(0)} & A_{\boldsymbol{x}(1)} & A_{\boldsymbol{x}(2)}\end{pmatrix},\ \boldsymbol{A}_{\boldsymbol{x}(0)} \in \mathbb{R}^n,\ A_{\boldsymbol{x}(1)} \in \mathbb{R}^{n\times d^*},\ A_{\boldsymbol{x}(2)} \in \mathbb{R}^{n\times(d-d^*)}$$

と分割して表す．同様に $\widetilde{\boldsymbol{\theta}}(\boldsymbol{x}) = \left(\widetilde{\theta}_0(\boldsymbol{x}) \quad \widetilde{\boldsymbol{\theta}}_{(1)}(\boldsymbol{x})^T \quad \widetilde{\boldsymbol{\theta}}_{(2)}(\boldsymbol{x})^T\right)^T$, $\widehat{\boldsymbol{\theta}}(\boldsymbol{x}) = \left(\widehat{\theta}_0(\boldsymbol{x}) \quad \widehat{\boldsymbol{\theta}}_{(1)}(\boldsymbol{x})^T \quad \widehat{\boldsymbol{\theta}}_{(2)}(\boldsymbol{x})^T\right)^T$ と分割する．理想的推定量 $\widetilde{\boldsymbol{\eta}}(\boldsymbol{x}), \widehat{\boldsymbol{\eta}}(\boldsymbol{x})$ を

$$\widetilde{\boldsymbol{\eta}}(\boldsymbol{x}) = \boldsymbol{\theta}^*(\boldsymbol{x}) + \begin{pmatrix} \Psi_{11}^{-1}\left\{\left(\boldsymbol{A}_{\boldsymbol{x}(0)} \ A_{\boldsymbol{x}(1)}\right)^T \boldsymbol{\varepsilon} - \lambda \overrightarrow{\mathrm{sign}}\begin{pmatrix} \theta_0^*(\boldsymbol{x}) \\ \boldsymbol{\theta}_{(1)}^*(\boldsymbol{x}) \end{pmatrix}\right\} \\ \mathbf{0}_{d-d^*} \end{pmatrix},$$

$$\widehat{\boldsymbol{\eta}}(\boldsymbol{x}) = \boldsymbol{\theta}^*(\boldsymbol{x}) + \begin{pmatrix} \Psi_{11}^{-1}\left\{\left(\boldsymbol{A}_{\boldsymbol{x}(0)} \ A_{\boldsymbol{x}(1)}\right)^T \boldsymbol{\varepsilon} - \lambda \overrightarrow{\mathrm{sign}}\begin{pmatrix} 0 \\ \boldsymbol{\theta}_{(1)}^*(\boldsymbol{x}) \end{pmatrix}\right\} \\ \mathbf{0}_{d-d^*} \end{pmatrix}$$

と定義したとき，ある増大事象列 $\mathcal{A}_n$ において

$$\widetilde{\boldsymbol{\theta}}(\boldsymbol{x}) = \widetilde{\boldsymbol{\eta}}(\boldsymbol{x}), \ \widehat{\boldsymbol{\theta}}(\boldsymbol{x}) = \widehat{\boldsymbol{\eta}}(\boldsymbol{x})$$

が成り立つ．ここで $\Psi_{11} = \left(\boldsymbol{A}_{\boldsymbol{x}(0)} \ A_{\boldsymbol{x}(1)}\right)^T \left(\boldsymbol{A}_{\boldsymbol{x}(0)} \ A_{\boldsymbol{x}(1)}\right)$ とする (Zhao and Yu (2006) を参照)．この理想的推定量を用いて，次の性質を示すことができる．

関数推定の精度：推定点 $\boldsymbol{x}$ における $f(\boldsymbol{x})$ を推定するという目的から，$\widetilde{\boldsymbol{\theta}}$ と $\widehat{\boldsymbol{\theta}}$ のどちらの方が良く $\theta_0^*(\boldsymbol{x}) = f(\boldsymbol{x})$ を推定できているか比較することは有用である．ある増大事象列 $\mathcal{B}_n \subset \mathcal{A}_n$ において

$$P\left(\left\{\left|\widehat{\theta}_0(\boldsymbol{x}) - \theta_0^*(\boldsymbol{x})\right| < \left|\widetilde{\theta}_0(\boldsymbol{x}) - \theta_0^*(\boldsymbol{x})\right|\right\} \cap \mathcal{B}_n\right) > P\left(\left\{\left|\widehat{\theta}_0(\boldsymbol{x}) - \theta_0^*(\boldsymbol{x})\right| > \left|\widetilde{\theta}_0(\boldsymbol{x}) - \theta_0^*(\boldsymbol{x})\right|\right\} \cap \mathcal{B}_n\right)$$

が成り立つことを示すことができる．

符号一致性：LASSO による変数選択の一致性や符号一致性に関しては，irrepresentable condition と呼ばれる条件を満たしている場合に成り立つことが知られている (Zhao and Yu (2006) 参照)．ここでは，どういった条件のもとで $\widetilde{\boldsymbol{\theta}}$ と $\widehat{\boldsymbol{\theta}}$ それぞれの符号一致性の確率，即ち，事象 $\left\{\overrightarrow{\mathrm{sign}}\left(\widetilde{\boldsymbol{\theta}}_{-0}(\boldsymbol{x})\right) = \overrightarrow{\mathrm{sign}}\left(\boldsymbol{\theta}_{-0}^*(\boldsymbol{x})\right)\right\}$ と事象 $\left\{\overrightarrow{\mathrm{sign}}\left(\widehat{\boldsymbol{\theta}}_{-0}(\boldsymbol{x})\right) = \overrightarrow{\mathrm{sign}}\left(\boldsymbol{\theta}_{-0}^*(\boldsymbol{x})\right)\right\}$ の確率に差が現れるのかについて考えた．特に，ある事象列 $\mathcal{C}_n \subset \mathcal{A}_n$ において

$$P\left(\left\{\overrightarrow{\mathrm{sign}}\left(\widehat{\boldsymbol{\theta}}_{-0}(\boldsymbol{x})\right) = \overrightarrow{\mathrm{sign}}\left(\boldsymbol{\theta}_{-0}^*(\boldsymbol{x})\right)\right\} \cap \mathcal{C}_n\right) > P\left(\left\{\overrightarrow{\mathrm{sign}}\left(\widetilde{\boldsymbol{\theta}}_{-0}(\boldsymbol{x})\right) = \overrightarrow{\mathrm{sign}}\left(\boldsymbol{\theta}_{-0}^*(\boldsymbol{x})\right)\right\} \cap \mathcal{C}_n\right)$$

が成り立つことを示すことができる．

適用例：シミュレーション実験の結果および実データへの適用については，シンポジウムにて報告した．

参考文献

Bertin, K. and Lecué, G. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics*, **2**:1224–1241.

Matsushima, Y. and Naito, K. (2022). Improvement on lasso-type estimator in nonparametric regression. *Journal of Nonparametric Statistics*, **34**(4):964–986.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, **7**(90):2541–2563.

Semiparametric regression with localized Bregman divergence

千葉大・融合理工学府　小杉 拓生
千葉大・理学研究院　内藤 貫太

はじめに：重回帰分析における局所推定量の漸近的性質に関し議論する．Bregman ダイバージェンスの局所的最小化により回帰推定量を構成する．Bregman ダイバージェンスを構成する狭義凸関数とモデルを構成するリンク関数のみを与えた一般的設定において，回帰推定量の漸近的性質を導出する．本講演では特に，提案推定量の漸近分布を導出するとともに，推定量のダイバージェンス・リスクに基づく性能評価について報告した．

設定：n 組のデータ $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n) \sim_{i.i.d.} f(y, \boldsymbol{x}) = p(y|\boldsymbol{x}) \cdot q(\boldsymbol{x})$ に基づき，回帰関数 $\mu(\boldsymbol{t}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{t}]$ の推定を行う．ただし，$(Y_i, \boldsymbol{X}_i) \in \mathbb{R} \times \mathbb{R}^d$ であり，f を $(Y, \boldsymbol{X})$ の同時密度，$p(\cdot|\boldsymbol{x})$ を $\boldsymbol{X} = \boldsymbol{x}$ が与えられたときの Y の条件付き密度，q を $\boldsymbol{X}$ の密度とする．

一般化線形回帰の流れから，リンク関数 G を用いて，

$$\eta(\boldsymbol{t}) = G(\mu(\boldsymbol{t})) = G(\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{t}])$$

と定義する．

推定点 $\boldsymbol{t}$ における局所化したモデルを，未知パラメータ $\boldsymbol{\theta}$ を用いて，

$$m(\boldsymbol{x}(\boldsymbol{t}), \boldsymbol{\theta}) = G^{-1}\left(\boldsymbol{\theta}^\top \widetilde{\boldsymbol{x}(\boldsymbol{t})}\right) \tag{1}$$

とする．ただし，$\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{d+1}$ であり，$\boldsymbol{x}(\boldsymbol{t}) = \boldsymbol{x} - \boldsymbol{t}$，$\widetilde{\boldsymbol{x}(\boldsymbol{t})} = [1\ \boldsymbol{x}(\boldsymbol{t})^T]^T$ である．

$\eta(\boldsymbol{X}_i)$ はテイラー展開より，

$$\eta(\boldsymbol{X}_i) \approx \eta(\boldsymbol{t}) + (\boldsymbol{X}_i - \boldsymbol{t})^T \frac{\partial}{\partial \boldsymbol{t}} \eta(\boldsymbol{t}) = \left[\eta(\boldsymbol{t}) \quad \frac{\partial}{\partial \boldsymbol{t}^T} \eta(\boldsymbol{t})\right] \widetilde{\boldsymbol{X}_i(\boldsymbol{t})}$$

のように近似できるので，推定点 $\boldsymbol{t}$ における真のパラメータ $\boldsymbol{\theta}^*(\boldsymbol{t})$ は，

$$\boldsymbol{\theta}^*(\boldsymbol{t}) = \begin{bmatrix} \theta_0^*(\boldsymbol{t}) \\ \theta_1^*(\boldsymbol{t}) \\ \vdots \\ \theta_d^*(\boldsymbol{t}) \end{bmatrix} = \begin{bmatrix} \eta(\boldsymbol{t}) \\ \\ \dfrac{\partial}{\partial \boldsymbol{t}} \eta(\boldsymbol{t}) \end{bmatrix} \tag{2}$$

と定義される．

μ とその一般的な推定量 $\bar{\mu}$ の functional Bregman ダイバージェンスは，狭義凸関数 $U : \mathbb{R} \to \mathbb{R}$ と $u = U'$，U の凸共役 U^* を用いて，

$$\begin{aligned} D_U[\mu, \bar{\mu}] &= \int_{\mathbb{R}^d} \left[U(\mu(\boldsymbol{x})) - U(\bar{\mu}(\boldsymbol{x})) - u(\bar{\mu}(\boldsymbol{x}))\{\mu(\boldsymbol{x}) - \bar{\mu}(\boldsymbol{x})\}\right] q(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \int_{\mathbb{R} \times \mathbb{R}^d} \left\{U^*(u(\bar{\mu}(\boldsymbol{x}))) - y \cdot u(\bar{\mu}(\boldsymbol{x}))\right\} f(y, \boldsymbol{x})\, dy d\boldsymbol{x} + Const. \end{aligned}$$

と定義される．これを用いて，推定点 $\boldsymbol{t}$ における局所ダイバージェンスは，核関数を組み込むことにより，

$$D_U[\mu, \bar{\mu} | \boldsymbol{t}] = \int_{\mathbb{R} \times \mathbb{R}^d} K_H(\boldsymbol{x}(\boldsymbol{t})) \left\{U^*(u(\bar{\mu}(\boldsymbol{x}))) - y \cdot u(\bar{\mu}(\boldsymbol{x}))\right\} f(y, \boldsymbol{x})\, dy d\boldsymbol{x} + Const. \tag{3}$$

と定義される．ここで，$K_H(\boldsymbol{z}) = |H|^{-1/2} K\left(H^{-1/2}\boldsymbol{z}\right)$，$K \geq 0$ は原点対称な d 次元核関数，H は正定値対称 $d \times d$ バンド幅行列である．(1) と (3) より，推定点 $\boldsymbol{t}$ における $\boldsymbol{\theta}$ の局所推定量 $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ は，

$$\hat{\boldsymbol{\theta}}(\boldsymbol{t}) = \arg\min_{\theta \in \Theta} \int_{\mathbb{R}\times\mathbb{R}^d} K_H\left(\boldsymbol{x}\left(\boldsymbol{t}\right)\right) \rho\left(y, \boldsymbol{x}\left(\boldsymbol{t}\right), \boldsymbol{\theta}\right) dF_n\left(y, \boldsymbol{x}\right)$$

で定義される．ただし，$\rho\left(y, \boldsymbol{x}, \boldsymbol{\theta}\right) = U^*\left(u\left(m\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)\right) - y \cdot u\left(m\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)$ であり，F_n は経験分布である．(2) より，$\mu(\boldsymbol{t})$ に関する提案推定量 $\hat{\mu}(\boldsymbol{t})$ は局所推定量 $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ を用いて，

$$\hat{\mu}(\boldsymbol{t}) = G^{-1}(\boldsymbol{e}_{d+1,1}^T \hat{\boldsymbol{\theta}}(\boldsymbol{t}))$$

で構成される．ただし，$\boldsymbol{e}_{k,i} \in \mathbb{R}^k$ を第 i 成分のみが 1 の単位ベクトルとする．推定方法を決定する Bregman ダイバージェンスを構成する狭義凸関数 U と，モデルを構成するリンク関数 G のみを与えた一般的設定は，局所多項式回帰の場合 ($d = 1$) において Zhang et al. (2009) で議論されている．本講演では特に，提案推定量 $\hat{\mu}(\boldsymbol{t})$ の漸近分布を $d \geq 2$ の場合で導出するとともに，多項式重回帰に拡張した結果についても紹介した．

問題：局所線形推定量 (Wand and Jones (1995),Ruppert and Wand (1994)) は提案推定量の特別な場合であり，狭義凸関数 $U(x) = x^2/2$, リンク関数 $G(x) = x$ とした場合に対応している．狭義凸関数 U とリンク関数 G により，どの程度豊かな推定量が得られるのか？そして，その性質，挙動について解明する．例えば，局所化のもとで，どのような U と G がロバストな推定量を導くかについて考察を与えた．

ダイバージェンス・リスク：狭義凸関数 U_1 とリンク関数 G_1 を用いて構成される提案推定量 $\hat{\mu} = \hat{\mu}(\cdot \mid U_1, G_1)$ の精度は，狭義凸関数 U によるダイバージェンス・リスク (Naito and Penev (2021))

$$\mathcal{R}_U(\hat{\mu}) = \mathbb{E}\left[D_U\left[\mu, \hat{\mu}\right]\right]$$

で評価される．ここで，U は U_1 と異なってもよい．$n \to \infty, H \to O$ のとき，

$$\mathcal{R}_U(\hat{\mu}) = \frac{1}{2}\int_{\mathbb{R}^d} \left[\frac{\mu_2^2}{4}\left\{\frac{\mathrm{tr}\left[\mathcal{H}_{\eta_1}(\boldsymbol{x})H\right]}{G_1'(\mu(\boldsymbol{x}))}\right\}^2 + \frac{\nu_0}{n|H|^{1/2}}\frac{v(\boldsymbol{x})}{q(\boldsymbol{x})}\right] U''(\mu(\boldsymbol{x}))q(\boldsymbol{x})d\boldsymbol{x} + o\left(\mathrm{tr}[H]^2 + \frac{1}{n|H|^{1/2}}\right)$$

で与えられる．ただし，$\eta_1(\boldsymbol{x}) = G_1(\mu(\boldsymbol{x}))$ であり,$\mathcal{H}_{\eta_1}(\boldsymbol{x})$ はヘッセ行列である．また，$\nu_0 = \int_{\mathbb{R}^d} K(\boldsymbol{z})^2 d\boldsymbol{z}$, $v(\boldsymbol{x}) = \mathrm{V}\left[Y \mid \boldsymbol{X} = \boldsymbol{x}\right]$ である．

適用例：本講演では，一般化線形回帰の範囲での適用例とダイバージェンス・リスクに基づいた提案推定量の性能比較を行った．特に，局所線形推定量との理論的比較について報告した．シミュレーション実験の結果及び実データへの適用例についても報告した．

参考文献

Naito, K. and Penev, S. (2021). Regression using localised functional Bregman divergence. *Electronic Journal of Statistics*, **15**(2):6544–6585.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, **22**(3):1346 – 1370.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall.

Zhang, C., Jiang, Y., and Shang, Z. (2009). New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation. *Canadian Journal of Statistics*, **37**(1):119–139.

二元分割表における divergence 型尺度を用いた対称性の可視化

浦崎 航[1], 中川 智之[2], 田畑 耕治[3]

[1]東京理科大学大学院 創域理工学研究科
[2]明星大学 データサイエンス学環
[3]東京理科大学 創域理工学部

1 はじめに

r 個のカテゴリと c 個のカテゴリから構成される 2 種類のカテゴリカル変数を考えたとき, それらの組合せから得られる r 行 c 列の表を二元分割表と呼ぶ. 分割表を用いる解析は, 2 つの変数に関係性があるか否か, つまり, 統計的独立性を持つかを評価することを目的に, 様々な手法が提案された. その手法の 1 つとして, カイ二乗統計量を用いた独立性の検定が有名であり, 広く一般的に用いられる. 分割表には行と列が同じ分類である場合が考えられ, そのような分割表を正方分割表と呼ぶ. 正方分割表は観測値が主対角セル付近に集中し, 主対角から離れるにつれて少なくなる傾向があるため, 2 種類の分類間の関連性は極めて高く, 独立性は成り立たない. そのために独立性と代わり, 対称なカテゴリの組合せの釣り合いを評価する対称性の解析が提案されてきた. 対称性の解析手法は大まかに 2 種類で, 具体的な代表例として Bowker [2] の対称モデルをもとにした検定による判断, Tomizawa *et al.* [3] の power-divergence 型尺度 $\Phi^{(\lambda)}$ による 0-1 区間での対称性からの逸脱度評価がある. これらを含めた先行研究による解析手法の提案は, 分割表全体の対称性に焦点を当て, 観測された分割表が持つ対称性の特徴, 対称性からの逸脱度を調べる手段として画期的だが, 全体より各カテゴリの対称性に着目したいときは不適切である.

本研究の目的は, 分割表全体の対称性の調査だけでは解釈が困難な正方分割表に対して, 各カテゴリの対称性の数値化と二次元座標上へ可視化をすることで, 解析結果の解釈性を容易にすることにある. また, power-divergence 型尺度を取り上げることで, power-divergence 内で一般化された距離の測り方でも, 本目的を達成できることを保証した.

2 対称性の可視化

近年の分割表解析は, 独立性, 対称性を問わず, 各カテゴリの評価と二次元座標上への可視化から, 直感的理解の獲得を目指す動きがある. 特に, Beh and Lombardo [1] は, Bowker のカイ二乗統計量 χ^2_S を用いた可視化を提案したが, Pearson 距離の提案にのみ留まる. そのため, power-divergence 型尺度 $\Phi^{(\lambda)}$ を用いることで, 様々な距離での可視化と最適な測り方を提案した.

power-divergece 型尺度 $\Phi^{(\lambda)}$ は, $\lambda > -1$ の範囲で以下のように定義される.

$$\Phi^{(\lambda)} = \sum_{i \neq j} \sum \frac{p^*_{ij} + p^*_{ji}}{2} \left[1 - \frac{\lambda 2^\lambda}{2^\lambda - 1} H^{(\lambda)}_{ij}(\{p^c_{ij}, p^c_{ji}\}) \right] = \sum_{i \neq j} \sum \phi^{(\lambda)}_{ij},$$

ただし, $\delta = \sum\sum_{i\neq j} p_{ij}$ として, $p^*_{ij} = p_{ij}/\delta$, $p^c_{ij} = p_{ij}/(p_{ij}+p_{ji})$ とおき,

$$H^{(\lambda)}_{ij}(\{p^c_{ij}, p^c_{ji}\}) = \frac{1}{\lambda}\left[1-(p^c_{ij})^{\lambda+1}-(p^c_{ji})^{\lambda+1}\right],$$
$$\phi^{(\lambda)}_{ij} = \frac{p^*_{ij}+p^*_{ji}}{2}\left[1-\frac{\lambda 2^\lambda}{2^\lambda-1}H^{(\lambda)}_{ij}(\{p^c_{ij}, p^c_{ji}\})\right],$$

とした. ここで, (i,j) セルの対称性からの逸脱を $\phi^{(\lambda)}_{ij}$ (≥ 0) が表すことに注意して, $\sqrt{\phi^{(\lambda)}_{ij}}$ を (i,j) 成分にもつ歪対称行列 $S_{skew(\lambda)}$ を考えると,

$$\Phi^{(\lambda)} = \sum_{i\neq j}\sum \phi^{(\lambda)}_{ij} = trace(S^T_{skew(\lambda)}S_{skew(\lambda)}),$$

として, 部分的な対称性の情報を持ちながらも $\Phi^{(\lambda)}$ が構成できた. 本研究は, $S_{skew(\lambda)}$ を用いて, 二次元座標上への可視化に必要な座標を導出した. 歪対称行列 $S_{skew(\lambda)}$ に特異値分解を行うと,

$$S_{skew(\lambda)} = AD_\alpha B^T,$$

と分解できた. ここで, A, B はそれぞれ $S_{skew(\lambda)}$ の右, 左特異ベクトルを持つ $S\times M$ 直交行列, D_α は対角成分に降順で特異値 α_m $(m=1,\ldots,M)$ を持つ $M\times M$ 対角行列である. また, M は行数 (または列数)S で変化し, S が偶数ならば $M=S$, 奇数ならば $M=S-1$ となる. これらの行列を用いて座標を得るが, 正方分割表についての座標のスケーリングとして,

$$F = D^{-1/2}AD_\alpha, \quad G = D^{-1/2}BD_\alpha,$$

から, 各行, 列カテゴリごとの対称性からの隔たりを示す座標が得られた. ここで, $D = (D_R + D_C)/2$ であり, D_R, D_C はそれぞれ行と列の周辺確率 $p_{i\cdot}, p_{\cdot j}$ を対角成分に持つ $S\times S$ 対角行列である. 一方で, 導出された M 次元の座標に関する m 番目の座標に対する主軸の寄与率は,

$$\Phi^{(\lambda)} = trace((AD_\alpha B^T)^T(AD_\alpha B^T)) = trace(D^2_\alpha),$$

であることから, $\alpha^2_m/\Phi^{(\lambda)}\times 100\,(\%)$ で求められた.

本講演では, power-divergence 型尺度 $\Phi^{(\lambda)}$ の可視化の様態をシミュレーションデータから観察したとともに, 実データ解析を通して, パラメータ λ を変化時の寄与率の変化やプロットされた各カテゴリの座標の位置から考察される結果の解釈を示した.

参考文献

[1] Eric J Beh and Rosaria Lombardo. Visualising departures from symmetry and bowker's x 2 statistic. *Symmetry*, Vol. 14, No. 6, p. 1103, 2022.

[2] Albert H Bowker. A test for symmetry in contingency tables. *Journal of the american statistical association*, Vol. 43, No. 244, pp. 572–574, 1948.

[3] Sadao Tomizawa, Takashi Seo, and Hideharu Yamamoto. Power-divergence-type measure of departure from symmetry for square contingency tables that have nominal categories. *Journal of Applied Statistics*, Vol. 25, No. 3, pp. 387–398, 1998.

複雑な構造をもつ高次元データに対する変数スクリーニング

田中 俊太郎（滋賀大学大学院データサイエンス研究科, 株式会社日本総合研究所）
松井 秀俊（滋賀大学データサイエンス学部）

１．概要

スクリーニング法は, 説明変数の数がサンプルサイズよりはるかに大きい場合に, 目的変数と関連する説明変数を選択するための有用なツールである．しかし, 説明変数が多重共線性を持つ場合, 適切に変数を選択できない場合が多い．これに対して, 因子分析を用いることで説明変数間の多重共線性を排除でき, 変数選択の性能を向上できる．本発表では, 既存手法に比べてより正確に多重共線性を排除する因子の数を選択する方法を提案する．提案手法は, 因子分析によって得られた情報から不要な部分を切り捨てることで, 変数選択の性能を向上させる．シミュレーションデータと実データを用いた解析により, 提案手法の性能を確認する．

２．変数スクリーニング

n 件のデータ $\{(y_i, \boldsymbol{x}_i);\ i = 1, \ldots, n\}$ が得られたとする．ここで, $y_i \in \mathbb{R}$ は目的変数, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$ は説明変数とする．これらはそれぞれ中心化, 標準化されているとする．y_i と $\boldsymbol{x}_i$ に線形回帰モデルの関係が成り立つと仮定すると, $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ と表せる．ここで, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ は回帰係数, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^T$ は誤差とする．Sure Independence Screening (SIS) [1] は, $\boldsymbol{\omega} = (\omega_1, ..., \omega_p)^T = X^T\boldsymbol{y} \in \mathbb{R}^p$ を計算し, j 番目の変数の重要度を $|\omega_j|$ $(1 \leq j \leq p)$ と定義し, $|\omega_j|$ の大きい順に変数を選択する．SIS は, 選択する変数が真に重要な変数を含む確率は漸近的に 1 に収束するという性質を持つ．しかし, SIS は説明変数が強い多重共線性を持つときには適切なスクリーニングができないことが多い．

３．因子分析を活用した変数スクリーニング

Factor Profiled Sure Independence Screening (FPSIS)[2] は, 因子分析を活用してデータから多重共線性を排除した後に SIS を適用する．X の d $(< n)$ 個の共通因子ベクトルを列としてもつ行列を $Z \in \mathbb{R}^{n\times d}$, 因子負荷行列を $B \in \mathbb{R}^{p\times d}$, 各列が互いに独立な独自因子からなる行列を $\check{X} \in \mathbb{R}^{n\times p}$ とおくと, $X = ZB^T + \check{X}$ と表せる．Z は回転の不定性により一意には定まらないが, $X = UDV^T$ と特異値分解し, U の最初の d 列を U_1 とすると, U_1 は Z の解の 1 つとみなせる（図 1）．U_1 の列ベクトルによって張られる線形部分空間の直交補空間への射影行列

$$Q_F = I_n - U_1(U_1^TU_1)^{-1}U_1^T$$

を線形回帰式の両辺に左から掛けることで, $Q_F\boldsymbol{y} = Q_FX\boldsymbol{\beta} + Q_F\boldsymbol{\varepsilon}$ が得られる．$\hat{\boldsymbol{y}} = Q_F\boldsymbol{y}$, $\hat{X} = Q_FX$ とおくと, $\hat{X}$ は独自因子 $\check{X}$ の近似値となる．$\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^T = \hat{X}^T\hat{\boldsymbol{y}}$ を計算し, $|\omega_j|$ を j 番目の変数の重要度とする．

Preconditioned Profiled Independence Screening (PPIS)[3] は FPSIS のデータ変換処理を改善した方法である．まず, $X = UDV^T$ の形に特異値分解した後, 行列 U, D, V それぞれを d 番目の列を境界として 2 つに分け, $U_1 = (\boldsymbol{u}_1, ..., \boldsymbol{u}_d) \in \mathbb{R}^{n\times d}$, $U_2 = (\boldsymbol{u}_{d+1}, ..., \boldsymbol{u}_n) \in \mathbb{R}^{n\times(n-d)}$, $D_1 = \mathrm{diag}(\mu_1, ..., \mu_d) \in \mathbb{R}^{d\times d}$, $D_2 = \mathrm{diag}(\mu_{d+1}, ..., \mu_n) \in \mathbb{R}^{(n-d)\times(n-d)}$, $V_1 = (\boldsymbol{v}_1, ..., \boldsymbol{v}_d) \in \mathbb{R}^{p\times d}$, $V_2 = (\boldsymbol{v}_{d+1}, ..., \boldsymbol{v}_n) \in \mathbb{R}^{p\times(n-d)}$ とする（図 1）．データ変換行列を

$$Q_P = U_2D_2U_2^T\left\{I_n - U_1(U_1^TU_1)U_1^T\right\}$$

で定め, FPSIS と同様に, $\hat{\boldsymbol{y}} = Q_P\boldsymbol{y}$, $\hat{X} = Q_PX$, $\boldsymbol{\omega} = (\omega_1, ..., \omega_p)^T = \hat{X}^T\hat{\boldsymbol{y}}$ として, $|\omega_j|$ を j 番目の変数の重要度とする．データから d 次元分の共通因子に関連する部分の影響を取り除くだけでは多重共線性が残ってしまうため, FPSIS で用いられていなかった $d+1$ 次元以降の情報も含めたデータ変換行列を用いることで, $\hat{X}$ が独自因子 $\check{X}$ により近い値となり, 多重共線性を取り除く効果が高まる．

4．提案手法

提案手法 Truncated Preconditioned Profiled Independence Screening (TPPIS) [4] は, データから多重共線性を取り除くために行われる PPIS でのデータ変換処理を改善した. α を $\alpha \in (0,1]$ かつ $d < [n\alpha]$ を満たす値とする. X を特異値分解することで得られる U, D, V を, d 番目の列と $[n\alpha]$ 番目の列を境界としてそれぞれ 3 つに分け, $U_1 = (\boldsymbol{u}_1, ..., \boldsymbol{u}_d)$, $U_{2a} = (\boldsymbol{u}_{d+1}, ..., \boldsymbol{u}_{[n\alpha]})$, $U_{2b} = (\boldsymbol{u}_{[n\alpha]+1}, ..., \boldsymbol{u}_n)$, $D_1 = \mathrm{diag}(\mu_1, ..., \mu_d)$, $D_{2a} = \mathrm{diag}(\mu_{d+1}, ..., \mu_{[n\alpha]})$, $D_{2b} = \mathrm{diag}(\mu_{[n\alpha]+1}, ..., \mu_n)$, $V_1 = (\boldsymbol{v}_1, ..., \boldsymbol{v}_d)$, $V_{2a} = (\boldsymbol{v}_{d+1}, ..., \boldsymbol{v}_{[n\alpha]})$, $V_{2b} = (\boldsymbol{v}_{[n\alpha]+1}, ..., \boldsymbol{v}_n)$ とする（図 2）. そして, データ変換行列を

$$Q_T = U_{2a} D_{2a}^{-1} U_{2a}^T \left\{ I_n - U_1 (U_1^T U_1) U_1^T \right\}$$

と定める. U_{2b} と D_{2b} は, 本来残すべき独自因子から余分に情報を排除してしまうため, 使用しない. 不要な U_{2b} と D_{2b} の部分を切り捨てることで, 独自因子部分をより正確に抽出でき, 変数選択の性能が向上する. 重要度の計算は, $\hat{\boldsymbol{y}} = Q_T \boldsymbol{y}$, $\hat{X} = Q_T X$, $\boldsymbol{\omega} = (\omega_1, ..., \omega_p)^T = \hat{X}^T \hat{\boldsymbol{y}}$ として, $|\omega_j|$ を j 番目の変数の重要度とする.

TPPIS の変数選択の性能は, 行列を 3 つに分割するための共通因子の次元数 d や α の値によって変化するため, これらを調整パラメータとみなし, その最適な値を高次元データに対応した BIC 型の基準 [2] を用いて選択する. シミュレーションデータおよび実データの分析を通して， 提案手法の有効性を確認する.

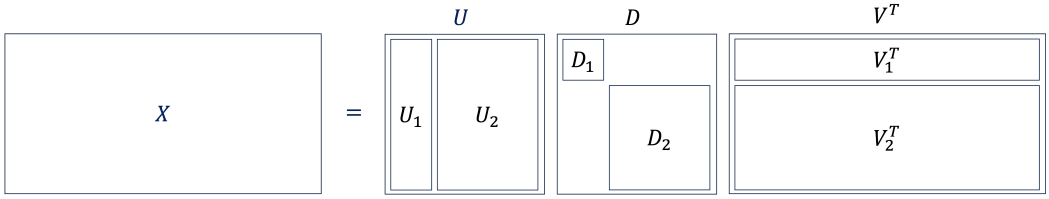


図 1　FPSIS と PPIS のデータ変換行列 Q_F, Q_P に用いられる行列のイメージ

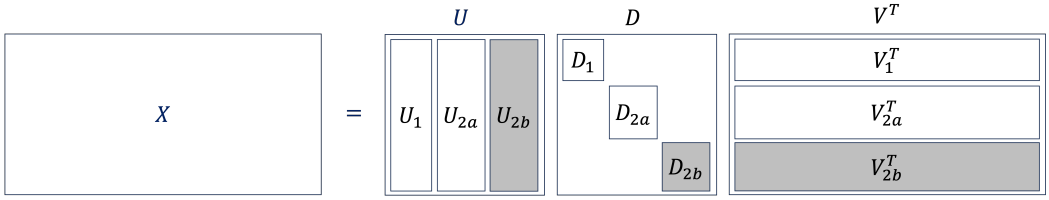


図 2　TPPIS のデータ変換行列 Q_T に用いられる行列のイメージ

References

[1] Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc B*. 2008;70(5):849–911.

[2] Wang H. Factor profiled sure independence screening. *Biometrika*. 2012;99(1):15–28.

[3] Zhao N, Xu Q, Tang ML, et al. High-dimensional variable screening under multicollinearity. *Stat*. 2020;9(1):e272.

[4] Tanaka S, Matsui H. Variable screening using factor analysis for high-dimensional data with multicollinearity. *arXiv preprint arXiv:2306.05702*. 2023;.

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推定

神戸大学大学院海事科学研究科 貝野友祐

1　モデル

$D=[0,1]^2$ 上の確率偏微分方程式

$$\begin{cases} \mathrm{d}X_t(y,z)=\left\{\theta_2\left(\dfrac{\partial^2}{\partial y^2}+\dfrac{\partial^2}{\partial z^2}\right)+\theta_1\dfrac{\partial}{\partial y}+\eta_1\dfrac{\partial}{\partial z}+\theta_0\right\}X_t(y,z)\mathrm{d}t \\ \qquad\qquad +\sigma\mathrm{d}W_t^Q(y,z), \quad (t,y,z)\in[0,1]\times D, \\ X_0(y,z)=\xi(y,z), \quad (y,z)\in D, \\ X_t(y,z)=0, \quad (t,y,z)\in[0,1]\times\partial D \end{cases} \tag{1}$$

を考える. ただし, $\{W_t^Q\}$ は D 上のソボレフ空間における Q-ウィーナー過程, 初期値 ξ は W^Q と独立, パラメータ空間 $\Theta\subset\mathbb{R}^3\times(0,\infty)^2$ はコンパクト凸集合, $(\theta,\sigma):=(\theta_0,\theta_1,\eta_1,\theta_2,\sigma)\in\Theta$ は未知パラメータ, $\theta^*=(\theta_0^*,\theta_1^*,\eta_1^*,\theta_2^*)$, σ^* は真値で $(\theta^*,\sigma^*)\in\operatorname{Int}\Theta$ とする. 既知の $\alpha\in(0,1)$ に対して Q-ウィーナー過程を

$$W_t^Q=\sum_{k,l\geq1}\lambda_{k,l}^{-\alpha/2}w_{k,l}(t)e_{k,l}$$

で定める. ただし $\{w_{k,l}\}_{k,l\geq1}$ は独立な1次元標準ブラウン運動であり,

$$e_{k,l}(y,z)=2\sin(\pi ky)\sin(\pi lz)\mathrm{e}^{-(\kappa y+\eta z)/2}, \quad \lambda_{k,l}=-\theta_0+\frac{\theta_1^2+\eta_1^2}{4\theta_2}+\pi^2(k^2+l^2)\theta_2$$

$(k,l\in\mathbb{N},(y,z)\in D)$ である. 高頻度時空間データとして $\mathbb{X}_{N,M}=\{X_{t_i}(y_{j_1},z_{j_2})\}_{0\leq i\leq N,0\leq j_1\leq M_1,0\leq j_2\leq M_2}$ を考える. ただし, $t_i=i\Delta=i/N$, $y_{j_1}=j_1/M_1$, $z_{j_2}=j_2/M_2$, $M:=M_1M_2$ とする.

2　$s=\sigma^2/\theta_2$, $\kappa=\theta_1/\theta_2$, $\eta=\eta_1/\theta_2$ の最小コントラスト推定

空間間引きデータ $\mathbb{X}_{N,m}^{(1)}=\{X_{t_i}(\widetilde{y}_{j_1},\widetilde{z}_{j_2})\}_{0\leq i\leq N,0\leq j_1\leq m_1,0\leq j_2\leq m_2}$ を考える. ただし, ある $0<\rho<1\wedge2(1-\alpha)$ と $b\in(0,1/2)$ について $m:=m_1m_2=O(N^\rho)$,

$$b\leq\widetilde{y}_0<\widetilde{y}_1<\cdots<\widetilde{y}_{m_1}\leq1-b, \quad b\leq\widetilde{z}_0<\widetilde{z}_1<\cdots<\widetilde{z}_{m_2}\leq1-b$$

とする.

$$Z_N(y,z)=\frac{1}{N\Delta^\alpha}\sum_{i=1}^N\Big\{X_{t_i}(y,z)-X_{t_{i-1}}(y,z)\Big\}^2, \quad f(y,z:s,\kappa,\eta)=\frac{\Gamma(1-\alpha)}{4\pi\alpha}s\,\mathrm{e}^{-(\kappa y+\eta z)}$$

とし, $\Xi\subset(0,\infty)\times\mathbb{R}^2$ はコンパクト凸集合, $(s^*,\kappa^*,\eta^*)\in\operatorname{Int}\Xi$ は真値とする. ただし, Γ はガンマ関数を表す. このとき (s,κ,η) の最小コントラスト推定量を空間間引きデータ $\mathbb{X}_{N,m}^{(1)}$ を用いて

$$U_{N,m}(s,\kappa,\eta)=\sum_{j_2=1}^{m_2}\sum_{j_1=1}^{m_1}\Big\{Z_N(\widetilde{y}_{j_1},\widetilde{z}_{j_2})-f(\widetilde{y}_{j_1},\widetilde{z}_{j_2}:s,\kappa,\eta)\Big\}^2,$$

$$(\hat{s},\hat{\kappa},\hat{\eta})=\operatorname*{argmin}_{(s,\kappa,\eta)\in\Xi}U_{N,m}(s,\kappa,\eta)$$

で定める. 正則条件の下, $N,m_1,m_2\to\infty$ のとき $N^{\frac{1}{2}\wedge(1-\alpha)}(\hat{s}-s^*,\hat{\kappa}-\kappa^*,\hat{\eta}-\eta^*)\xrightarrow{p}0$ が成り立つ.

3　係数パラメータ $\theta = (\theta_0, \theta_1, \eta_1, \theta_2)$ および σ^2 の適応的推定

時間間引きデータ $\mathbb{X}^{(2)}_{n,M} = \{X_{\widetilde{t}_i}(y_{j_1}, z_{j_2})\}_{0\le i\le n, 0\le j_1\le M_1, 0\le j_2\le M_2}$ を考える. ただし, $n \le N$ に対して

$$\widetilde{t}_i = \left\lfloor \frac{N}{n} \right\rfloor \frac{i}{N}, \quad i = 0, 1, \ldots, n$$

とする. 座標過程 $x_{k,l}(t)$ を時間間引きデータ $\mathbb{X}^{(2)}_{n,M}$ を用いて

$$\hat{x}_{k,l}(\widetilde{t}_i) = \frac{2}{M} \sum_{j_2=1}^{M_2} \sum_{j_1=1}^{M_1} X_{\widetilde{t}_i}(y_{j_1}, z_{j_2}) \sin(\pi k y_{j_1}) \sin(\pi l z_{j_2}) \mathrm{e}^{(\hat{\kappa} y_{j_1} + \hat{\eta} z_{j_2})/2}$$

で近似する. ただし $\{x_{k,l}\}_{k,l\ge1}$ は次で表される独立な 1 次元オルンシュタイン・ウーレンベック過程である.

$$\begin{cases} \mathrm{d}x_{k,l}(t) = -\lambda_{k,l} x_{k,l}(t)\mathrm{d}t + \sigma \lambda_{k,l}^{-\alpha/2} \mathrm{d}w_{k,l}(t), \\ x_{k,l}(0) = \langle \xi, e_{k,l} \rangle_\theta = \int_0^1 \int_0^1 \xi(y,z) e_{k,l}(y,z) \mathrm{e}^{\kappa y + \eta z} \mathrm{d}y \mathrm{d}z. \end{cases}$$

このとき座標過程のボラティリティパラメータ $\sigma_{k,l} := \sigma \lambda_{k,l}^{-\alpha/2}$ の推定量を

$$\hat{\sigma}^2_{k,l} = \sum_{i=1}^{n} \left\{ \hat{x}_{k,l}(\widetilde{t}_i) - \hat{x}_{k,l}(\widetilde{t}_{i-1}) \right\}^2$$

で定める. SPDE (1) の係数パラメータの適応的推定量を

$$\begin{aligned} \hat{\theta}_2 &= \left\{ \frac{3\pi^2}{\hat{s}^{1/\alpha}} \left(\frac{1}{(\hat{\sigma}^2_{1,2})^{1/\alpha}} - \frac{1}{(\hat{\sigma}^2_{1,1})^{1/\alpha}} \right)^{-1} \right\}^{\alpha/(1-\alpha)}, \\ \hat{\theta}_1 &= \hat{\kappa}\hat{\theta}_2, \quad \hat{\eta}_1 = \hat{\eta}\hat{\theta}_2, \quad \hat{\sigma}^2 = \hat{s}\hat{\theta}_2, \\ \hat{\theta}_0 &= -\left(\frac{\hat{s}\hat{\theta}_2}{\hat{\sigma}^2_{1,1}} \right)^{1/\alpha} + \left(\frac{\hat{\kappa}^2 + \hat{\eta}^2}{4} + 2\pi^2 \right) \hat{\theta}_2 \end{aligned}$$

で定め, $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1, \hat{\eta}_1, \hat{\theta}_2)$ とする. 正則条件の下, $n, M_1, M_2 \to \infty$ のとき $\sqrt{n}(\hat{\theta} - \theta^*, \hat{\sigma}^2 - (\sigma^*)^2) \xrightarrow{d} \mathcal{N}(0, \mathcal{J})$ が成り立つ. $\mathcal{J}$ の詳細については [5] を参照.

4　数値シミュレーション

$N = 10^3$, $M_1 = M_2 = 200$ とし, 高頻度データに基づいた大規模数値シミュレーションを行う. 具体的には，$(\hat{s}, \hat{\kappa}, \hat{\eta})$ および $(\hat{\theta}, \hat{\sigma}^2)$ を計算し，その漸近挙動を検証する.

参考文献

[1] M. Bibinger and M. Trabs. (2020). Volatility estimation for stochastic PDEs using high-frequency observations. *Stochastic Processes and their Applications*, 130(5):3005–3052.

[2] F. Hildebrandt and M. Trabs. (2021). Parameter estimation for SPDEs based on discrete observations in time and space. *Electronic Journal of Statistics*, 15(1):2716–2776.

[3] Y. Kaino and M. Uchida. (2020). Parametric estimation for a parabolic linear SPDE model based on discrete observations. *Journal of Statistical Planning and Inference*, 211:190–220.

[4] L. Piterbarg and A. Ostrovskii. *Advection and diffusion in random media: implications for sea surface temperature anomalies*. Springer Science & Business Media, 1997.

[5] Y. Tonaki, Y. Kaino, and M. Uchida. (2023). Parameter estimation for linear parabolic SPDEs in two space dimensions based on high frequency data. To appear in Scandinavian Journal of Statistics.

[6] H.C. Tuckwell. Stochastic partial differential equations in neurobiology: linear and nonlinear models for spiking neurons, In *Stochastic biomathematical models*, pages 149–173, 2013. Springer.

外れ値と過剰なゼロを含む計数データの事後ロバスト分析

入江 薫（東京大），羽村 靖之（京都大），菅澤 翔之助（慶應義塾大）

外れ値に対して敏感である線形回帰モデルを頑健に推定する方法論はロバスト統計と呼ばれ多くの研究がある．一方で，現代的な統計分析においては多数のパラメータを含む階層モデル，状態空間モデル，時空間モデルが用いられ，外れ値が推定に及ぼす影響も相対的に増大している．さらに，統計分析は点推定のみならず，区間推定量による不確実性の評価や予測，意思決定を含むことが一般的であり，そのような統計分析全体を外れ値から保護する野心的な頑健性も求められている．パラメータの点推定量ではなく，パラメータの事後分布に関する頑健性は事後ロバスト性（Posterior robustness）と呼ばれ，成立のための（簡易な）十分条件が明らかになったのは比較的最近のことである（Desgagné, 2015; Gagnon et al., 2020; Hamura et al., 2022）．その条件とは「誤差項の従う分布が厚い裾を持ち，誤差項によって外れ値を説明できる」という常識的なものであるが，事後ロバスト性が成立するためには誤差分布は超裾厚（super heavy-tailed）である，または対数正規変動（log-regularly varying）であることが要求される．ゆえに，t 分布では十分に裾が厚いとは言えず，事後ロバスト性は成立しない（Gagnon and Hayashi, 2023）．代わりに対数パレート分布などが活用されている．

一般化線形モデルにおいては，外れ値が特に問題となるのは計数データ（非負の整数値を取るデータ）を扱うポアソン回帰や負の二項分布回帰である．外れ値とは回帰項に比べて大きい，または小さい観測値を指すが，計数データにおいて後者はゼロとなることが多い．この問題は過剰なゼロ（zero inflation）と呼ばれ研究されてきたが，本研究では過剰なゼロを外れ値の一種とみなし，事後ロバスト性を達成することを目指す．

具体的には，階層構造を持つポアソン回帰モデル

$$y_i \sim \mathrm{Po}(\eta_i \exp\{x_i^\top \beta + g_i^\top \xi_i\}), \qquad i = 1, \dots, n,$$

を考え，極端に大きい値や過剰なゼロを誤差項 η_i で説明する．そのため，η_i の従う確率分布は超裾厚であり，かつゼロ付近で密度が発散している必要がある．そこで，我々はベータ分布を変数変換することにより得られる Rescaled-beta 分布の使用を提案する．その密度関数は以下で与えられる：

$$p(\eta_i; a, b) = \frac{1}{B(a,b)} \frac{\{\log(1+\eta_i)\}^{a-1}}{1+\eta_i} \frac{1}{\{1+\log(1+\eta_i)\}^{a+b}}, \quad \eta_i > 0.$$

この分布は超裾厚であり（$p(\eta_i; a, b) \sim \eta_i^{-1}\{\log \eta_i\}^{-(b+1)}$ as $\eta_i \to \infty$），かつ $0 < a < 1$ のとき原点で密度が発散する．密度関数は対数項を含み，そのままでは計算に不向きであるが，いくつかの混合表現が知られている：

$$\eta_i | u_i, v_i, w_i \sim \mathrm{Ex}(u_i), \quad u_i | v_i, w_i \sim \mathrm{Ga}(v_i, 1), \quad v_i | w_i \sim \mathrm{Ga}(b, w_i), \quad \text{and} \quad w_i \sim \mathrm{Be}(a, 1-a).$$

このモデルの η_i の周辺分布は Rescaled-beta 分布となる．ゆえに，適当な潜在変数を導入することで，ギブス・サンプラーにより簡易に事後分析が可能となる．本研究（Hamura et al., 2021）では上述のモデルについて，過剰なゼロに対する事後分布のふるまいや，事後ロバスト性の理論的証明を行った上で，階層ポアソン回帰モデルを用いた数値実験や犯罪統計データの空間モデルによる頑健な解析を行う．

以上の内容に沿って，講演当日は研究の背景を述べた後に，提案するモデル，計算手法，理論的性質を説明し，数値実験および実データ分析の結果を紹介した．

質疑応答では，外れ値の種類や，外れ値の構造によって提案手法の理論的性質に影響はあるか質問があった．本研究で用いる事後ロバスト性の定義は「極端な値が観測される時のロバスト性」であり，外れ値の構造に依存せずにロバスト性が成立することを指摘した．そのうえで，本研究では外れ値の構造を明らかにしたり，外れ値を特定したりすることは目的としていないことを述べ，手法の限界について議論した．また，使用する Rescaled beta 分布について，導出の方法や，確かに確率密度関数になっていることについての確認の質問が

あり，解説を行った．バイナリデータやカテゴリカルデータにおける外れ値についても質問があったが，連続値データやカウントデータとは異なる扱いになるであろうことを述べた．

Desgagné, A. (2015). “Robustness to outliers in locationscale parameter model using log-regularly varying distributions.” *The Annals of Statistics*, 43(4).

Gagnon, P., Desgagné, A., & Bdard, M. (2020). “A New Bayesian Approach to Robustness Against Outliers in Linear Regression.” *Bayesian Analysis*, 15(2), 389-414.

Hamura, Y., Irie, K., & Sugasawa, S. (2022). “Log-regularly varying scale mixture of normals for robust regression.” *Computational Statistics & Data Analysis*, 173, 107517.

Hamura, Y., Irie, K., & Sugasawa, S. (2021). “Robust hierarchical modeling of counts under zero-inflation and outliers.” *arXiv preprint* arXiv:2106.10503.

Gagnon, P., & Hayashi, Y. (2023). “Theoretical properties of Bayesian Student-t linear regression.” *Statistics & Probability Letters*, 193, 109693.

Jeffreys prior と MLE に共通した役割とその含意

柳 本 武 美: 統計数理研究所

1. 序

ベイズ推定における Jeffreys prior の仮定と、母数推定における MLE には著しい共通点が観察される。しかし、多くの文献では Jeffreys prior の仮定には批判的である一方、MLE は広く受け入れられているように見える。この乖離を整理して理解を深めることは、これらを乗り越えて新しい発展を試みる基礎として重要である。以下の議論では、ごく弱い正則条件は断り無く仮定する。

2. 役割の共通性

標本密度 $p(\boldsymbol{x}|\theta), \theta \in \Theta \subset R^p$ に従う母集団からの大きさ n の標本を $\boldsymbol{x}$ と書く。ベイズ推定では事前密度 (関数) $\pi(\theta)$ を仮定する。

2.1. 歴史的な共通性：ごく常識的な無情報事前密度は一様事前密度 $\pi(\theta) \propto 1$ である。初期の研究者 Bayes, Laplace らは一様事前分布を仮定した。しかし、母数の変数変換に依存してしまう。その対案として提出された Jeffreys prior は、I(θ) を Fisher 情報量、として $\pi_J(\theta) \propto |I(\theta)|^{1/2}$ で定義される。分布族の構造を利用している点でより理論的であり、変数変換に関して不変である。

一方、誤差論としての初期の母数推定では当初はモーメント法とか最小二乗法が用いられていた。最尤推定量は $\hat{\theta}_{ML} = \mathrm{Argmax}_\theta\ p(\boldsymbol{x}|\theta)$ で定義される。分布族の構造を利用している点でより理論的である。

2.2. その革新的な役割：Jeffreys prior 変数変換に伴う Jacobian の補正項と理解され、母数の変数変換に関して不変な事前密度である。事前密度の選択のために、主観的な考察とか標本密度に関する制約は殆どない。ベイズ推定量の最適性が担保される。MLE はモデルが定められると一意に定まる。二次モーメントを利用するような恣意性がない。実際の計算も最適化プログラムを利用して実行される。尤度推定方程式は不偏であり、漸近的には最適性が成り立つ。

2.3. 正規分布の特徴付け：一次元位置母数族 $p(\boldsymbol{x}|\theta) = \prod g(x_i - \theta)$ では、MLE が常に標本平均に一致する条件から正規分布が導かれる。なお、このモデルでは Jeffreys prior は一様分布になる。この特徴付けは、正規分布の良さと共に MLE が良いとも受け止められた。

一方、一次元指数分布族 $p(\boldsymbol{x}|\theta) = \prod \exp(x_i\theta - M(\theta))a(x_i)$ では Jeffreys prior が一様分布に一致する条件から正規分布が導かれる。なお、このモデルでは MLE は常に標本平均になる。

2.4. 実用性：Jeffreys prior と MLE は共に定義が明確で多様なモデルに対して形式的に適用できる。前提条件が殆ど無いことは実用的には好都合である。また、直感的な一様事前分布とかモーメ

ント推定量より性能が優れている。従って、第一選択として推奨される。

3. 共通の欠陥と対策

今日における両者に共通した欠陥は、母数の次元が高いときに導出される推定量の振る舞いの悪さにある。次元 p が極めて高いと破滅的ある。次元が 2, 3 のように低くても、仔細に調べるとしばしば改善の余地がある。正規分布 $\mathrm{N}(\mu, \sigma^2)$ とか二つの二項分布が例として挙げられる。その欠陥は本質的である。

Jeffreys prior の改良として α prior などが提案されている。また、reference prior (Bernaldo, 1979) は広く受け入れられている。MLE の改良としては条件付き尤度、部分尤度あるいは周辺尤度の尤度を最大化させる提案がなされているが、改良の選択肢は限られている。

4. 評価の乖離

上述のように、改めて整理すると Jeffreys prior と MLE の役割はその歴史的革新性と致命的欠陥に関して著しい共通性が観察される。その一方で、文献上での評価は大きく異なる。Jeffreys prior に関しては否定的な評価が圧倒的である。MLE は maximum likelihood principle の用語が見られる (例えば Goodfellow ら, 2016) ように広く受け入れられている。ベイズ推定でも MAP 解 (事後尤度最大化解) が受け入れられている。また、数値計算でも最適化法適用できる。

5. 性能の優劣

上に記述した一般的な評価とは逆に、Jeffreys prior に基づいた推定量 $\hat{\theta}_J = \mathrm{E}_{\mathrm{post}}[\theta; \pi_J(\theta|\boldsymbol{x})$ が MLE より性能が優れていることが観察される。指数分布族では自然母数、位置母数族では位置母数の事後平均で推定するとして議論する。MLE の長所として漸近的な最適性があり尤度方程式の不偏性がある。ところが指数分布族の下では、推定量 $\hat{\theta}_J$ と MLE は漸近的に $O(1/n^2)$ のオーダーで漸近一致する (Yanagimoto and Ohinishi, 2021)。従って漸近的な最適性は担保される。不偏性に関しても様々な例で観察される (Sakumura and Yanagimoto, in press)

より一般の場合の性能を比較する際に先ず注意することは、有限標本の下で $\hat{\theta}_J$ には最適性が認められるが、MLE に認められない事実がある。Yanagimoto and Ohnishi (2011) は鞍点条件の下で $\hat{\theta}_J$ の最適性と MLE の最悪性を示している。具体的な例で性能をシミュレーションで比較すると、$\hat{\theta}_J$ が MLE に優越することが観察される。

文献：1) Bernardo, J.M. (1979). *J. Roy. Statist. Soc. Ser. B*, **41**, 113-147. 2)Goodfellow, I., Bengio, Y. and Courville, A. (2016). Deep Learning, MIT press Cambridge. 3) Sakumura, T. and Yanagimoto, T. To appear in *Commun. In Statist. - Simul. Comput.* 4) Yanagimoto, T. and Ohnishi, T. (2011). *J. Statist. Plann. Inf.*, **41**, 1990-2000. 5) Yanagimoto, T. and Ohnishi, T. (2021). *Pioneering Works on Distribution Theory: In Honor of Masaaki Sibuya*, 103-121, Springer.

共役ガンマ事前分布を用いた
ワイブル分布のベイズ推定

作村建紀 *　　柳本武美 †

1　はじめに

ワイブル分布からのサイズ n の標本 $\boldsymbol{x} = (x_1, \cdots, x_n)$ を考える．ワイブル分布は次の密度を持つ．

$$p(x|\mu,\tau) = \tau\mu^{-\tau}x^{\tau-1}e^{-(x/\mu)^\tau}$$

ここで，$x > 0$, $\mu > 0$, $\tau > 0$ である．この密度は，τ が固定のとき指数型分布族に属す．自然母数は $1/\mu^\tau$ であり，対応する十分統計量は $\sum x_i^\tau$ である．x^a $(a > 0)$ の期待値は $\mu^a\Gamma(1 + a/\tau)$ となる．

本研究では，ワイブル分布のパラメータ (μ, τ) のベイズ推定を考える．ベイズ推定では事前分布の選択が重要である．過去の十分なデータがあれば，安定した適切な事前分布が導かれる．一方そうでなければ，無情報事前分布が考えられる．たとえば，参照事前分布などが候補になる (Bernardo, 1979; Sun, 1997; Xu, Fu, Tang, & Guan, 2015)．本研究では，弱い情報を持たせた事前分布として，μ に対して共役な事前分布を，τ に対してガンマ事前分布を考え，特殊なケースとして無情報事前分布を内包する形で導入する．

ベイズ推定で重要なもう一つの点は，推定対象の選択である．たとえば，指数型分布族に対しては自然母数を推定対象とし，その事後平均を考えることで最適性を有する推定量を構築できる (Yanagimoto & Ohnishi, 2009)．これは，適切な推定対象の選択が重要であることを示唆するものである．本研究では，この自然母数に着想を得た推定量を考える．その性能を頻度論の観点から評価する．

2　共役ガンマ事前分布

標本 $\boldsymbol{x}$ の幾何平均を $\tilde{x}$ とする．また，$\eta = 1/\mu^\tau$, $\tilde{\mu}^\tau = \sum x_i^\tau/n$ とする．$\tilde{\mu}$ は μ の最尤推定量である．すると，ワイブル分布の標本密度は以下のようになる．

$$p(\boldsymbol{x}|\eta,\tau) = \eta^n\tau^n\tilde{x}^{n(\tau-1)}e^{-n\eta\tilde{\mu}^\tau}. \qquad (1)$$

τ を固定したとき η の共役事前分布を次で定義する．

$$\begin{aligned}&\pi(\eta|\tau; m, \delta_1)\\ &= e^{-\delta_1(m^\tau\eta - \log(m^\tau\eta)-1)}\pi_J(\eta)e^{-k(m^\tau,\delta_1)} \ (\delta_1 > 0),\\ &\propto \pi_J(\eta) \ (\delta_1 = 0).\end{aligned}$$

ここで，$\pi_J(\eta) = 1/\eta$ はジェフリース事前分布である．また，m は平均パラメータである．τ の事前分布は次で定義する．

$$\begin{aligned}\pi(\tau; t, \delta_2) &= (t\delta_2)^{\delta_2}\tau^{\delta_2-1}e^{-t\delta_2\tau}/\Gamma(\delta_2) \ (t, \delta_2 > 0),\\ &\propto 1/\tau \ (\delta_2 = 0)\end{aligned}$$

以上を踏まえて，(η, τ) の事前分布を $\pi(\eta, \tau) = \pi(\eta|\tau; m, \delta_1)\pi(\tau; t, \delta_2)$ と仮定する．$\delta_1 = 0$ の場合を考えると，(η, τ) の事前分布は，

$$\pi(\eta, \tau) \propto \eta^{-1}(t\delta_2)^{\delta_2}\tau^{\delta_2-1}e^{-t\delta_2\tau}/\Gamma(\delta_2) \qquad (2)$$

となる．$\delta_1 = \delta_2 = 0$ の場合，$\pi(\eta, \tau) \propto \eta^{-1}\tau^{-1}$ となる．

3　提案推定量

本節では，前節で定義した $\delta_1 = 0$ のときの事後分布 $\pi(\eta, \tau|\boldsymbol{x})$ のもとで，ワイブル分布のパラメータ (μ, τ) に対するベイズ推定量を提案する．

まず τ を固定したときの自然母数 η の事後平均を考えると，$\hat{\eta} = \hat{\eta}(\tau) = 1/\tilde{\mu}^\tau$ が得られる．このとき，τ の推定量 $\hat{\tau}$ は $\pi(\eta, \tau)$ のもとでの τ の事後平均として求める．$\delta_1 = 0$ のとき，周辺事後密度 $\int \pi(\eta, \tau|\boldsymbol{x})d\eta$ は，$\tau^{n+\delta_2-1}\tilde{x}^{n\tau}\tilde{\mu}^{-n\tau}e^{-t\delta_2\tau}$ に比例することから，τ の事後平均は次式で得られる．

$$\hat{\tau} = \frac{\int_0^\infty \left(\frac{\tilde{x}^\tau}{\tilde{\mu}^\tau}\tau\right)^n \tau^{\delta_2}e^{-t\delta_2\tau}d\tau}{\int_0^\infty \left(\frac{\tilde{x}^\tau}{\tilde{\mu}^\tau}\tau\right)^n \tau^{\delta_2-1}e^{-t\delta_2\tau}d\tau}.$$

また，μ の推定量は $\hat{\mu} = 1/\hat{\eta}^{1/\hat{\tau}}$ として求める．

4　リスク比較

提案推定量と既存推定量はリスクで比較される．既存推定量としては最尤推定量（MLE）を採用する．

*法政大学理工学部：〒 184-8584 東京都小金井市梶野町 3-7-2.
†統計数理研究所：〒 190-8562 東京都立川市緑町 10-3.

表 1: μ の二乗損失

		Proposed			
n	τ	$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$	MLE
3	0.5	3.491	3.508	3.525	3.568
	1	0.377	0.374	0.371	0.381
	2	0.087	0.086	0.085	0.087
	5	0.015	0.015	0.015	0.015
	10	0.004	0.004	0.004	0.004
5	0.5	1.554	1.563	1.572	1.593
	1	0.223	0.222	0.221	0.225
	2	0.053	0.053	0.053	0.053
	5	0.009	0.009	0.009	0.009
	10	0.002	0.002	0.002	0.002
10	0.5	0.627	0.629	0.632	0.638
	1	0.113	0.113	0.113	0.114
	2	0.027	0.027	0.027	0.027
	5	0.004	0.005	0.005	0.004
	10	0.001	0.001	0.001	0.001
20	0.5	0.254	0.254	0.255	0.257
	1	0.054	0.054	0.054	0.055
	2	0.013	0.013	0.013	0.013
	5	0.002	0.002	0.002	0.002
	10	0.001	0.001	0.001	0.001

表 2: カルバックライブラー損失

		Proposed			
n	τ	$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$	MLE
3	0.5	1.494	1.464	1.439	1.510
	1	1.494	1.417	1.353	1.510
	2	1.494	1.334	1.213	1.510
	5	1.494	1.142	0.958	1.509
	10	1.494	0.958	0.904	1.510
5	0.5	1.236	1.229	1.223	1.251
	1	1.236	1.212	1.189	1.251
	2	1.236	1.178	1.127	1.251
	5	1.236	1.094	1.000	1.251
	10	1.236	1.001	0.952	1.251
10	0.5	1.107	1.105	1.103	1.116
	1	1.107	1.098	1.090	1.116
	2	1.107	1.085	1.066	1.116
	5	1.107	1.052	1.013	1.116
	10	1.107	1.014	0.989	1.116
20	0.5	1.035	1.034	1.033	1.040
	1	1.035	1.031	1.028	1.040
	2	1.035	1.025	1.017	1.040
	5	1.035	1.011	0.993	1.040
	10	1.035	0.994	0.981	1.041

シミュレーションでは $\mu = 1$ とする．また，$t = 1$ とする．すべてのシミュレーションにおいて，10,000 回行う．

4.1 二乗損失

μ についての二乗損失 $L_s(\hat{\mu}, \mu) = (\hat{\mu} - \mu)^2$ を考える．推定されたリスクの結果を表 1 に示す．τ が小さいところでは，MLE よりも提案推定量の推定リスクが小さい．τ が大きくなると，両者の推定リスクに大きな差はなくなる．これはどのサンプルサイズ n でも同じ傾向である．また，δ_2 の違いに対して，推定リスクの値に大きな差は見受けられないことが観察される．

4.2 カルバックライブラー損失

カルバック・ライブラー損失の結果を表 2 にまとめる．ここでも，MLE のリスクよりも提案推定量のリスクのほうが全体的に小さい傾向が見受けられる．δ_2 を大きくするとリスクは小さくなる．特に，τ が大きくなるとその傾向は顕著になる．逆に，$\tau = 0.5$ のときの δ_2 の大きさによる影響は小さい．また，サンプルサイズによるリスクの変化は，二乗損失の場合と比べると小さいことが観察される．

5 考察

今回の比較では $\delta_1 = 0$ という限定的なものを扱い，δ_2 を変化させて検証を行った．τ の事前分布の選択を誤ったとしても，μ の推定は良い性能を示す．特に，パラメータ (μ, τ) の提案推定量の性能は MLE を上回り有望な結果である．

References

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(2), 113–147.

Sun, D. (1997). A note on noninformative priors for Weibull distributions. *Journal of Statistical Planning and Inference*, *61*(2), 319-338. doi: https://doi.org/10.1016/S0378-3758(96)00155-3

Xu, A., Fu, J., Tang, Y., & Guan, Q. (2015). Bayesian analysis of constant-stress accelerated life test for the Weibull distribution using noninformative priors. *Applied Mathematical Modelling*, *39*(20), 6183-6195. doi: https://doi.org/10.1016/j.apm.2015.01.066

Yanagimoto, T., & Ohnishi, T. (2009). Bayesian prediction of a density function in terms of e-mixture. *Journal of Statistical Planning and Inference*, *139*(9), 3064-3075. doi: https://doi.org/10.1016/j.jspi.2009.02.005

ROBUSTIFYING GAUSSIAN QUASI-LIKELIHOOD INFERENCE FOR VOLATILITY

SHOICHI EGUCHI AND HIROKI MASUDA

1. Setup and objective

Given a complete filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, P)$ for a fixed time horizon $[0,T]$, we consider the process

$$Y_t^\star = Y_0^\star + \int_0^t \mu_s ds + \int_0^t \sigma(X_{s-}, \theta) dw_s + J_t, \qquad t \in [0,T],$$

where μ and X are càdlàg processes in $\mathbb{R}^d$ and $\mathbb{R}^q$, respectively ($X_{s-} := \lim_{u\uparrow s} X_u$), w is a standard r-dimensional Wiener process, and J is the jump part of $Y^\star$ which is assumed to be finite-activity: $J_t = \sum_{0<s\le t} \Delta Y_s^\star$ with $\sum_{0<s\le t} 1_{\{\Delta Y_s^\star \neq 0\}} < \infty$ a.s., where $\Delta Y_s^\star := Y_s^\star - Y_{s-}^\star$ denotes the jump size of $Y^\star$ at time s. All the processes are $(\mathcal{F}_t)$-adapted and the components of the covariate process X may contain those of Y. The diffusion coefficient $\sigma : \mathbb{R}^q \times \Theta \to \mathbb{R}^d \otimes \mathbb{R}^r$ is known except for the finite-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^p$, the parameter space Θ assumed to be a bounded convex domain. Let $t_j = t_j^n := jh$ with $h = h_n := T/n$; we largely omit the dependence on n from the notation. For each $n \in \mathbb{N}$, we define the stochastic process $Y = (Y_t)_{t\in[0,T]}$ as follows:

$$Y_t = \begin{cases} Y_0^\star & t = 0 \\ Y_t^\star & t \in (t_{j-1}, t_j) \\ Y_{t_j}^\star + b_j \Lambda_j & t = t_j \end{cases} \tag{1.1}$$

where $b_j = b_{n,j} \in \{0,1\}$ a.s and $\Lambda_j = \Lambda_{n,j} \in \mathbb{R}^d$ are $\mathcal{F}_{t_j}$-measurable random variables. Note that a sample path of Y is a.s. Riemann integrable while is not right-continuous if it $Y^\star$ is blurred by a "spike noise" $(b_j\Lambda_j)$, the number of which is much smaller than n.

Our objective is to estimate the true value $\theta_0 \in \Theta$, assumed to exist, based on a discrete-time sample $\{(X_{t_j}, Y_{t_j})\}_{j=0}^n$. Obviously, the "ideal" situation is the case where $b_j \equiv 0$ and $J \equiv 0$ so that the model (X,Y) equals the continuous regression model considered in [10] (see also [2]):

$$Y_t = Y_0 + \int_0^t \mu_s ds + \int_0^t \sigma(X_{s-}, \theta) dw_s.$$

We regard both $\{(b_j, \Lambda_j)\}_{n,j}$ and J as "contamination", leaving the drift function μ unknown.

Denote by $\Delta_j \xi = \xi_{t_j} - \xi_{t_{j-1}}$ the jth increment of a process ξ, and write $f_{j-1}(\theta) = f(X_{t_{j-1}}, Y_{t_{j-1}}; \theta)$ for any measurable function f defined on $\mathbb{R}^q \times \mathbb{R}^d \times \overline{\Theta}$. When ignoring the drift, the jump component, and the conditional-mixture structure (1.1), the Euler approximation scheme is given by

$$Y_{t_j} \overset{P_\theta}{\approx} Y_{t_{j-1}} + h\sigma_{j-1}(\theta)\Delta_j w.$$

In this case, the conventional Gaussian quasi-(log-)likelihood function (GQLF) is given by, with $S := \sigma^{\otimes 2}$,

$$\theta \mapsto -\frac{1}{2}\sum_{j=1}^n \left(\log\det(S_{j-1}(\theta)) + \frac{1}{h} S_{j-1}(\theta)^{-1}[(\Delta_j Y)^{\otimes 2}] \right),$$

which would miserably behave if the contaminations are present. We refer to [2] and [10] for the asymptotics of this random function in case where $Y = Y^\star$ and $Y^\star$ has no jumps. It is well-known that the unbounded Gaussian quasi-score function $\partial_\theta \mathbb{H}_n(\theta)$ is inevitably much influenced by outliers related to $\mathcal{L}\{(X,Y)\}$. We show that the density-power tapering [1] (also [5]) effectively works in asymptotic inference, as a handy alternative to the threshold estimation where several sensitive fine tunings are inevitable (see [7] and [3]). To the best of our knowledge, the only previous studies on the robust divergence-based inference for stochastic differential equation models are [6], [8], and [9], all of which are concerned with ergodic Markovian diffusions without theoretical consideration in the presence of the contaminations.

2. Density-power divergence and Gaussian quasi-likelihood inference

2.1. **Construction.** Write $\phi_j(\theta) = \phi_{n,j}(\theta) := \phi\left(Y_{t_j}; Y_{t_{j-1}}, hS_{j-1}(\theta)\right)$ and $\psi_j(\theta) := \partial_\theta \log \phi_j(\theta)$, where $\phi(\cdot; \mu, \Sigma)$ denotes the d-dimensional normal $N_d(\mu, \Sigma)$-density. We will consider the density-power weighting [1] in our framework, by multiplying the (non-predictable) weight $\phi_j(\theta)^\beta$ to each summand, namely $\sum_{j=1}^n \phi_j(\theta)^\beta \psi_j(\theta)$; the tuning parameter β is assumed to take its values in $[0, \overline{\beta}]$ for a given $\overline{\beta} \in (0, \infty)$ and should be pre-assigned by a user. Since the weighting entails a bias in the quasi-likelihood equation, we consider the compensation to obtain the associated martingale estimating function:

$$\theta \mapsto \sum_{j=1}^n \left(\phi_j(\theta)^\beta \psi_j(\theta) - E_\theta^{j-1}[\phi_j(\theta)^\beta \psi_j(\theta)] \right). \tag{2.1}$$

The conditional distribution $\mathcal{L}(Y_{t_j} | \mathcal{F}_{t_{j-1}})$ can be rarely explicitly given, so that (2.1) itself cannot be of direct use for estimating θ. We need to reasonably approximate the term $E_\theta^{j-1}[\phi_j(\theta)^\beta \psi_j(\theta)]$ in an explicit way, which will turn out to be possible because of the high-frequency nature of the sample.

By approximating the conditional expectation in (2.1), we introduce the explicit *density-power GQLF*:

$$\mathbb{H}_n(\theta; \beta) := \frac{h^{d\beta/2}}{\beta} \sum_{j=1}^n \phi_j(\theta)^\beta - \frac{1}{(2\pi)^{d\beta/2}(\beta+1)^{1+d/2}} \sum_{j=1}^n \det(S_{j-1}(\theta))^{-\beta/2}.$$

Given a value $\beta > 0$, we define the *density-power GQMLE* by any element $\hat{\theta}_n(\beta) \in \operatorname{argmax}_{\theta \in \overline{\Theta}} \mathbb{H}_n(\theta; \beta)$. We are interested in the asymptotic behavior of the rescaled density-power GQMLE $\sqrt{n}(\hat{\theta}_n(\beta) - \theta_0)$ with handling the effects of jumps caused by J and possible measurement errors.

2.2. **Asymptotic mixed normality.** With some structural assumptions on the covariate process X, we can deduce the asymptotic mixed normality of $\hat{\theta}_n(\beta)$; our model may be non-Markovian, and the key tool is the stable-convergence result due to Jacod [4], where certain filtration structure is essential to verify the conditional Gaussian character of the asymptotic distribution. Under some regularity conditions, we obtain the asymptotic mixed normality

$$\sqrt{n}(\hat{\theta}_n(\beta) - \theta_0) \xrightarrow{\mathcal{L}} MN_p\left(0, \Gamma_0(\beta)^{-1} \Sigma_0(\beta) \Gamma_0(\beta)^{-1}\right),$$

where the random matrices $\Gamma_0(\beta)$ and $\Sigma_0(\beta)$ are explicitly given through the Riemann integrals of the form $T^{-1} \int_0^T g(X_s, \beta, \theta_0) dt$. Hence it is straightforward to provide consistent estimators of the random asymptotic covariance matrix, followed by a practical recipe for approximate confidence set for θ.

References

[1] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[2] V. Genon-Catalot and J. Jacod. On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Ann. Inst. H. Poincaré Probab. Statist.*, 29(1):119–151, 1993.

[3] H. Inatsugu and N. Yoshida. Global jump filters and quasi-likelihood analysis for volatility. *Ann. Inst. Statist. Math.*, 73(3):555–598, 2021.

[4] J. Jacod. On continuous conditional Gaussian martingales and stable convergence in law. In *Séminaire de Probabilités, XXXI*, volume 1655 of *Lecture Notes in Math.*, pages 232–246. Springer, Berlin, 1997.

[5] M. C. Jones, N. L. Hjort, I. R. Harris, and A. Basu. A comparison of related density-based minimum divergence estimators. *Biometrika*, 88(3):865–873, 2001.

[6] S. Lee and J. Song. Minimum density power divergence estimator for diffusion processes. *Ann. Inst. Statist. Math.*, 65(2):213–236, 2013.

[7] Y. Shimizu and N. Yoshida. Estimation of parameters for diffusion processes with jumps from discrete observations. *Stat. Inference Stoch. Process.*, 9(3):227–277, 2006.

[8] J. Song. Robust estimation of dispersion parameter in discretely observed diffusion processes. *Statist. Sinica*, 27(1):373–388, 2017.

[9] J. Song. Robust test for dispersion parameter change in discretely observed diffusion processes. *Comput. Statist. Data Anal.*, 142:106832, 17, 2020.

[10] M. Uchida and N. Yoshida. Quasi likelihood analysis of volatility and nondegeneracy of statistical random field. *Stochastic Process. Appl.*, 123(7):2851–2876, 2013.

Faculty of Information Science and Technology, Osaka Institute of Technology, 1-79-1 Kitayama, Hirakata City, Osaka, 573-0196, Japan.

Email address: shoichi.eguchi@oit.ac.jp

Graduate School of Mathematical Sciences, University of Tokyo, 3-8-1 Komaba Meguro-ku Tokyo 153-8914, Japan.

Email address: hmasuda@ms.u-tokyo.ac.jp

Estimation of the number of relevant factors from high-frequency data

小池 祐太 *

2023 年度科研費シンポジウム
「データサイエンスにおける統計的理論・方法論の新展開」
報告書

d 個の金融資産の対数価格過程 $Y = (Y_t)_{t\in[0,1]}$ が以下の連続時間ファクターモデルに従っているものとする:

$$Y_t = \beta F_t + Z_t, \qquad t \in [0,1]. \tag{1}$$

ここに,

- β は非ランダムな $d \times r$ 行列で, 各列が因子負荷量に対応する.
- $F = (F_t)_{t\in[0,1]}$ は r 次元セミマルチンゲールでファクター過程を表す.
- $Z = (Z_t)_{t\in[0,1]}$ は d 次元セミマルチンゲールで個別変動を表す成分に対応する.

r, β, F, Z はすべて未知であるとし, Y の等間隔時点 $t_i = i/n$ $(i = 0, 1, \ldots, n)$ での観測データ $(Y_{t_i})_{i=0}^n$ のみが与えられているとする. また, モデル (1) に「弱い」ファクターがいくつか含まれることを許容する. すなわち, 行列 $d^{-1}\beta^\top\beta$ の固有値のいくつかが $d \to \infty$ のとき 0 に収束することを許容する. このような状況において, 高次元高頻度データの漸近論 $d, n \to \infty$ の下で「実用上意味がある程度に十分強いファクター」の数を推定するのが本報告の目的である.

目的をより正確に述べるために, ファクター過程と因子負荷量に関する次の仮定を導入する:

[A1] (i) F および r は d, n に依存せず固定されている. また, F の二次共変動行列 $[F,F]_1$ は確率 1 で可逆である.
(ii) 定数 $1 \geq \alpha_1 \geq \cdots \geq \alpha_r > 0$ が存在して, 各 $j = 1, \ldots, r$ について $d \to \infty$ のとき $\lambda_j(\beta^\top\beta) \asymp d^{\alpha_j}$ が成り立つ.

この仮定は, 離散時間モデルにおいて弱いファクターをモデル化するための典型的な仮定 (Bai & Ng (2023) や Uematsu & Yamagata (2023a,b) 参照) の連続時間版とみなせる. このとき, $\tau \in (0,1)$ を 1 つ決めた下で,

$$r_\tau := \max\{j \in \{1, \ldots, r\} : \alpha_j > \tau\}$$

* 東京大学大学院数理科学研究科, CREST JST

を推定するのが本報告の目的である. 以下, 便宜上 $\max \emptyset = 0$ と定める. Freyaldenhoven (2022) で論じられているように, 金融資産価格に対するモデルについては $\tau = 1/2$ の場合が応用上興味のあるファクターに対応する.

Y の実現共分散行列

$$[Y,Y]_1^n := \sum_{i=1}^{n} (Y_{i/n} - Y_{(i-1)/n})(Y_{i/n} - Y_{(i-1)/n})^\top$$

を考える. $[Y,Y]_1^n$ の固有値を降順に重複度込みで並べたものを $\lambda_1([Y,Y]_1^n) \geq \cdots \geq \lambda_d([Y,Y]_1^n)$ とする. このとき, 適当な正則条件の下で以下のことが示せる:

(a) 任意の $j \leq r_\tau$ に対して, $\lambda_j\left([Y,Y]_1^n\right) \asymp_p d^{\alpha_j}$ $(n \to \infty)$.
(b) $\max_{j:j>r_\tau} \lambda_j\left([Y,Y]_1^n\right) = O_p(d^\tau)$ $(n \to \infty)$.

この事実を用いると, r_τ の自然な推定量として以下のものが考えられる:

$$\hat{r}_\tau^{BN} := \max\left\{j \in \{1,\ldots,r_{max}\} : \lambda_j([Y,Y]_1^n) > d^\tau g(d)\right\}.$$

ここに, r_{max} は r_τ の任意の上界で, $g : (0,\infty) \to (0,\infty)$ は緩変動関数である. この推定量は適当な正則条件の下で r_τ の強一致推定量となることが示せるが, 実用上はパフォーマンスが g の選び方に非常に左右されるため実装が難しい. そのため, Pelger (2019) において提案された摂動付き固有値の比に基づく推定量も考えてみる. これは,

$$ER_j([Y,Y]_1^n) := \frac{\lambda_j([Y,Y]_1^n) + d^\tau g(d)}{\lambda_{j+1}([Y,Y]_1^n) + d^\tau g(d)} \qquad (j = 1,\ldots,d-1)$$

としたとき, $ER_j([Y,Y]_1^n)$ は $j = r_\tau$ ならば 1 より大きく $j > r_\tau$ ならば 1 に近い値をとるという事実に基づく推定量で, 以下のように定義される:

$$\hat{r}_\tau^{P}(\gamma) := \max\left\{j \in \{1,\ldots,r_{max}\} : ER_j([Y,Y]_1^n) > 1+\gamma\right\}.$$

ここに, $\gamma > 0$ はチューニングパラメータである. この推定量も適当な正則条件の下で r_τ の強一致推定量となることを示すことができて, 有限標本でのパフォーマンスも悪くないことが確認できる.

参考文献

Bai, J. & Ng, S. (2023). Approximate factor models with weaker loadings. J. Econometrics **235**, 1893–1916.

Freyaldenhoven, S. (2022). Factor models with local factors — determining the number of relevant factors. J. Econometrics **229**, 80–102.

Pelger, M. (2019). Large-dimensional factor modeling based on high-frequency observations. J. Econometrics **208**, 23–42.

Uematsu, Y. & Yamagata, T. (2023a). Estimation of sparsity-induced weak factor models. J. Bus. Econom. Statist. **41**, 213–227.

Uematsu, Y. & Yamagata, T. (2023b). Inference in sparsity-induced weak factor models. J. Bus. Econom. Statist. **41**, 126–139.

Simultaneous Confidence Region of an Embedded One-Dimensional Curve in Multi-Dimensional Space

千葉大・理学研究院　内藤 貫太

はじめに:　本研究の源は，Naito et al. (2010) に遡る．“ヒト胎児発生過程の多次元スタンダードをどのように構築するのか？”という課題に対し，多次元空間に埋め込まれた 1 次元曲線の推定問題として問題設定がなされ，成果が得られた．しかしながら，1 次元曲線の “同時信頼領域”の構築は未解決のままになっていた．この未解決課題に再び目を向けることになる動機はネットワークデータである．道路交通網上の交通事故や，脳神経細胞網における樹状突起スピンの分布といった，枝分かれした 1 次元曲線上で観測されるデータをネットワークデータを呼ぶ (Liu and Ruppert, 2021; McSwiggan et al., 2017)．興味深い研究があるものの，ネットワークを構成する 1 次元曲線の同時信頼領域を構築する方法については今だ研究がないようである．本研究では，ノンパラメトリック回帰のアプローチから，埋め込み 1 次元曲線の同時信頼領域の構築問題に取り組んだ．

問題:　$I=[a,b]\subset\mathbb{R}$ を閉区間とし，$\varphi_i : I\to\mathbb{R}$ $(i=1,\dots,d)$ を滑らかな関数とする．このとき，$\mathbb{R}^d$ に埋め込まれた 1 次元曲線を

$$\boldsymbol{\varphi}(t)=\begin{bmatrix}\varphi_1(t)\\ \vdots\\ \varphi_d(t)\end{bmatrix}\in\mathbb{R}^d$$

で表す．確率ベクトル $\boldsymbol{Y}\in\mathbb{R}^d$ と共変量 $T\in I$ の組 $(T,\boldsymbol{Y})$ からの n 個の観測値 $(T_1,\boldsymbol{Y}_1),\dots,(T_n,\boldsymbol{Y}_n)$ に基づき 1 次元曲線 $\boldsymbol{\varphi}$ の同時信頼領域を構築したい．つまり，$M_{\boldsymbol{\varphi}}=\left\{\boldsymbol{\varphi}(t)\in\mathbb{R}^d \mid t\in I\subset\mathbb{R}\right\}$ としたとき，十分小さい $\alpha>0$ に対して，

$$\mathbb{P}\left(\mathfrak{D}\supset M_{\boldsymbol{\varphi}}\right)\geq 1-\alpha$$

となるような有界閉領域 $\mathfrak{D}\subset\mathbb{R}^d$ の構築を目指す．

設定:　1 次元曲線の推定の関連研究である Hastie and Stuetzle (1989) を参照しつつ，モデルの仮定を行う．$\mathbb{R}^d$ に埋め込まれた 1 次元曲線 $\boldsymbol{\varphi}$ に対して，任意の点 t における勾配ベクトル $\boldsymbol{\varphi}'(t)$ を $\boldsymbol{\varphi}'(t)=[\varphi_1'(t)\ \cdots\ \varphi_d'(t)]^T$ で表す．また，$\boldsymbol{n}_1(t),\dots,\boldsymbol{n}_{d-1}(t)$ を $\{\boldsymbol{\varphi}'(t)/\left\|\boldsymbol{\varphi}'(t)\right\|,\boldsymbol{n}_1(t),\dots,\boldsymbol{n}_{d-1}(t)\}$ が $\mathbb{R}^d$ の正規直交基底となるようにとる．さらに，T を I 上の密度 f_T に従う確率変数とし，$\boldsymbol{V}=[V_1\ \cdots\ V_{d-1}]^T$ を原点中心半径 r の $(d-1)$ 次元球 B_r^{d-1} 上の密度 $f_{\boldsymbol{V}}$ に従う確率ベクトルとする．

$(T_1,\boldsymbol{V}_1),\dots,(T_n,\boldsymbol{V}_n)\overset{\text{i.i.d.}}{\sim} f_T\times f_V$ と $E[\boldsymbol{V}_1]=\boldsymbol{0}$ を仮定する．確率ベクトル $\boldsymbol{Y}$ の d 次元標本 $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n$ として，

$$\boldsymbol{Y}_i=\boldsymbol{\varphi}(T_i)+N(T_i)\boldsymbol{V}_i$$

を考える．ただし，$\boldsymbol{Y}_i=[Y_{i1}\ \cdots\ Y_{id}]^T,\boldsymbol{V}_i=[V_{i1}\ \cdots\ V_{id-1}]^T,N(t)=[\boldsymbol{n}_1(t)\ \cdots\ \boldsymbol{n}_{d-1}(t)]$ である．

同時信頼領域の構築:　共変量 $\boldsymbol{T}=[T_1\ \cdots\ T_n]^T$ と $\boldsymbol{Y}$ の第 j-成分のデータ $\widetilde{\boldsymbol{Y}}_j=[Y_{1j}\ \cdots\ Y_{nj}]^T$ との組 $(\boldsymbol{T},\widetilde{\boldsymbol{Y}}_i)$ $(i=1,\dots,d)$ に基づき，t における $\varphi_i(t)$ の局所線形推定量 $\widehat{\varphi}_i(t)(i=1,\dots,d)$ が得られる．これらを

並べることにより $\boldsymbol{\varphi}(t)$ の推定量 $\widehat{\boldsymbol{\varphi}}(t)$ が構築される．さらに $\widehat{\boldsymbol{\varphi}}(t)$ を用いてその勾配ベクトルを求める：

$$\widehat{\boldsymbol{\varphi}}(t) = \begin{bmatrix} \widehat{\varphi}_1(t) \\ \vdots \\ \widehat{\varphi}_d(t) \end{bmatrix}, \quad \widehat{\boldsymbol{\varphi}}'(t) = \frac{d}{dt}\widehat{\boldsymbol{\varphi}}(t).$$

局所線形推定量については Wand and Jones (1995); Fan and Gijbels (1996) などを参照されたい．$\widehat{\boldsymbol{n}}_1(t), \ldots, \widehat{\boldsymbol{n}}_{d-1}(t)$ を $\{\widehat{\boldsymbol{\varphi}}'(t)/\left\|\widehat{\boldsymbol{\varphi}}'(t)\right\|, \widehat{\boldsymbol{n}}_1(t), \ldots, \widehat{\boldsymbol{n}}_{d-1}(t)\}$ が $\mathbb{R}^d$ の正規直交基底となるようにとり，

$$\begin{aligned}
\widehat{\mathfrak{D}}_a(r) &= \left\{ \widehat{\boldsymbol{\varphi}}(a) - R\frac{\widehat{\boldsymbol{\varphi}}'(a+)}{\|\widehat{\boldsymbol{\varphi}}'(a+)\|} + \widehat{N}(a)\boldsymbol{v} \;\middle|\; 0 \le R \le r, \boldsymbol{v} \in B^{d-1}_{\sqrt{r^2-R^2}} \right\}, \\
\widehat{\mathfrak{D}}_J(r) &= \left\{ \widehat{\boldsymbol{\varphi}}(t) + \widehat{N}(t)\boldsymbol{v} \;\middle|\; t \in J, \boldsymbol{v} \in B^{d-1}_{r} \right\}, \\
\widehat{\mathfrak{D}}_b(r) &= \left\{ \widehat{\boldsymbol{\varphi}}(b) + R\frac{\widehat{\boldsymbol{\varphi}}'(b-)}{\|\widehat{\boldsymbol{\varphi}}'(b-)\|} + \widehat{N}(b)\boldsymbol{v} \;\middle|\; 0 \le R \le r, \boldsymbol{v} \in B^{d-1}_{\sqrt{r^2-R^2}} \right\}, \\
\widehat{N}(t) &= \begin{bmatrix} \widehat{\boldsymbol{n}}_1(t) & \cdots & \widehat{\boldsymbol{n}}_{d-1}(t) \end{bmatrix}
\end{aligned}$$

とする．3 つの排反な領域を用いて同時信頼領域 $\widehat{\mathfrak{D}}(r) \subset \mathbb{R}^d$ を

$$\widehat{\mathfrak{D}}(r) = \widehat{\mathfrak{D}}_a(r) \cup \widehat{\mathfrak{D}}_J(r) \cup \widehat{\mathfrak{D}}_b(r)$$

として定義する (図 1 参照)．本講演では $\widehat{\boldsymbol{\varphi}}(t)$ と $\widehat{\mathfrak{D}}(r)$ の理論的性質および，信頼係数 $1-\alpha$ から $\widehat{\mathfrak{D}}(r)$ の r を決定する方法について解説を与え，実データへの適用例についても報告した．

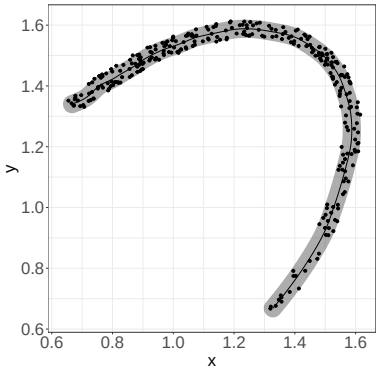


図 1　推定曲線 $\widehat{\boldsymbol{\varphi}}(t)$ から作られた同時信頼領域 $\widehat{\mathfrak{D}}(r)$ の例

参考文献

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. Taylor & Francis.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84, 502–516.

Liu, Y. and Ruppert, D. (2021). Density estimation on a network. *Computational Statistics & Data Analysis*, 156, 107128.

McSwiggan, G., Baddeley, A., and Nair, G. (2017). Kernel density estimation on a linear network. *Scandinavian Journal of Statistics*, 44, 324–345.

Naito, K., Udagawa, J., and Otani, H. (2010). Multidimensional standard curve for the development process of human fetuses. *Statistics in Medicine*, 29, 2235–2245.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall.