科研費シンポジウム

データサイエンスと周辺領域の双方向的理解への挑戦

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)

開催責任者 : 片山 翔太 (慶應義塾大学)
開催日時 : 2022 年 9 月 9 日(金), 9 月 10 日(土)
開催場所 : 慶應義塾大学日吉キャンパス来往舎中会議室 + Zoom (ハイブリッド開催)

内容・目的 : 今やデータサイエンスはその周辺領域ー例えば因果推論, 計量経済, AI・機械学習, バイオインフォマティクス, 情報理論などーにおける共通言語となっており, 各領域でそれぞれ独自の発展も見せています. 本シンポジウムでは, 統計科学の理論・応用に限らず, 周辺領域の研究報告や, 最先端の研究動向・問題提起などをデータサイエンスに関連する限り広く募集します. 各領域における独自的な発想や, 共通項について共有・議論しながら, データサイエンスをさらに深く理解し, それを周辺領域へと還元することを目指します. 若手からベテランまで幅広い層の講演・参加を歓迎します.

プログラム

9月9日 (金)

13:10 ~ 13:15 Opening

13:20 ~ 13:55 対面 森川 耕輔 (大阪大学), Jae Kwang Kim (アイオワ州立大学)
Semiparametric adaptive estimation under informative sampling

13:55 ~ 14:30 対面 米倉 頌人 (千葉大学), 菅澤 翔之助 (東京大学)
Adaptation of the Tuning Parameter in General Bayesian Inference with Robust Divergence

14:30 ~ 15:05 対面 高橋 慎 (法政大学), 山内 雄太 (名古屋大学), 渡部 敏明 (一橋大学), 大森 裕浩 (東京大学)
Realized stochastic volatility models with skew-t distributions

---

15:05 ~ 15:25 休憩

15:25 ~ 16:00 対面 鈴木 洋一 (慶應義塾大学量子コンピューティングセンター)
ノイズのある量子コンピュータにおける量子振幅推定法

16:00 ~ 16:35 対面 藤森 洸 (信州大学)
The Lasso-based principal component analysis for high-dimensional stationary time series

16:35 ~ 17:10 対面 星野 匡郎 (早稲田大学), 柳 貴英 (京都大学)
A Randomization Test for the Specification of Interference Structure

---

17:10 ~ 17:30 休憩

特別講演

17:30 ~ 18:30 Zoom 大津 泰介 (London School of Economics)
Likelihood inference under alternative asymptotics

---

9 月 10 日 (土)

10:00 ~ 10:35[Zoom] 佐々木 俊一 (岩手大学), 川崎 秀二 (岩手大学)

Erdos-Renyi ランダムグラフの大数法則に関連する漸近評価について

10:35 ~ 11:10 [対面] 栗栖 大輔 (横浜国立大学), 深見 陸 (東京大学), 小池 祐太 (東京大学)

深層学習による時系列データの適応的推定

11:10 ~ 11:45 [対面] 寺田 吉壱 (大阪大学), 山本 倫生 (大阪大学)

ベクトル量子化による大規模クラスタリングの近似法とその性質

11:45 ~ 13:10 Lunch

特別講演

13:10 ~ 14:10 [対面] 二宮 嘉行 (統計数理研究所)

傾向スコア解析のための情報量規準および差分の差法への展開

14:10 ~ 14:30 休憩

14:30 ~ 15:05 [対面] 原田 和治 (東京医科大学), 藤澤 洋徳 (統計数理研究所)

外れ値に頑健な因果効果の推定量

15:05 ~ 15:40 [対面] 劉 言 (早稲田大学), 谷口 正信 (早稲田大学), Hernando Ombao (King Abdullah University of Science and Technology)

Statistical Inference for Local Granger Causality

15:40 ~ 16:15 [対面] 山本 倫生 (大阪大学)

弱い共通サポート条件下での確率的介入に基づく因果効果の推定

16:15 ~ 16:20 Closing

Semiparametric Adapative Estimation Under Informative Sampling

大阪大学大学院基礎工学研究科　森川 耕輔
アイオワ州立大学統計学部　Jae Kwang Kim

1. はじめに

標本調査では，標本を抽出した母集団が必ずしも母集団全体を代表せず，偏った標本 (baised sampling) しか得られないという問題が生じる．このような問題は欠測値データ解析や因果推論においても生じるが，標本調査においては特に，データが抽出される確率である包含確率（inclusion probabililty）の情報も推測に用いることができる．本報告では包含確率の情報を有効に用いた，母集団分布に対するセミパラメトリック漸近有効推定量を提案する．

2. 設定と従来法

結果変数を Y, 説明変数を $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ とする．ある無限母集団からの N 個の無作為標本 $\mathcal{F}_N = \{(\boldsymbol{x}_i, y_i)_{i=1}^N\}$ を考える．ここで，最終的に得られる標本は包含確率 π_i $(i = 1, \ldots, n)$ で抽出された n 個の標本 $\{(\boldsymbol{x}_i, y_i)_{i=1}^n\}$ とする．ただし，この包含確率 π_i は説明変数や結果変数，また観測されない変数に依存していても良い (informative sampling) とし，N は未知と既知の両方の場合を考えるが，当面 N は未知とする．また説明の都合上，確率変数 $W_i := \pi_i^{-1}$ $(i = 1, \ldots, n)$ を定義する．ここで推定対象として次の 3 つを考える:

(a) 結果変数の期待値: $\theta = E(Y)$;

(b) 回帰係数: $\mu(\boldsymbol{x}; \theta) = E(Y \mid \boldsymbol{X} = \boldsymbol{x}; \theta)$;

(c) 条件付き密度: $f(y \mid \boldsymbol{x}; \theta)$.

例えば，標本から計算可能な (c) に対する推定量として，次の Hortitz-Thompson 型の推定量 (Horvitz and Thmopson, 1952) が考えられる:

$$\sum_{i=1}^{N} \delta_i W_i S_\theta(X_i, Y_i; \theta) = 0.$$

ここで，δ_i は有限母集団 $\mathcal{F}_N$ から抽出されていれば 1，そうでなければ 0 をとる確率変数であり，$S_\theta(x, y; \theta) = \partial \log f(y \mid x; \theta)/\partial\theta$ は θ のスコア関数である．この推定方程式は不偏推定方程式となっているため，その解は θ に対する一致性を持つが，有効推定量ではない．

3. セミパラメトリックモデルと最適な推定方程式

推定対象 (a)–(c) に対して，興味のあるパラメータ以外の特定を必要としない次のセミパラメトリックモデルを考える:

$$\begin{aligned} f(w, x, y \mid \delta = 1; \theta, \eta_1, \eta_2, \eta_3) &= \frac{P(\delta = 1 \mid w, x, y) f(w \mid x, y; \eta_1) f(y \mid x; \eta_2, \theta) f(x; \eta_3)}{\int P(\delta = 1 \mid w, x, y) f(w \mid x, y; \eta_1) f(y \mid x; \eta_2, \theta) f(x; \eta_3) \mathrm{d}w\mathrm{d}x\mathrm{d}y} \\ &= \frac{w^{-1} f(w \mid x, y; \eta_1) f(y \mid x; \eta_2, \theta) f(x; \eta_3)}{\int w^{-1} f(w \mid x, y; \eta_1) f(y \mid x; \eta_2, \theta) f(x; \eta_3) \mathrm{d}w\mathrm{d}x\mathrm{d}y}. \end{aligned}$$

ただし，η_1, η_2, η_3 はそれぞれ $[w \mid x, y]$, $[y \mid x]$, $[x]$ の確率分布を規定する未知の無限次元パラメータである．ここで推定対象が (c) の場合は $f(y \mid x; \eta_2, \theta) = f(y \mid x; \theta)$ となることに注意する．

本研究では，本設定下におけるセミパラメトリック漸近有効下限に到達するセミパラメトリック漸近有効推定量を導出した．セミパラメトリック有効下限とは，セミパラメトリックモデルにおける推定量の漸近分散の下限でありパラメトリックモデルにおける Cramér-Rao の下限のようなものである．N が未知の場合，次の推定方程式に基づいた推定量はセミパラメトリック漸近有効推定量となる:

$$S_{\text{eff}} = \sum_{i=1}^{N} \delta_i W_i D^*_{\text{eff}}(X_i, Y_i) = 0. \qquad (*)$$

ここで，D^*_{eff} は推定対象ごとに以下の形で表される．

(a) $D^*_{\text{eff}}(X, Y) = \theta - Y$;

(b) $D^*_{\text{eff}}(X, Y) = \dfrac{1}{E(W\varepsilon^2 \mid X)} \dfrac{\partial}{\partial\theta}\mu(X;\theta)$;

(c) $D^*_{\text{eff}}(X, Y) = \bar{\pi}(X, Y)\left[S_\theta(X, Y) - \dfrac{E\{\bar{\pi}(X, Y)S_\theta(X, Y) \mid X\}}{E\{\bar{\pi}(X, Y) \mid X\}}\right]$.

ただし，$\varepsilon = Y - \mu(X;\theta)$, $\bar{\pi}(x, y) = 1/E(W \mid x, y)$ である．ここで，推定対象 (b) に対する推定関数は，Kim and Skinner (2013) で導出されているものと同一であることに注意する．

また，母集団サイズ N が既知である場合の最適な推定方程式は，次の形で表される:

$$S_{\text{eff}} = \sum_{i=1}^{N} \left\{\delta_i W_i \tilde{D}^*_{\text{eff}}(X_i, Y_i) + (1 - \delta_i W_i)\tilde{c}^*_{\text{eff}}\right\} = 0. \qquad (**)$$

$\tilde{D}^*_{\text{eff}}$ と $\tilde{c}^*_{\text{eff}}$ の詳細については当日報告する．母集団サイズ N が既知である場合，その情報を用いた $(**)$ の解を用いることで，$(*)$ の解である Kim and Skinner (2013) の推定量を優越する推定量が構成可能となる．

4. 適応的推定量

最適な推定方程式 $(*)$ は未知関数を含むため，作業用モデルを要する．例えば，推定対象が (c) の場合，$\bar{\pi}(x, y) = 1/E(W \mid x, y)$ という関数をモデリングし推定する必要がある．そこで，Ferrari and Cribari-Neto (2004) のベータ回帰のアイディアを利用した柔軟なモデリング方法を提案する．

未知関数を作業用モデルを用いて推定したものを用いた最適な推定方程式 $(*)$ および $(**)$ の θ に対する解は，作業用モデルが正しい場合，セミパラメトリック漸近有効推定量となり，仮に誤っていたとしても一致性のある推定量となる．当日は，1999 年に実施された Canadian Workplace and Employee Survey データ (Fuller, 2009) に提案手法を適用し，Horvitz-Thompson 型の推定量と比較することで，その有用性を示す．

5. 参考文献

Ferrari, S. and Chibari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 407–419.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken: John Wiley & Sons, Inc.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Kim, J. K. and Skinner, C. J. (2013). Weighting in Survey Analysis Under Informative Sampling. *Biometrika*, **100**, 385–398.

Morikawa, K. and Kim, J. K. (2022). Semiparametric adaptive estimation under informative sampling. arXiv: 2208.06039.

Adaptation of the Tuning Parameter in General Bayesian Inference with Robust Divergence

Shouto Yonekura[1,3] and Shonosuke Sugasawa[2,3]

[1]Graduate School of Social Sciences, Chiba University
[2]Center for Spatial Information Science, The University of Tokyo
[3]Nospare Inc.

August 28, 2022

We introduce a novel methodology for robust Bayesian estimation with robust divergence (e.g., density power divergence or γ-divergence), indexed by tuning parameters. It is well known that the posterior density induced by robust divergence gives highly robust estimators against outliers if the tuning parameter is appropriately and carefully chosen. In a Bayesian framework, one way to find the optimal tuning parameter would be using evidence (marginal likelihood). However, we theoretically and numerically illustrate that evidence induced by the density power divergence does not work to select the optimal tuning parameter since robust divergence is not regarded as a statistical model. To overcome the problems, we treat the exponential of robust divergence as an unnormalisable statistical model, and we estimate the tuning parameter by minimising the Hyvarinen score. We also provide adaptive computational methods based on sequential Monte Carlo (SMC) samplers, enabling us to obtain the optimal tuning parameter and samples from posterior distributions simultaneously. The empirical performance of the proposed method through simulations and an application to real data are also provided.

The talk is organised as follows. We first set up the framework and then show theoretically and numerically that evidence induced by density power divergence to select the tuning parameter. Instead, we propose to estimate it based on the H-score (Hyvärinen, 2005; Dawid et al., 2015) and characterise its asymptotic behaviour. As mentioned earlier,

our method involves functions for which it is difficult to obtain an analytic representation. Therefore, we develop an adaptive and efficient Markov chain Monte Carlo (MCMC) algorithm based on SMC samplers (Del Moral et al., 2006). Numerical applications of the developed methodology are also provided, then conclusions and directions for future research are stated.

References

Dawid, A. P., Musio, M., et al. (2015). Bayesian model selection based on proper scoring rules. *Bayesian analysis*, 10(2):479–499. 1

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436. 2

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4). 1

Realized Stochastic Volatility Models with Skew-t Distributions

Makoto Takahashi*[1], Yuta Yamauchi[2], Toshiaki Watanabe[3], and Yasuhiro Omori[4]

[1]Faculty of Business Administration, Hosei University

[2]Graduate School of Economics, Nagoya University

[3]Center for the Promotion of Social Data Science Education and Research, Hitotsubashi University

[4]Faculty of Economics, University of Tokyo

Abstract

Forecasting volatility and quantiles of financial returns is essential to measure the financial tail risk such as value-at-risk and expected shortfall. There are two important aspects of volatility and quantile forecasts: the distribution of financial returns and the estimation of the volatility. Building on the traditional stochastic volatility model, the realized stochastic volatility model incorporates realized volatility as the precise estimator of the volatility. Using three types of skew-t distributions, the model is extended to capture the well-known characteristics of the return distribution, namely skewness and heavy tails. In addition to the normal and Student's t distributions, included as the special cases of the skew-t distributions, two of them contain the skew-normal, and hence allow more flexible modeling of the return distribution. The Bayesian estimation scheme via a Markov chain Monte Carlo method is developed and applied to major stock indices. The empirical study using the US and Japanese stock indices data suggests that incorporating both skewness and heavy tail to daily returns is important for volatility and quantile forecasts.

1 Summary

Financial volatility, defined as a standard deviation or variance of asset returns, changes stochastically over time and hence it is important to forecast time-varying volatility for financial risk management. Time-varying volatility used to be estimated and forecasted by two classes of time-series models using financial returns: the generalized autoregressive conditional heteroskedasticity (GARCH) methodologies of Engle (1982) and Bollerslev (1986), and the stochastic volatility (SV) model (Taylor, 1994). These models are consistent with volatility clustering (high volatility persistence), and have been extended to accommodate a phenomenon called volatility asymmetry or leverage effect, i.e., the negative correlation between today's return and tomorrow's volatility that is observed in stock markets. One such extension is the exponential GARCH (EGARCH) model proposed by Nelson (1991).

In recent years, realized volatility (RV) has replaced these models for volatility estimation. Daily RV is calculated as the sum of squared intraday returns over a day. Andersen and Benzoni (2009) and McAleer and Medeiros

*Corresponding author. Email: m-takahashi@hosei.ac.jp

(2008) provide RV reviews. To predict volatility using RV, we must model its dynamics. Many studies such as Andersen et al. (2003) have documented that daily RV may follow a long-memory process; thus, they use autoregressive fractionally integrated moving average (ARFIMA) models (see Beran (1994) for long-memory and ARFIMA models). The heterogeneous autoregressive (HAR) model proposed by Corsi (2009) is used more widely for the dynamics of RV. Although it is not a long-memory model, it is known to approximate a long-memory process well.

Although ARFIMA and HAR models have been shown to significantly improve volatility forecasts relative to GARCH and SV models, the RV is subject to the bias caused by market microstructure noise and non-trading hours. There are some methods for mitigating the bias in RV: the multiscale estimators by Zhang et al. (2005) and Lan Zhang (2006), the pre-averaging approach by Jacod et al. (2009), and the realized kernel (RK) method by Barndorff-Nielsen et al. (2008, 2009). See, for example, Aït-sahalia and Mykland (2009), Ubukata and Watanabe (2014), and Liu et al. (2015) for details.

To account for the bias in RV, two classes of hybrid models have been proposed. One is the realized SV (RSV) model proposed by Takahashi et al. (2009), Dobrev and Szerszen (2010), and Koopman and Scharth (2013), while the other is the realized GARCH (RGARCH) model and realized EGARCH (REGARCH) models proposed by Hansen et al. (2012) and (Hansen and Huang, 2016), respectively. In the existing literature, the volatility forecasting abilities of these volatility models have been compared within each class, while a comprehensive comparison across different classes of volatility models has not been conducted extensively. An exception is Takahashi et al. (2021) which compare volatility forecasting abilities across different classes of volatility models including the SV, EGARCH, HAR, RSV, and REGARCH models.

To measure the financial tail risk such as value-at-risk (VaR) and expected shortfall (ES), it is important to forecast not only volatility but also quantiles of financial returns. The empirical return distribution is more peaked and has heavier tails than the normal distribution. Such a heavy tail or leptokurtosis can be partly explained by the time-varying volatility, but the return distribution conditional on the volatility may still be leptokurtic. Moreover, the return distribution may also be skewed. Therefore, several types of skew Student's t distributions have been used in the volatility models. For example, the generalized hyperbolic (GH) skew Student's t distribution (Aas and Haff, 2006) has been used for the SV model (Nakajima and Omori, 2012; Leão et al., 2017) and the RSV model (Takahashi et al., 2016).

Along these lines, the RSV model is extended with three types of skew- t distributions and Bayesian estimation scheme via a Markov chain Monte Carlo method is developed. The skew-t distributions include the ones proposed by Azzalini (1985) and Fernández and Steel (1998) as well as the GH skew-t distribution. In the empirical analysis using the Dow Jones Industrial Average (DJIA) and the Nikkei 225 (N225) data, the volatility, VaR, and ES forecasts of the RSV models as well as the SV, EGARCH, and REGARCH models are estimated. The volatility forecasts are evaluated by mean squared error and quasi likelihood loss functions, while the VaR and ES forecasts are jointly evaluated by the loss function proposed by Fissler and Ziegel (2016) with the specification suggested by Patton et al. (2019). The predictive ability test (Giacomini and White, 2006) for these loss functions and the model confidence set (Hansen et al., 2011) are implemented to compare the forecasting performance. The results show that the RV improves the forecasting performance significantly, the RSV models are better than the other models, and the skew-t distributions improve the forecasting performance. These results confirm and complement the previous studies mentioned above.

ノイズのある量子コンピュータにおける量子振幅推定法

慶應義塾大学量子コンピューティングセンター
鈴木洋一

量子コンピュータは, 様々な科学技術分野において, 古典的な計算と比較し, 効率的な計算を可能にすると期待されている. 現在, 量子アルゴリズムが実際の量子コンピュータで実行できる環境が整いつつある一方, 現在の実量子コンピュータはいわゆる Noisy Intermidiate-Scale Quantum (NISQ) デバイスであり, ゲート演算数や利用可能な量子ビットの数など, 実用上いくつかの制約がある. そこで, これらの制約を考慮したカスタムサブルーチンがいくつか提案されている.

本講演では、化学, 金融, 機械学習など, 様々な応用分野で量子計算の中核を担う量子振幅推定アルゴリズムに焦点を当てる. 特に, 量子振幅推定によるモンテカルロサンプリングの効率化は, これらの応用の中核をなすものであり, その重要性に鑑み, 前述の方向性に沿って, NISQ デバイスで実行可能な新しい量子振幅推定アルゴリズムの開発が近年盛んに行われている.

量子振幅推定は, その名の通り本質的にパラメータ推定問題であり, 量子状態 $|\psi\rangle = \sin\theta\,|\mathrm{good}\rangle + \cos\theta\,|\mathrm{bad}\rangle$ に含まれる, 未知パラメータ θ を推定する問題である. ここで $|\mathrm{good}\rangle$ と $|\mathrm{bad}\rangle$ は, 直行する量子状態ベクトルであり, $|\mathrm{good}\rangle$ 状態が観測される確率 $p^{(0)}_{\mathrm{good}}(\theta)$ は, $|\mathrm{good}\rangle$ 状態の確率振幅の二乗, 即ち $\sin^2\theta$ で与えられる. 従来の量子振幅推定は, 量子振幅増幅 (amplitude amplification) と, 位相推定 (phase estimation) を組み合わせ実現される [1]. 量子振幅増幅では, 増幅演算子 G が, 量子状態 $|\psi\rangle = \sin\theta\,|\mathrm{good}\rangle + \cos\theta\,|\mathrm{bad}\rangle$ に m_k 回に作用した際, $G^{m_k}\,|\psi\rangle = \sin((2m_k+1)\theta)\,|\mathrm{good}\rangle + \cos((2m_k+1)\theta)\,|\mathrm{bad}\rangle$ のように, ターゲットとなる状態 $|\mathrm{good}\rangle$ を観測する確率を $p^{(k)}_{\mathrm{good}}(\theta) = \sin^2((2m_k+1)\theta)$ のように増幅することができる. ここで, m_k は, $k = 0, 1, 2, \cdots, M$ における増幅演算の実行回数である. 一方, 量子振幅推定のもう 1 つの構成要素である位相推定は, 大量の量子ゲート操作を必要とするため, NISQ デバイスでの実行は現実的ではない. そこで我々は, 量子振幅増幅と, 古典的統計手法 (最尤推定) を組み合わせ, 量子振幅推定を行う方法を提案した [2]. 図 1 に提案手法のスキームを示す. 具体的には, 異なる回数 (図 1 では, m_0, m_1, $\cdots$, m_M 回) の量子振幅増幅を行なって得たれた結果を組み合わせ最尤推定を行い, パラメータ推定を行う. この手法を用いてモンテカルロ積分を行ったところ, 古典モンテカルロ積分と比較し, 高速に実行可能であること (二乗加速があること) を確認し, 大幅に量子ゲート操作を削減できることを示した.

更に, 我々はこの方法を拡張し, 現実的なノイズ効果を考慮に入れ, 超伝導 IBM 量子デバイスを用いた実験の解析を行った. 尚, 詳細については講演で述べる. 実験結果の解析にあたり, 我々は次のようなモデルを導入した. $|\mathrm{good}\rangle$ 状態を観測する確率が, $p^{(k)}_{\mathrm{good}}(\theta, \beta_k)$ 及び, $q^{(k)}_{\mathrm{good}}(\theta, \beta_k)$ である 2 種類の量子回路を用意する. 但し, β_k はノイズの効果を表すパラメータであり, $p^{(k)}_{\mathrm{good}}(\theta, \beta_k) = 1/2 - 1/2\beta_k \cos(2(2m_k+1)\theta)$, $q^{(k)}_{\mathrm{good}}(\theta, \beta_k) = 1/2 - 1/2\beta_k \cos(2(2m_k-3)\theta)$ とした. β_k は, $0 \le \beta \le 1$ の実数であり, $\beta_k = 1$ が ノイズのない場合に対応する. 我々

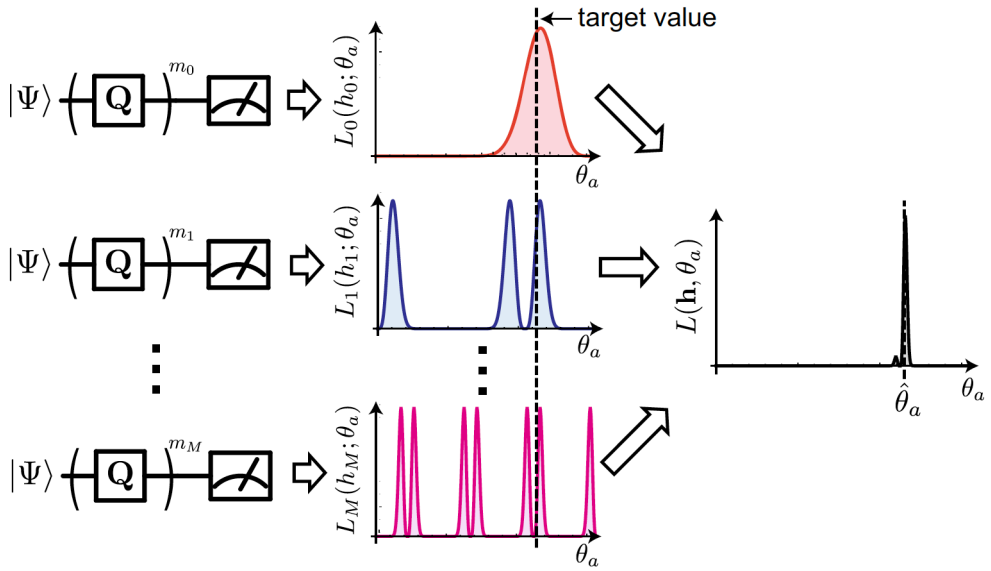

図 1　最尤推定を用いる量子振幅推定スキーム．異なる回数 $(m_0, m_1, \cdots, m_M)$ の量子振幅増幅 (左)．それぞれの実験結果から得られる尤度関数 (中央)．異なる回数の量子振幅増幅を組み合わせた尤度関数 (右)．

は，この $p^{(k)}_{\mathrm{good}}(\theta, \beta_k)$ 及び，$q^{(k)}_{\mathrm{good}}(\theta, \beta_k)$，但し $k = 0, 1, 2, \cdots, M$ を用いて尤度関数を構成し，最尤推定を行った．ここで問題となるのは，確率分布モデルに多くのノイズパラメータ β_k，$k = 0, 1, 2, \cdots, M$(局外パラメータ) が含まれるため，最適なパラメータの推定が効率的でないという本質的な問題がある．この問題に対して我々は，パラメータ直交化法 (parameter orthogonalization method)[3, 4] を適用し，他のノイズパラメータ β_k，$k = 0, 1, 2, \cdots, M$，を除去して，目的の振幅パラメータ θ のみの推定を行った [5]．講演では，本手法の詳細や，実際の超伝導量子デバイスを用いた実験の解析結果などについて紹介する予定である．

参考文献

[1] G. Brassard, P. Høyer, M. Mosca, A. Tapp, Quantum amplitude amplification and estimation. Contemp. Math. Ser. Millenn. 305, 53-74 (2002)

[2] Y. Suzuki, S. Uno, R. Raymond, T. Tanaka, T. Onodera, and N. Yamamoto, Amplitude estimation without phase estimation, Quant. Info. Proc. 19, 75 (2020).

[3] D. R. Cox and N. Reid, Parameter orthogonality and approximate conditional inference, J. R. Stat. Soc., Ser. B 49(1), 1 (1987).

[4] J. Suzuki, Nuisance parameter problem in quantum estimation theory: Tradeoff relation and qubit examples, J. Phys. A: Math. Theor. 53, 264001 (2020).

[5] T. Tanaka, S. Uno, T. Onodera, N. Yamamoto, and Y. Suzuki, Noisy quantum amplitude estimation without noise estimation, Phys. Rev. A 105, 012411 (2022).

The Lasso-based principal component analysis for high-dimensional stationary time series

藤森　洸

信州大学　経法学部

1　序論

高次元の設定において，従来の主成分分析は精度が悪いことが知られており，特に，スパースな主成分ベクトルの推定には，より良い性質を持った推定量が，主に独立標本の設定において提案されてきた．本講演では Gaussian よりも裾の重いケースを含む高次元定常時系列に対して Lasso 型のスパース主成分分析手法を適用し，得られる推定量の漸近的な挙動について議論する．さらに，従来の主成分分析手法や独立標本に対する Lasso 型主成分分析との比較を行っていく．

2　主結果

$\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ を確率空間 $(\Omega, \mathcal{F}, P)$ 上の $\mathbb{R}^p$-値, 平均 0 の定常過程とする．観測系列 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, $n \in \mathbb{N}$ に対して，次の $p \times p$ 標本共分散行列及び，共分散行列を考える．

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{X}_t \boldsymbol{X}_t^\top, \quad \Sigma_0 = E[\boldsymbol{X}_t \boldsymbol{X}_t^\top].$$

以下，Σ_0 の最大固有値 $\phi_{\max}^2$ に対応する，規格化された第一主成分ベクトル $\boldsymbol{q}^0$ の推定問題を考える．推定の対象となるパラメータは，$\boldsymbol{\beta}^0 = \phi_{\max}\boldsymbol{q}^0$ として特徴づけることができるが，これは，次の最適化問題の解として与えられる：

$$\boldsymbol{\beta}^0 = \arg\min_{\boldsymbol{\beta}} \frac{1}{4}\|\Sigma_0 - \boldsymbol{\beta}\boldsymbol{\beta}^\top\|_F^2 \iff \Sigma_0\boldsymbol{\beta}^0 = \|\boldsymbol{\beta}^0\|_2^2\boldsymbol{\beta}^0,$$

ただし，$\|\cdot\|_F$ は行列のフロベニウスノルムである．さて，$\boldsymbol{\beta}^0$ の推定量として，次の Lasso 型推定量を定義する (van de Geer (2016))：

$$\hat{\boldsymbol{\beta}}_n^1 := \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \left\{ \frac{1}{4}\|\hat{\Sigma}_n - \boldsymbol{\beta}\boldsymbol{\beta}^\top\|_F^2 + \lambda_1\|\boldsymbol{\beta}\|_1 \right\}, \quad \mathcal{B} := \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_2 \le \eta\}.$$

ただし，λ_1 は正則化パラメータであり，η は適当な定数である．本講演では，適当な定数 $q > 0$ に対して，$p = qn$ が成立することを仮定し，推定量 $\hat{\boldsymbol{\beta}}_n^1$ の挙動について議論する．

時系列における従属構造の影響は，混合係数を用いて評価することになる．例えば，ガウス型過程 $\{\boldsymbol{X}\}_{t\in\mathbb{Z}}$ に対しては次の ρ-混合係数を用いた評価を行う．

$$\begin{aligned}\rho(l) \;&:=\; \sup\{|\operatorname{Cov}(f(\boldsymbol{X}_t), g(\boldsymbol{X}_{t+l}))| : \\ &\qquad E[f] = E[g] = 0,\ E[f^2] = E[g^2] = 1\}, \quad l \in \mathbb{Z}.\end{aligned}$$

ρ-混合係数を用いることで，ガウス型確率過程の標本共分散行列の推定誤差評価に際して，Hanson-Wright の不等式と呼ばれる集中不等式を利用することができる．いま，$\boldsymbol{\beta}^0$ の非零成分の個数を s_0 とする．このとき，共分散行列の固有値に対するスパイク性に関する条件や混合係数に関する条件をはじめとする適当な条件の下で，推定量の収束レートは次のように導出される．

Theorem 2.1. 確率過程 $\{\boldsymbol{X}_t\}$ はガウス型であるとし，その ρ-混合係数を $\rho(\cdot)$ とする．, $\sum_{l=0}^{\infty}\rho(l) < \infty$, $\phi_{\max}^2 = O(1)$, $n \to \infty$ を仮定し，正則化パラメータ λ_1 は次を満たすとする：

$$\lambda_1 \asymp \sqrt{s_0 \frac{\log p}{n}}.$$

このとき，適当な正則条件の下で，次が成立する．

$$\|\hat{\boldsymbol{\beta}}_n^1 - \boldsymbol{\beta}^0\|_1 = O_p\left(s_0^{3/2}\sqrt{\frac{\log p}{n}}\right), \quad n \to \infty$$

$$\|\hat{\boldsymbol{\beta}}_n^1 - \boldsymbol{\beta}^0\|_2^2 = O_p\left(s_0^3 \frac{\log p}{n}\right), \quad n \to \infty.$$

上述の定理において，ρ-混合係数の総和に関する条件を仮定しているが，これは推定量の非漸近的な評価を行った際に，時系列の従属構造の影響が現れることに由来している．すなわち，独立標本のケースとは異なった挙動を観察することができる．

本講演ではガウス型確率過程だけではなく，より裾の重いケースを含むクラスである sub-Weibull 型確率過程をより強い混合係数に対する条件の下で扱い，Lasso 型推定量の漸近的挙動を比較する．なお，講演内容は，Fujimori et al. (2016) に基づく．

References

Fujimori, K., Goto, Y., Liu, Y., Taniguchi, M. (2021). Sparse principal component analysis for high-dimensional stationary time series. *Submitted.*

van de Geer, S. A. (2016). *Estimation and testing under sparsity.* Springer.

A Randomization Test for the Specification of Interference Structure

Tadao Hoshino* and Takahide Yanagi†

Recent studies in the causal inference literature have highlighted the importance of accounting for treatment spillovers from other units via empirical applications in many fields, including political science, epidemiology, education, economics, etc. In a most general case, the treatment of all units in the population may affect one's outcome. The principle of causal inference is to compare similar individuals with different treatment status. However, when treatment spillovers are present, since each individual has his/her unique social network with others, no such similar counterparts might exist. Thus, it is generally not possible to identify meaningful causal parameters without some simplifying assumptions on the interference structure. Indeed, Imbens and Rubin (2015) states *"causal inference is generally impossible without such assumptions, and thus it is critical about their content and their justifications."*

To address this problem, a common approach in the literature is to assume the existence of an *exposure mapping* that summarizes the impacts from others' treatments into lower dimensional sufficient statistics. For example, Hong and Raudenbush (2006) study the impact of school retention on later academic performance, assuming that other students' retention may affect one's own performance only through whether their school has a higher or lower retention rate. As another example, Leung (2020) considers a treatment spillover model in which one's potential outcome is a function not only of one's treatment but also of the number of treated neighbors and that of all neighbors.

Clearly, if the chosen exposure mapping is not appropriate, the resulting causal inference may be misleading (e.g., failure of detecting treatment spillovers).[1] As stated in the quote from Imbens and Rubin (2015), in order for the obtained spillover effects to be valid, it is required to justify the specification of the exposure mapping. A direct approach to accomplish this is to statistically test the specification, which is exactly the aim of this study. That is, we propose a randomization specification test for the form of general exposure mappings.

We are not the first to develop this type of randomization test; see, e.g., Athey *et al.* (2018), Basse *et al.* (2019), and Puelz *et al.* (2022). The major difference between our approach and theirs lies in the construction of the test statistic. The aforementioned studies are almost agnostic about the construction of test statistics with sufficient power, or they only provide examples that work well

*School of Political Science and Economics, Waseda University, 1-6-1 Nishi-waseda, Shinjuku-ku, Tokyo 169-8050, Japan. Email: thoshino@waseda.jp.

†Graduate School of Economics, Kyoto University, Yoshida Honmachi, Sakyo, Kyoto, 606-8501, Japan. Email: yanagi@econ.kyoto-u.ac.jp.

[1]Some recent studies uncover under what conditions one can estimate meaningful causal parameters even when the exposure mapping is misspecified or is not explicitly specified (Hoshino and Yanagi, 2021; Sävje *et al.*, 2021; Leung, 2022). A common finding in these studies is that, only if the network dependence is sufficiently weak, we can identify some composite causal parameters.

for certain situations. In contrast, by introducing the notion of *coarseness* of exposure mappings, we propose a method that automatically derives test statistics equipped with reasonable power in most situations.

The basic idea is as follows. Suppose that we have two exposure mappings E^0 and E^1. We say that E^0 is *coarser* than E^1 if there exists a surjective mapping c such that $c(E^1) = E^0$. For example,

$$E^0 = \mathbf{1}\{\#\text{ treated neighbors} > 0\}, \qquad E^1 = \#\text{ treated neighbors}$$

Now, suppose that we would like to test the following null hypothesis:

$$\mathbb{H}_0\text{: } E^0 \text{ is a correct exposure mapping.}$$

Then, if $\mathbb{H}_0$ is true, E^1 is also a correct exposure mapping. This enables us to perform a randomization test by shuffling the treatment assignment while keeping the value of E^0 fixed but allowing the value of E^1 to alternate. Under $\mathbb{H}_0$, the distributions of potential outcomes for different E^1-values should be identical; otherwise, $\mathbb{H}_0$ should be rejected. The result of numerical simulations shows that our approach works satisfactorily compared to the existing method.

References

Athey, S., Eckles, D., and Imbens, G.W., 2018. Exact p-values for network interference, *Journal of the American Statistical Association*, 113 (521), 230–240.

Basse, G.W., Feller, A., and Toulis, P., 2019. Randomization tests of causal effects under interference, *Biometrika*, 106 (2), 487–494.

Hong, G. and Raudenbush, S.W., 2006. Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data, *Journal of the American Statistical Association*, 101 (475), 901–910.

Hoshino, T. and Yanagi, T., 2021. Causal inference with noncompliance and unknown interference, *arXiv:2108.07455*.

Imbens, G.W. and Rubin, D.B., 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

Leung, M., 2020. Treatment and spillover effects under network interference, *The Review of Economics and Statistics*, 102 (2), 368–380.

Leung, M., 2022. Causal inference under approximate neighborhood interference, *Econometrica*, 90 (1), 267–293.

Puelz, D., Basse, G., Feller, A., and Toulis, P., 2022. A graph-theoretic approach to randomization tests of causal effects under general interference, *Journal of the Royal Statistical Society: Series B*, 84 (1), 174–204.

Sävje, F., Aronow, P.M., and Hudgens, M.G., 2021. Average treatment effects in the presence of unknown interference, *The Annals of Statistics*, 49 (2), 673–701.

LIKELIHOOD INFERENCE UNDER ALTERNATIVE ASYMPTOTICS

TAISUKE OTSU

In this talk, I discuss a unified inference framework for various alternative or non-standard asymptotic inference problems, including sparse network, degeneracy under multiway data, small bandwidth for semiparametric inference, many weak instruments, and many covariates asymptotics. The talk is based on three papers

(1) Empirical likelihood for network data (with Yukitoshi Matsushita)
(2) Multiway empirical likelihood (with Harold Chiang and Yukitoshi Matsushita)
(3) Jackknife empirical likelihood: small bandwidths, sparse network and high-dimensional asymptotics (with Yukitoshi Matsushita)

The first paper is summarized as follows. Analysis on network data is becoming increasingly important in various fields of data science, and the literature on statistical modelling and estimation algorithms for networks is rapidly growing. However, general statistical inference methods for networks are still less developed. This article develops concept of nonparametric likelihood for network data based on the network moments, and proposes general inference methods by adapting the theory of jackknife empirical likelihood. Our methodology can be used not only to conduct inference on population network moments and parameters in network formation models, but also to implement goodness-of-fit testing, such as testing block size for stochastic block models. Theoretically we show that the jackknife empirical likelihood statistic loses its asymptotic pivotalness under the sparse network asymptotics and develop a modified statistic which converges to a chi-squared distribution under both the sparse and dense network asymptotics.

The second paper is summarized as follows. This paper develops a general methodology to conduct statistical inference for observations indexed by multiple sets of entities. We propose a novel multiway empirical likelihood statistic that converges to a chi-square distribution under the non-degenerate case, where corresponding Hoeffding type decomposition is dominated by linear terms. Our methodology is related to the notion of jackknife empirical likelihood but the leave-out pseudo values are constructed by leaving out columns or rows. We further develop a modified version

of our multiway empirical likelihood statistic, which converges to a chi-square distribution regardless of the degeneracy, and discover its desirable higher-order property compared to the t-ratio by the conventional Eicker-White type variance estimator. The proposed methodology is illustrated by several important statistical problems, such as bipartite network, two-stage sampling, generalized estimating equations, and three-way observations.

The third paper is summarized as follows. This paper sheds light on inference problems for statistical models under alternative or nonstandard asymptotic frameworks from the perspective of jackknife empirical likelihood (JEL). Examples include small bandwidth asymptotics for semiparametric inference, many covariates asymptotics for regression models, many-weak instruments asymptotics for instrumental variable regression, and sparse network asymptotics. We first establish Wilks' theorem for the JEL statistic on a general semiparametric inference problem under the conventional asymptotics. We then show that the JEL statistics lose asymptotic pivotalness under the above nonstandard asymptotic frameworks, and argue that these phenomena are understood as emergence of Efron and Stein's (1981) bias of the jackknife variance estimator in the *first* order. Finally we propose a modification of JEL to recover asymptotic pivotalness under both the conventional and nonstandard asymptotics. Our modification works for all above examples and provides a unified framework to investigate nonstandard asymptotic problems.

Erdős-Rényi ランダムグラフの大数法則に関連する漸近評価について

岩手大学総合科学研究科理工学専攻　佐々木俊一
岩手大学理工学部　川崎秀二

1　はじめに：Erdős-Rényi ランダムグラフ

ランダムグラフ（以下，RG）は n 個のノード（頂点）とそれらの間を繋ぐエッジ（辺）から成り，ノード間に確率 $p \in (0,1)$ でエッジが形成される．1 つのノードがエッジにより連結するノード数は非負整数値の確率変数で，その期待値を $\lambda > 0$ とすると，$\lambda = np$ の関係が仮定される．RG は $\lambda = 1$ を境界として，次のような相転移を呈する事が知られている：

$$\begin{cases} \lambda < 1: & \text{劣臨界} \cdots \text{確率 } 1 \text{ で有限ノード数 } (n < \infty) \text{ の系となる（「絶滅」確率 } \eta = 1) \\ \lambda = 1: & \text{臨界} \cdots \text{条件により} \eta = 1 \text{ あるいは} \eta = 0 \text{ の系となる} \\ \lambda > 1: & \text{優臨界} \cdots \text{確率 } \zeta = 1 - \eta \text{ で無限ノード数 } (n \to \infty) \text{ の系となる（「生存」確率 } \zeta > 0) \end{cases}$$

RG のうち特に，1 つのノードからエッジにより連結するノードの数が独立同分布で，それぞれ Poisson 分布 $\mathrm{Pois}(\lambda)$ に従うものを **Erdős-Rényi** RG（ERG）と呼ぶ．本研究では，ERG で $n \to \infty$ の時の漸近的性質について調べる．

また一般に，RG においてエッジによって連結したノードの集合 (**クラスター**) が複数個生成される．全ノード数 n の ER RG において生成された複数のクラスターのうち，ノード数が最も多いクラスターを最大クラスターと呼び，$\mathcal{C}_{\max}$ と表す．$|\mathcal{C}_{\max}|$ は最大クラスターのノード数である．

優臨界での最大クラスターサイズについては，次の形の大数の法則が成り立つ；

定理 1 ([1])

$\lambda > 1$ とする．任意の $\nu \in (\frac{1}{2}, 1)$ に対して，$\delta = \delta(\nu, \lambda)$ が存在して，次を満たす：

$$P\Big(\big| |\mathcal{C}_{\max}| - \zeta_\lambda n \big| \geqq n^\nu \Big) = O\big(n^{-\delta} \big), \qquad \text{as} \quad n \to \infty. \tag{1}$$

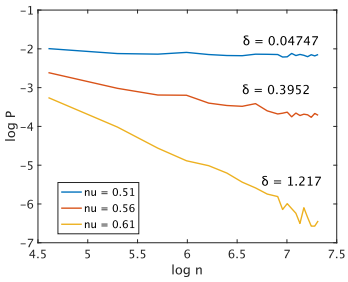

図 2.　(1) で ν を固定した時の P 対 n

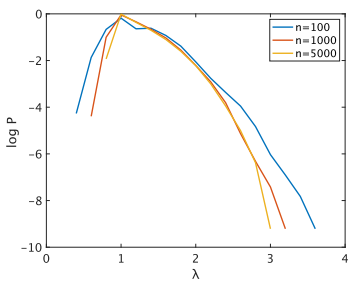

図 3　(1) で $\nu = 0.51$ を固定した時の $\log P$ 対 λ

n が大きくなると (1) 式左辺の確率 P が小さくなるのは当然だが，図 3 から分かるように，λ に関しても，P は $|\lambda - 1|$ の減少関数となっている事が見受けられる．この事を説明できるよう，P を λ の関数として評価する事を目指す．

さて，上記の定理 1 に対応する中心極限定理 (CLT) も，ある $\sigma^2 = \sigma_\lambda^2$ に対し，次のように知られている [1]：

$$\frac{|\mathcal{C}_{\max}| - \zeta n}{\sqrt{n}} \implies N\big(0, \sigma^2\big), \qquad \text{as} \quad n \to \infty. \tag{2}$$

つまり，個々のミクロ因子である各ノードのリンク数の分布も分散が有限であり，またマクロにも CLT が成り立つような分散が有限の系である．では NW の構造やサイズは，単純に独立確率変数列の和として捉えられるかというと一般にはそうではなく，各ノードからのリンクが現時点の最大クラスタに接続するか否かのランダム性がある．接続すれば，その新規ノードが $\mathcal{C}_{\max}$ へ追加され，$|\mathcal{C}_{\max}|$ が増大する．そのランダム性の結果，$|\mathcal{C}_{\max}|$ は優臨界では n の線形関数 ζn に漸近する．劣臨界では，$\dfrac{|\mathcal{C}_{\max}|}{\log n} \longrightarrow I_\lambda^{-1}$（確率収束）となる事が知られている．ただし $I_\lambda = \lambda - 1 - \log \lambda$ である．

$|\mathcal{C}_{\max}|$ 自体の分布は，必ずしも正規分布に従わない．幾つかの λ に対する $|\mathcal{C}_{\max}|$ のヒストグラムの変化を図 4 に示す．

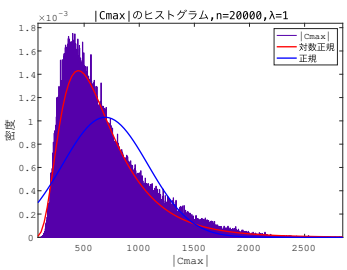

図 4(a)： $\lambda = 1$

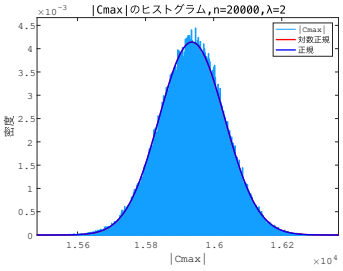

図 4(b)： $\lambda = 2$

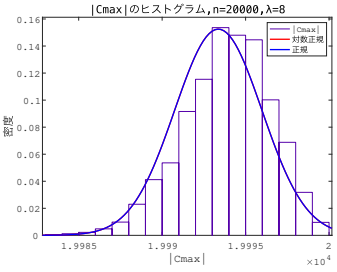

図 4(c)： $\lambda = 8$

図 4 ： 幾つかの λ に対する P のヒストグラム．λ が小さい方から大きい方へ向かって，右歪み → 左歪みと変化している．ビンの幅が異なるのは，横軸のレンジが変化している事による．

図 4 から，$|\mathcal{C}_{\max}|$ の分布は，λ が小さい方から右歪み → 左右対称 → 左歪みと変化している．左右対称のケースでは，正規分布が良くフィットしている．また，右歪みあるいは左歪みのうち，パラメータの値によっては対数正規分布がフィットする時もあれば，いずれもフィットしない時もある．これらの分布の変化について，劣臨界／臨界／優臨界を包摂した分布の確率的メカニズムや構造について調べることを目標としている．

参考文献

[1] R. van der Hofstad, *Random Graphs and Complex Networks*, Cambridge Univ. Press, 2017.

ADAPTIVE DEEP LEARNING FOR NONPARAMETRIC TIME SERIES REGRESSION

DAISUKE KURISU, RIKU FUKAMI, AND YUTA KOIKE

ABSTRACT. In this paper, we develop a general theory for adaptive nonparametric estimation of mean functions of nonstationary and nonlinear time series using deep neural networks (DNNs). We first consider two types of DNN estimators, non-penalized and sparse-penalized DNN estimators, and establish their generalization error bounds for general nonstationary time series. We then derive minimax lower bounds for estimating mean functions belonging to a wide class of nonlinear autoregressive (AR) models that include nonlinear generalized additive AR, single index, and threshold AR models. Building upon the results, we show that the sparse-penalized DNN estimator is adaptive and attains the minimax optimal rates up to a poly-logarithmic factor for many nonlinear AR models. Through numerical simulations, we demonstrate the usefulness of the DNN methods for estimating nonlinear AR models with intrinsic low-dimensional structures and discontinuous or rough mean functions, which is consistent with our theory.

1. INTRODUCTION

Let $(\Omega, \mathcal{G}, \{\mathcal{G}_t\}_{t\geq 0}, \mathrm{P})$ be a filtered probability space. Consider the following nonparametric time series regression model:

$$Y_t = m(X_t) + \eta(X_t)v_t,\ t = 1, \ldots, T, \tag{1.1}$$

where $T \geq 3$, $(Y_t, X_t) \in \mathbb{R} \times \mathbb{R}^d$, and $\{X_t, v_t\}_{t=1}^T$ is a sequence of random vectors adapted to the filtration $\{\mathcal{G}_t\}_{t=1}^T$. We assume $C_\eta := \sup_{x\in[0,1]^d} |\eta(x)| < \infty$. In this paper we investigate nonparametric estimation of the mean function m on the compact set $[0,1]^d$, that is, $f_0 := m\mathbf{1}_{[0,1]^d}$.

2. DEEP NEURAL NETWORKS

To estimate the mean function m of the model (1.1), we fit a deep neural network (DNN) with a nonlinear activation function $\sigma : \mathbb{R} \to \mathbb{R}$. The network architecture $(L, \mathbf{p})$ consists of a positive integer L called the *number of hidden layers* or *depth* and a *width vector* $\mathbf{p} = (p_0, \ldots, p_{L+1}) \in \mathbb{N}^{L+2}$. A DNN with network architecture $(L, \mathbf{p})$ is then any function of the form

$$f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}},\ x \mapsto f(x) = A_{L+1} \circ \sigma_L \circ A_L \circ \sigma_{L-1} \circ \cdots \circ \sigma_1 \circ A_1(x), \tag{2.1}$$

where $A_\ell : \mathbb{R}^{p_{\ell-1}} \to \mathbb{R}^{p_\ell}$ is an affine linear map defined by $A_\ell(x) := W_\ell x + \mathbf{b}_\ell$ for given $p_{\ell-1} \times p_\ell$ weight matrix W_ℓ and a shift vector $\mathbf{b}_\ell \in \mathbb{R}^{p_\ell}$, and $\sigma_\ell : \mathbb{R}^{p_\ell} \to \mathbb{R}^{p_\ell}$ is an element-wise nonlinear activation map defined as $\sigma_\ell(z) := (\sigma(z_1), \ldots, \sigma(z_{p_\ell}))'$. We assume that the activation function σ is C-Lipschitz for some $C > 0$, that is, there exists $C > 0$ such that $|\sigma(x_1) - \sigma(x_2)| \leq C|x_1 - x_2|$ for any $x_1, x_2 \in \mathbb{R}$. Examples of C-Lipschitz activation functions include the rectified linear unit

D. Kurisu is partially supported by JSPS KAKENHI Grant Number 20K13468. Y. Koike is partially supported by JST CREST Grant Number JPMJCR2115 and JSPS KAKENHI Grant Number 19K13668.

(ReLU) activation function $x \mapsto \max\{x, 0\}$ and the sigmoid activation function $x \mapsto 1/(1 + e^{-x})$. For a neural network of the form (2.1), we define

$$\theta(f) := (\mathrm{vec}(W_1)', \mathbf{b}_1', \ldots, \mathrm{vec}(W_{L+1})', \mathbf{b}_{L+1}')'$$

where $\mathrm{vec}(W)$ transforms the matrix W into the corresponding vector by concatenating the column vectors.

3. Main results

We establish

- generalization error bounds for non-penalized DNN and sparse-penalized DNN estimator,
- minimax lower bounds for estimating mean functions f_0 of a nonlinear AR model that belong to (i) composition structured functions and (ii) ℓ^0-bounded affine class,
- minimax optimality of sparse-penalized DNN estimator over those two classes of mean functions.

In addition to the theoretical results, we also conduct simulation studies to investigate the finite sample performance of the DNN estimators and compare their performance with other estimators (kernel ridge regression, k-nearest neighbors, and random forest). We find that the DNN methods work well for the models with (i) intrinsic low-dimensional structures and (ii) discontinuous or rough mean functions. These results are consistent with our main results.

References

(D. Kurisu) Graduate School of International Social Sciences, Yokohama National University, 79-4 Tokiwadai, Hodogaya-ku, Yokohama 240-8501, Japan.

E-mail address: kurisu-daisuke-jr@ynu.ac.jp

(R. Fukami) Graduate School of Mathematical Science, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8914, Japan.

E-mail address: rick.h.azuma@gmail.com

(Y. Koike) Graduate School of Mathematical Science, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8914, Japan.

E-mail address: kyuta@ms.u-tokyo.ac.jp

ベクトル量子化による大規模クラスタリングの近似法とその性質

寺田吉壱 [1,3], 山本 倫生 [2,3]
[1] 大阪大学大学院基礎工学研究科, [2] 大阪大学大学院人間科学研究科
[3] 理化学研究所革新知能統合研究センター

1. はじめに

近年，データの大規模化と複雑化が進み, データから仮説や有益な情報を獲得することが課題となっている. そして, 探索的なデータ解析の中でも教師なし学習の重要性が再認識されている．クラスタリング法は, データの背後のクラスタ構造を明らかにするための教師なし学習の方法であり, 様々な分野で広く応用されている. 最も代表的なクラスタリング法として, k-means 法が挙げられる. k-means 法はその簡便性と計算コストの低さから多用されるが, その単純さ故にデータの背後にある複雑なクラスタ構造を十分に捉えられない可能性がある. そのため, spectral clustering (von Luxburg, 2007) など, より柔軟にクラスタ構造を捉えられる方法の利用が望ましい. しかし, これらの方法は, 一般に計算コストが高く, 計算コストの削減が大きな課題となっている.

カーネル法におけるクラスタリング法に対しては, カーネル法に特化した近似法である Nyström 近似や Random Fourier Feature などの計算コスト緩和法が適用出来る. 一方で, クラスタリング法において, これらの方法による近似は安定的に機能しないという問題点がある. Yan et al. (2009) では, spectral clustering の k-means 法に基づく近似法 (KASP) が提案されている．KASP は, spectral clustering に限らず任意のクラスタリング法に対して適用することができる. この方法の大きな利点は, 安定性, 簡便性, 計算コストの低さ, 汎用性である.

本発表では, KASP の問題点を明らかにし, その問題点を緩和した一般のベクトル量子化 (vector quantization) を用いた近似方法を提案する.

2. KASP の問題点と提案手法

KASP では, クラスタ数を多く設定した k-means 法を大規模データに適用し, 得られたクラスタ中心をデータを代表する点とする. そして, 得られた代表点のみに spectral clustering など複雑なクラスタリング法を適用し, 代表点に対するラベルをその代表点に近いデータ点のラベルとする. しかし, k-means 法によって生成した代表点から構成される経験分布は, 母集団分布を代表する点とはならない. 実際に, 代表点の経験分布は, 母集団分布よりも裾が重い分布に収束することが示せる. そのため, KASP を用いて近似を行うと, グラスタリング結果にズレが生じてしまう.

この問題点に対する最もシンプルな解決法は, 各代表点に適切な重みを与えることである．K-means 法に対応するベクトル量子化は, 母集団分布との L_2-Wasserstein 距離を最小にするような $\#(\mathrm{supp}(Q)) \leq K$ を満たす離散測度を求める問題に対応している. このことに注目すると, 代表点に適切な重みを与えることで, KASP の問題点が解消されることがわかる. 一方で, 代表点の経験分布のズレから, 母集団分布において密度の低い点も代表点として生成される. それらの代表点には, 低い重みが割り振られるため, 効率が悪い.

この問題を解決するために, 発表者らのこれまでの研究で, 新しい代表点の生成方法である Density-Preserving Vector Quantization (DPVQ) を提案している． DPVQ が生成する代表点の経験分布は, 漸近的にデータの背後の分布へ収束することが示せる. そのため, DPVQ によって生成された代表点には平等な重みが割り振られるため, 効率的な近似が期待できる. 一方で, DPVQ は, 密度推定を必要とするため, 次元が高くなると不安定になるという問題がある.

Algorithm 1 $\mathrm{VQ}_n(\mu \,|\, r, K)$ の最適化アルゴリズム

1: $t \leftarrow 0$ とし, クラスタ中心 $\mu_1^{(0)}, \ldots, \mu_K^{(0)}$ を初期化する.

2: **for** $t = 0, \ldots, T$ **do**

3: 各 i $(i = 1, \ldots, n)$ に対して, $\left\|x_i - \mu_k^{(t)}\right\|$ を最小にするクラスタ k を割り当て, 帰属行列 $U^{(t)} = \left(u_{ij}^{(t)}\right)_{n \times K}$ を得る.

$$u_{ik}^{(t)} = \begin{cases} 1 & \text{if } \forall j;\ \left\|x_i - \mu_k^{(t)}\right\| \leq \left\|x_i - \mu_j^{(t)}\right\|, \\ 0 & \text{otherwise.} \end{cases}$$

4: クラスタ平均を以下で更新する.

$$\hat{\mu}_k^{(t+1)} = \frac{1}{\sum_{j=1}^n u_{jk}^{(t)} w_{jk}^{(t)}} \sum_{i=1}^n u_{ik}^{(t)} w_{ik}^{(t)} x_i, \quad w_{ik}^{(t)} = \begin{cases} \left\|x_i - \hat{\mu}_k^{(t)}\right\|^{r-2} & \text{if } \left\|x_i - \hat{\mu}_k^{(t)}\right\| > 0, \\ \delta & \text{if } \left\|x_i - \hat{\mu}_k^{(t)}\right\| = 0. \end{cases}$$

ここで, $\delta > 0$ は小さい正の定数である.

5: 収束判定条件を満たせば停止し, 満たさなければ $t \leftarrow t + 1$ とする.

6: **end for**

そこで, 本発表では, 以下で定義される order r のベクトル量子化器を用いた近似法を考える.

$$\mathrm{VQ}_n(\mu \,|\, r, K) := \frac{1}{n} \sum_{i=1}^n \min_{1 \leq k \leq K} \|x_i - \mu_k\|^r$$

ここで, $r \in (0, 2]$ は定数, $\|\cdot\|$ は $\mathbb{R}^d$ 上のノルム, $x_1, \ldots, x_n \in \mathbb{R}^d$ は各データ点, $\mu_k \in \mathbb{R}^d$ は k 番目のクラスタ中心, $\mu = (\mu_1, \ldots, \mu_K)$ である. 本発表では, 計算の簡便性のために, $\|\cdot\|$ は Euclid ノルムとする. $r = 2$ とすれば $\mathrm{VQ}_n(\mu \,|\, r, K)$ は k-means 法と一致するが, r を小さくとることで代表点の経験分布と母集団分布のズレを小さくすることができる. 本発表では, $\mathrm{VQ}_n(\mu \,|\, r, K)$ に対する最適化問題を高速に解くために, k-means like なアルゴリズムを提案する. Algorithm 1 において, $\mathrm{VQ}_n(\mu^{(t)} \,|\, r, K) > \mathrm{VQ}_n(\mu^{(t+1)} \,|\, r, K)$ という単調減少性が成り立つため, 停留点への収束性が保証できる.

詳細な理論的性質や $r < 2$ としたベクトル量子化器を用いた近似と k-means 法や DPVQ による近似の数値実験による比較は当日報告する. 特に, spectral clustering に対応する normalized cut (e.g., Terada and Yamamoto, 2019) に対して, 代表点の分布と母集団分布のズレの影響を明らかにする.

参考文献

Terada, Y. and Yamamoto, M. (2019). Kernel Normalized Cut: a Theoretical Revisit. *In Proceedings of the 36th International Conference on Machine Learning*, PMLR **97**, 6206–6214.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, **17**, 395–416.

Yan, D., Huang, L., and Jordan, M. I. (2009) Fast approximate spectral clustering. *In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 907–916.

傾向スコア解析のための情報量規準および差分の差法への展開

二宮 嘉行
統計数理研究所 数理・推論研究系

本講演では，基本的な因果推論モデル

$$y = \sum_{h=1}^{H} t^{(h)} y^{(h)}, \qquad y^{(h)} \sim f(\cdot \mid \boldsymbol{x}^{(h)}; \boldsymbol{\theta})$$

を扱う．ここで，$t^{(h)}$ $(\in \{0,1\})$ は h 番目の処置を割り当てられたときに 1 となる割り当て変数 $(\sum_{h=1}^{H} t^{(h)} = 1)$，$y^{(h)}$ $(\in \mathbb{R})$ は h 番目の処置を割り当てられたときの潜在結果変数，$\boldsymbol{x}^{(h)}$ $(\in \mathbb{R}^r)$ は $y^{(h)}$ のための説明変数，$f(y^{(h)} \mid \boldsymbol{x}^{(h)}; \boldsymbol{\theta})$ は $y^{(h)}$ に対する $\boldsymbol{x}^{(h)}$ の回帰モデル，$\boldsymbol{\theta}$ $(\in \mathbb{R}^p)$ はそこで用いられるパラメータである（$h \in \{1, 2, \ldots, H\}$）．$\boldsymbol{x}^{(h)}$ は交絡変数の一部を含んでもよい．このモデルでは，$t^{(h)} = 0$ なる $H-1$ 個の $y^{(h)}$ が欠測していると考えられる．そして，一般には $\mathrm{E}[y^{(h)}] \neq \mathrm{E}[y^{(h)} \mid t^{(h)} = 1]$ であるため，観測値のみから単純に $\boldsymbol{\theta}$ を推定するとバイアスが生じてしまう．そこで，$y^{(h)}$ と $t^{(h)}$ の交絡変数 $\boldsymbol{z}$ $(\in \mathbb{R}^s)$ が観測されていることを想定する．そして，このバイアスの除去を可能とするためのものである，無視できる割り当て条件

$$\{y^{(1)}, y^{(2)}, \ldots, y^{(H)}\} \perp\!\!\!\perp \{t^{(1)}, t^{(2)}, \ldots, t^{(H)}\} \mid \boldsymbol{z}$$

を仮定する．また，正値条件 $\mathrm{P}(t^{(h)} = 1) > 0$ を仮定する．このモデルにしたがう N 個のサンプルがあるとし，第 i サンプルの変数には添え字 i を付けることにする．$\boldsymbol{u}_i = (y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(H)}, t_i^{(1)}, t_i^{(2)}, \ldots, t_i^{(H)}, \boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}, \ldots, \boldsymbol{x}_i^{(H)}, \boldsymbol{z}_i)$ と書き，サンプルは独立，つまり

$$\boldsymbol{u}_i \perp\!\!\!\perp \boldsymbol{u}_j \qquad (i \neq j;\ i, j \in \{1, 2, \ldots, N\})$$

であることも仮定する．推定対象は，第 k 処置群が実際より $d^{(k)}$ 倍存在する母集団におけるパラメータ $\boldsymbol{\theta}$ とする．以降，パラメータの真値は $\boldsymbol{\theta}^*$ のように表すこととする．

$y^{(h)}$ と $\boldsymbol{z}$ の関係を正しくモデリングできれば，最尤法から因果効果を一致推定することができるが，一般にこのモデリングは難しい．そのため，このモデリングを必ずしも必要としない，傾向スコア $e^{(h)}(\boldsymbol{z}; \boldsymbol{\alpha}) \equiv \mathrm{P}(t^{(h)} = 1 \mid \boldsymbol{z}; \boldsymbol{\alpha})$ を用いたセミパラメトリックアプローチが用いられることも多い．ここで，$\boldsymbol{\alpha}$ $(\in \mathbb{R}^q)$ は傾向スコアの関数を特徴付けるパラメータである．そのアプローチの中での基本は，逆確率重み付け推定である．$w^{(h)}(\boldsymbol{z}; \boldsymbol{\alpha}) \equiv \sum_{k=1}^{H} d^{(k)} e^{(k)}(\boldsymbol{z}; \boldsymbol{\alpha}) / e^{(h)}(\boldsymbol{z}; \boldsymbol{\alpha})$ とし，

$$\sum_{i=1}^{N} \sum_{h=1}^{H} t_i^{(h)} w^{(h)}(\boldsymbol{z}_i; \boldsymbol{\alpha}) \frac{\partial}{\partial \boldsymbol{\theta}} \log f(y_i^{(h)} \mid \boldsymbol{x}_i^{(h)}; \boldsymbol{\theta}) = \mathbf{0}_p$$

を解くことで逆確率重み付け推定量 $\hat{\boldsymbol{\theta}}^{\mathrm{IPW}}$ は与えられる．無視できる割り当て条件のもと，逆確率重み付け推定量は一致性をもつ．

本設定において $y^{(h)}$ は $\boldsymbol{z}$ と相関があるが，逆確率重み付け推定はその期待値の推定に $\boldsymbol{z}$ の情報を直接は用いていない．二重頑健推定は，それを実行して逆確率重み付け推定を改良したものである．$\boldsymbol{z}$ を与えたもとでの $y^{(h)}$ の条件付き分布をさらなるパラメータ $\boldsymbol{\beta}$ $(\in \mathbb{R}^r)$ を用いて $p^{(h)}(y^{(h)} \mid \boldsymbol{z}; \boldsymbol{\beta})$ と表したとき，$\log f(y^{(h)} \mid \boldsymbol{x}^{(h)}; \boldsymbol{\theta})$ を $p^{(h)}(y^{(h)} \mid \boldsymbol{z}; \boldsymbol{\beta})$ で期待値をとり，その条件付き期待値を $g^{(h)}(\boldsymbol{x}^{(h)}, \boldsymbol{z}; \boldsymbol{\theta}, \boldsymbol{\beta})$ と書いて用いる．実際に $\boldsymbol{\beta}$ に対しては，なんらかの一致推定量 $\hat{\boldsymbol{\beta}}$ を代入する．具体的には，二重頑健推定量 $\hat{\boldsymbol{\theta}}^{\mathrm{DR}}$ は，上記推定方程式の左辺に

$$\sum_{i=1}^{N} \sum_{h=1}^{H} \left\{ \sum_{k=1}^{H} d^{(k)} t_i^{(k)} - t_i^{(h)} w^{(h)}(\boldsymbol{z}_i; \boldsymbol{\alpha}) \right\} \frac{\partial}{\partial \boldsymbol{\theta}} g^{(h)}(\boldsymbol{x}_i^{(h)}, \boldsymbol{z}_i; \boldsymbol{\theta}, \boldsymbol{\beta})$$

を加え，$\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ に推定量を代入し，それイコール $\mathbf{0}_p$ を $\boldsymbol{\theta}$ について解くことで与えられる．この推定方程式の i に依存する部分を $\boldsymbol{m}(\boldsymbol{u}_i;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta})$ と表しておく．傾向スコアか条件付き期待値のどちらかさえ正しくモデリングされていれば $\hat{\boldsymbol{\theta}}^{\mathrm{DR}}$ は一致性をもち，それ故に二重頑健と呼ばれる．

因果推論における情報量規準を導くためのリスクを定義する．通常の AIC タイプの情報量規準を考えるときと同様，$(y_i^{(h)\dagger}, t_i^{(h)\dagger}, \boldsymbol{x}_i^{(h)\dagger}, \boldsymbol{z}_i^{\dagger})$ を $(y_i^{(h)}, t_i^{(h)}, \boldsymbol{x}_i^{(h)}, \boldsymbol{z}_i)$ のコピーとし，逆確率重み付け推定で用いられる損失関数から自然に定義される

$$-2\sum_{i=1}^{N}\sum_{h=1}^{H}\mathrm{E}[t_i^{(h)\dagger}w^{(h)}(\boldsymbol{z}_i^{\dagger})\log f(y_i^{(h)\dagger}\mid \boldsymbol{x}_i^{(h)\dagger};\hat{\boldsymbol{\theta}}^{\mathrm{DR}})]$$

を考える．ここで，$w^{(h)}(\boldsymbol{z})$ は $w^{(h)}(\boldsymbol{z};\hat{\boldsymbol{\alpha}})$ の収束先とする．このリスクは，第 k 処置群が実際より $d^{(k)}$ 倍存在する母集団における，真の分布と推定された分布の Kullback-Leibler ダイバージェンスに基づくものとみなせる．このリスクの自然な推定量は $-2\sum_{i=1}^{N}\sum_{h=1}^{H}t_i^{(h)}w^{(h)}(\boldsymbol{z}_i)\log f(y_i^{(h)}\mid \boldsymbol{x}_i^{(h)};\hat{\boldsymbol{\theta}}^{\mathrm{DR}})$ であるが，それは言わずもがな過小評価する．そこで，バイアスを

$$-2\mathrm{E}\left[\sum_{i=1}^{N}\sum_{h=1}^{H}t_i^{(h)}w^{(h)}(\boldsymbol{z}_i)\log f(y_i^{(h)}\mid \boldsymbol{x}_i^{(h)};\hat{\boldsymbol{\theta}}^{\mathrm{DR}})-\sum_{i=1}^{N}\sum_{h=1}^{H}t_i^{(h)\dagger}w^{(h)}(\boldsymbol{z}_i^{\dagger})\log f(y_i^{(h)\dagger}\mid \boldsymbol{x}_i^{(h)\dagger};\hat{\boldsymbol{\theta}}^{\mathrm{DR}})\right]$$

と表現し，この期待値の中身の収束先を b^{limit} とし，$\mathrm{E}(b^{\mathrm{limit}})$ を漸近バイアスとして補正に用いる．漸近バイアスを導出する前に，準備として以下を与えておく．

補題.「結果変数に対する交絡変数のモデル」か「割り当て変数に対する交絡変数のモデル」のどちらか一方が正しければ，二重頑健推定量の誤差は

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{\mathrm{DR}}-\boldsymbol{\theta}^{*}=\frac{1}{N}\sum_{i=1}^{N}\Bigg\{&\boldsymbol{A}(\boldsymbol{\theta}^{*},\boldsymbol{\alpha}^{\dagger},\boldsymbol{\beta}^{\dagger})^{-1}\boldsymbol{m}(\boldsymbol{u}_i;\boldsymbol{\theta}^{*},\boldsymbol{\alpha}^{\dagger},\boldsymbol{\beta}^{\dagger})+\boldsymbol{C}_1(\boldsymbol{\theta}^{*},\boldsymbol{\alpha}^{\dagger},\boldsymbol{\beta}^{\dagger})\sum_{h=1}^{H}t_i^{(h)}\frac{\partial}{\partial\boldsymbol{\alpha}}\log e^{(h)}(\boldsymbol{z}_i;\boldsymbol{\alpha}^{\dagger})\\&+\boldsymbol{C}_2(\boldsymbol{\theta}^{*},\boldsymbol{\alpha}^{\dagger},\boldsymbol{\beta}^{\dagger})\sum_{h=1}^{H}\frac{\partial}{\partial\boldsymbol{\beta}}t_i^{(h)}\log p^{(h)}(y_i^{(h)}\mid \boldsymbol{z}_i;\boldsymbol{\beta}^{\dagger})\Bigg\}\{1+\mathrm{o_P}(1)\}\end{aligned}$$

の形で表される．$\boldsymbol{\alpha}^{\dagger}$，$\boldsymbol{\beta}^{\dagger}$ のどちらか一方は真値 $\boldsymbol{\alpha}^{*}$，$\boldsymbol{\beta}^{*}$ である．

$\boldsymbol{A}$，$\boldsymbol{C}_1$，$\boldsymbol{C}_2$ の定義はここでは省略するが，$\boldsymbol{C}_1(\boldsymbol{\theta}^{*},\boldsymbol{\alpha}^{\dagger},\boldsymbol{\beta}^{*})=\boldsymbol{C}_2(\boldsymbol{\theta}^{*},\boldsymbol{\alpha}^{*},\boldsymbol{\beta}^{\dagger})=\boldsymbol{O}$ が成立している．つまり，上記誤差には妥当でない可能性のあるスコア関数が含まれているが，妥当でないときは係数が $\boldsymbol{O}$ となっており，問題は存在しない．この補題より，以下の主結果が得られる．

定理.「結果変数に対する交絡変数のモデル」か「割り当て変数に対する交絡変数のモデル」のどちらか一方が正しければ，二重頑健推定のための情報量規準の漸近バイアスは

$$\mathrm{E}(b^{\mathrm{limit}})=\mathrm{tr}[\boldsymbol{A}(\boldsymbol{\theta}^{*},\boldsymbol{\alpha}^{*})^{-1}\{\boldsymbol{B}(\boldsymbol{\theta}^{*},\boldsymbol{\alpha}^{*})+\boldsymbol{D}_1(\boldsymbol{\theta}^{*},\boldsymbol{\alpha}^{*},\boldsymbol{\beta}^{*},\boldsymbol{\alpha}^{\dagger},\boldsymbol{\beta}^{\dagger})\}+\boldsymbol{D}_2(\boldsymbol{\theta}^{*},\boldsymbol{\alpha}^{*},\boldsymbol{\beta}^{*},\boldsymbol{\alpha}^{\dagger},\boldsymbol{\beta}^{\dagger})]$$

の形で与えられる．

$\boldsymbol{A}$，$\boldsymbol{B}$，$\boldsymbol{D}_1$，$\boldsymbol{D}_2$ の定義はここでは省略するが，既存の情報量規準から想像できるように，やはり漸近バイアスはパラメータの真値に依存していることがわかる．$\boldsymbol{\alpha}^{*}$ か $\boldsymbol{\beta}^{*}$ のどちらかを推定できない本設定では，漸近バイアスの評価はできないと思われるかもしれないが，真値を推定せずに経験推定をうまく用いることで，$\boldsymbol{D}_1$ と $\boldsymbol{D}_2$ の一致推定量 $\hat{\boldsymbol{D}}_1$ と $\hat{\boldsymbol{D}}_2$ を与えることができる．結果，

$$\begin{aligned}\mathrm{DRIC}\equiv&-2\sum_{i=1}^{N}\sum_{h=1}^{H}\frac{t_i^{(h)}}{e^{(h)}(\boldsymbol{z}_i;\hat{\boldsymbol{\alpha}})}\log f(y_i^{(h)}\mid \boldsymbol{x}_i^{(h)};\hat{\boldsymbol{\theta}}^{\mathrm{DR}})\\&+2\mathrm{tr}[\boldsymbol{A}(\hat{\boldsymbol{\theta}}^{\mathrm{DR}},\hat{\boldsymbol{\alpha}})^{-1}\{\boldsymbol{B}(\hat{\boldsymbol{\theta}}^{\mathrm{DR}},\hat{\boldsymbol{\alpha}})+\hat{\boldsymbol{D}}_1(\hat{\boldsymbol{\theta}}^{\mathrm{DR}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})\}+\hat{\boldsymbol{D}}_2(\hat{\boldsymbol{\theta}}^{\mathrm{DR}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})]\end{aligned}$$

という形の情報量規準を与えることができる．情報量規準自体が二重頑健性をもっているため，二重頑健基準と呼ぶことにする．

外れ値に頑健な因果効果の推定量

原田 和治[a], 藤澤 洋徳[b]
a) 東京医科大学 医療データサイエンス分野
b) 統計数理研究所 数理・推論研究系

1 はじめに

統計学的推定において,外れ値はバイアスの主な原因のひとつである．Canavire-Bacarrez らは平均因果効果の様々な推定量について，外れ値の影響を実験的に評価した [1]．特に，vertical outlier と呼ばれる，結果変数に入るタイプの外れ値が，平均因果効果の推定に深刻なバイアスをもたらすことが指摘されている．

ロバスト統計は，外れ値に頑健な統計学的推論の方法を扱う分野である．平均因果効果に対するロバストな推定量としては，例えば，中央値因果効果に対する推定量が挙げられる．具体的には，Zhang らの IPW および二重頑健（Doubly Robust; DR）中央値などがある [2]．平均因果効果とは推定対象が異なるものの，後述する潜在アウトカムが対称分布に従うならば，中央値因果効果に対する推定量を平均因果効果に対するロバスト推定量として用いることができる．しかし，一般に中央値は平均に比べて外れ値に頑健ではあるものの，影響を取り除けるわけではない．特に，外れ値が分布の中心から見て片方に偏っている場合に，外れ値の割合が高まるほど，その影響は顕著となる．

本発表では，DR 推定量を拡張し，平均因果効果に対する中央値よりも外れ値に頑健な推定量を提案した．提案手法の基本的なアイデアは，中央値よりも頑健な密度べき重みづけ推定方程式に，二重頑健な M 推定量（DR M 推定量）の枠組みを適用するものである．しかし，単に DR M 推定量を当てはめるだけでは，モデル選択に関する二重頑健性を失うことを示す．さらに，外れ値の割合による補正により，二重頑健性を保ちながら，外れ値にも頑健な推定量を構成できることを示した．なお，本発表は査読付き学術誌に掲載された論文 [3] をもとにしている．

2 提案手法

提案した 3 手法（DP-IPW, DP-DR, εDP-DR）のうち，二重頑健性を保ちながら，外れ値に対しても頑健な ε DP-DR 法の概略を示す．

DP-DR 法は，DR 推定量を一般の M 推定に拡張した DR M 推定量において，推定関数に密度べき型推定関数を選択したものであった．しかし，DP-DR 法は外れ値存在下で二重頑健性を保てないことを報告した．DP-DR 法の外れ値存在下でのバイアスを補正し，外れ値存在下でも二重頑健性を満たすように構成したものが，εDP-DR 推定量である．次の推定

方程式の解を εDP-DR 推定量と定義する．

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\pi(X_i;\hat{\alpha})}h(Y_i;\mu)^{\gamma}(Y_i-\mu)-(1-\hat{\varepsilon})\frac{T_i-\pi(X_i;\hat{\alpha})}{\pi(X_i;\hat{\alpha})}\mathbb{E}_{\hat{q}}[h(Y_i;\mu)^{\gamma}(Y_i-\mu)|T_i=1,X_i]\right\}=0 \tag{1}$$

外れ値の割合 ε は，例えば先述の [4] によって，条件付分布 q と同時に一致推定できる．εDP-DR 推定量は，外れ値がない場合には DP-DR 推定量と同様に二重頑健性を持つ．さらに，外れ値がある場合でも，推定方程式の不偏性の意味で二重頑健性を保持している．傾向スコアが正しい場合の証明は DP-DR 推定量と同様であり，条件付分布モデルが正しい場合にも，補正により，推定方程式の不偏性を証明できる．$\hat{\varepsilon}$ による補正は傾向スコアが正しい場合は必要なく，条件付分布モデルだけが正しい場合には，$\hat{\varepsilon}$ も同時に正しく推定されるため，補正のために，新たなモデルを特定する必要はない．

3 数値実験

提案手法の外れ値耐性や推定効率について，実験的に検討した．性能評価のため，外れ値耐性なしの IPW/DR 推定量のほか，各種提案されている IPW/DR 中央値と比較した．本発表の設定に沿ったシミュレーション用データセットを，サンプルサイズ $n=100$ で 10,000 セット生成した．真の $\mu^{(1)}$ は 3 とした．

全般的な傾向として，提案手法は外れ値耐性・推定効率の両面で中央値よりも優れる傾向にあった．また，γ の値を大きくすると外れ値耐性は増すものの，推定効率は低下する傾向にあった．この傾向は他の密度べき重みを用いたロバスト推定量と同様である．

また，本稿が想定する混合分布型の外れ値とは別に，誤差項を Cauthy 分布とすることで，対称に外れ値が発生する場合についても検討した．外れ値が対称の場合，中央値も一致性を失わないため，上述の結果に比べると中央値に有利な設定である．それにも関わらず，提案手法は中央値に基づく各手法よりも RMSE が小さかった．

加えて，DP-DR 法と εDP-DR 法を比較し，ε による補正の効果を検証した結果についても報告した．条件付分布モデルだけを正しく特定し，各手法を適用した結果を表 3 に示す．DP-DR 推定量は，γ を十分大きくしても，外れ値の増加に伴いバイアスが拡大している．一方，εDP-DR 法はつねに真値に近い値を推定できている．この結果からも，$\hat{\varepsilon}$ による補正がバイアスの軽減に機能していると考えられる．

参考文献

[1] Gustavo Canavire-Bacarreza, Luis Castro Peñarrieta, and Darwin Ugarte Ontiveros. Outliers in Semi-Parametric estimation of treatment effects. *Econometrics*, 9(2):19, April 2021.

[2] Zhiwei Zhang, Zhen Chen, James F Troendle, and Jun Zhang. Causal inference on quantiles with an obstetric application. *Biometrics*, 68(3):697–706, September 2012.

[3] Kazuharu Harada and Hironori Fujisawa. Outlier-Resistant estimators for average treatment effect in causal inference. *Statistica Sinica*, 34(2), April 2024. (to appear).

[4] Takafumi Kanamori and Hironori Fujisawa. Robust estimation under heavy contamination using unnormalized models. *Biometrika*, 102(3):559–572, May 2015.

Statistical Inference for Local Granger Causality

Yan Liu, Masanobu Taniguchi, Hernando Ombao [1]

Abstract

Granger causality has been employed to investigate causality relations between components of stationary multiple time series. We generalize this concept by developing statistical inference for local Granger causality for multivariate locally stationary processes. Our proposed local Granger causality approach captures time-evolving causality relationships in nonstationary processes. The proposed local Granger causality is well represented in the frequency domain and estimated based on the parametric time-varying spectral density matrix using the local Whittle likelihood. Under regularity conditions, we demonstrate that the estimators converge to multivariate normal in distribution. Additionally, the test statistic for the local Granger causality is shown to be asymptotically distributed as a quadratic form of a multivariate normal distribution. For practical demonstration, the proposed local Granger causality method uncovered new functional connectivity relationships between channels in brain signals. Moreover, the method was able to identify structural changes in financial data.

keywords: Brain signals, Local Granger causality, Local Whittle likelihood, Multivariate locally stationary processes, Time-varying spectral density matrix, Topological data analysis

1. Local Granger Causality

Let $\boldsymbol{X}_{t,T} = (X_{t,T}^{[1]}, \ldots, X_{t,T}^{[p]})^\top$ be a sequence of p-dimensional multivariate stochastic processes

$$\boldsymbol{X}_{t,T} = \sum_{j=-\infty}^{\infty} A_{t,T}(j)\boldsymbol{\epsilon}_{t-j}, \tag{1.1}$$

where the sequences $\{A_{t,T}(j)\}_{j\in\mathbb{Z}}$ satisfy the regularity conditions. The process (1.1) is usually referred to as the *multivariate locally stationary process*.

Let m and M be two positive integers such that $p = m+M$. Suppose $\boldsymbol{X}_{t,T} = \left(\boldsymbol{X}_{t,T}^{(1)^\top}, \boldsymbol{X}_{t,T}^{(2)^\top}\right)^\top$, $\boldsymbol{X}_{t,T}^{(1)} \in \mathbb{R}^m$, $\boldsymbol{X}_{t,T}^{(2)} \in \mathbb{R}^M$, has the time-varying spectral density matrix $\boldsymbol{f}(u,\lambda)$ with the partition

$$\boldsymbol{f}(u,\lambda) = \begin{pmatrix} \boldsymbol{f}(u,\lambda)_{11} & \boldsymbol{f}(u,\lambda)_{12} \\ \boldsymbol{f}(u,\lambda)_{21} & \boldsymbol{f}(u,\lambda)_{22} \end{pmatrix} := \frac{1}{2\pi} A(u,\lambda) K A(u,-\lambda)^\top, \qquad u \in [0,1],$$

where $A(u,\lambda) := \sum_{j=-\infty}^{\infty} A(u,j)\exp(\mathrm{i}j\lambda)$. Let $\Sigma(u)$ be the one-step-ahead prediction error covariance matrix based on the time-varying spectral density matrix $\boldsymbol{f}(u,\lambda)$ with the same partition.

[1] this work was supported by JSPS Grant-in-Aid for Scientific Research (C) 20K11719 (Liu, Y.) and JSPS Grant-in-Aid for Scientific Research (S) 18H05290 (Taniguchi, M.)

Let $H(\tau_1,\tau_2)=\overline{\mathrm{sp}}(\boldsymbol{X}_{t,T}^{(1)},1\le t\le\tau_1;\boldsymbol{X}_{t,T}^{(2)},1\le t\le\tau_2)$ be the closed linear subspace generated by $\{\boldsymbol{X}_{t,T}^{(1)},1\le t\le\tau_1;\boldsymbol{X}_{t,T}^{(2)},1\le t\le\tau_2\}$. Especially, we use $H(\tau_1,0)$ and $H(0,\tau_2)$ to express the closed linear subspace generated by $\{\boldsymbol{X}_{t,T}^{(1)},t\le\tau_1\}$, $\{\boldsymbol{X}_{t,T}^{(2)},t\le\tau_2\}$, respectively. Introducing the companion process

$$\boldsymbol{Y}_{t,T}^{(2)}=\boldsymbol{X}_{t,T}^{(2)}-E\big(\boldsymbol{X}_{t,T}^{(2)}\mid H(t,t-1)\big),$$

we propose the local Granger causality measure from $\{\boldsymbol{X}_{t,T}^{(2)}\}$ to $\{\boldsymbol{X}_{t,T}^{(1)}\}$ as

$$\mathrm{GC}^{(2\to1)}(u)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\mathrm{FGC}(u,\lambda)d\lambda,$$

where

$$\mathrm{FGC}(u,\lambda)=\ln\frac{|\boldsymbol{f}(u,\lambda)_{11}|}{\Big|\boldsymbol{f}(u,\lambda)_{11}-2\pi\boldsymbol{g}(u,\lambda)_{12}\tilde{\Sigma}(u)_{22}^{-1}\boldsymbol{g}(u,\lambda)_{21}\Big|}.$$

Here, $\boldsymbol{g}(u,\lambda)$ is the time-varying spectral density matrix of the process $\{\big(\boldsymbol{X}_{t,T}^{(1)\top},\boldsymbol{Y}_{t,T}^{(2)\top}\big)^{\top}\}$, and $\tilde{\Sigma}(u)$ is an $(M\times M)$-matrix

$$\tilde{\Sigma}(u)_{22}=\Sigma(u)_{22}-\Sigma(u)_{21}\Sigma(u)_{11}^{-1}\Sigma(u)_{12}.$$

Regarding the local hypothesis $H_0^{(2\to1)}:\mathrm{GC}^{(2\to1)}(u)=0$, we use the following test statistic:

$$S^{\dagger}(u):=2T\,b_T\,\mathrm{GC}^{(2\to1)}(u;\hat{\boldsymbol{\theta}}_T),$$

where $\hat{\boldsymbol{\theta}}_T$ is the Whittle estimate. Then we have the following result.

Theorem 1.1. *Under the null hypothesis $H_0^{(2\to1)}:\mathrm{GC}^{(2\to1)}(u)=0$, if assumptions in Liu et. al. (2021) hold with $b_T^{-1}=o\big(T(\ln T)^{-6}\big)$ and $b_T=o(T^{-1/5})$, we have*

$$S^{\dagger}(u)\xrightarrow{\mathcal{L}}\mathcal{N}\big(0,\mathbb{V}(u)\big)^{\top}\mathcal{H}(u)\mathcal{N}\big(0,\mathbb{V}(u)\big),$$

where $\mathcal{N}\big(0,\mathbb{V}(u)\big)$ and $\mathcal{H}(u)$ are defined in Liu et. al. (2021).

References

Dahlhaus, R. (2000). A likelihood approximation for locally stationary processes. *The Annals of Statistics* **28** 1762–1794.

Dahlhaus, R. (2009). Local inference for locally stationary time series based on the empirical spectral measure. *Journal of Econometrics* **151** 101–112.

Granger, C. W. J. (1963). Economic processes involving feedback. *Information and Control* **6** 28–48.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* **37** 424–438.

Hosoya, Y. (1991). The decomposition and measurement of the interdependency between second-order stationary processes. *Probability Theory and Related Fields* **88** 429–444.

Liu, Y., Taniguchi, M. and Ombao, H. (2021). Statistical inference for local Granger causality. *arXiv: https://arxiv.org/abs/2103.00209*.

Ombao, H., von Sachs, R. and Guo, W. (2005). SLEX analysis of multivariate nonstationary time series. *Journal of the American Statistical Association* **100** 519–531.

Taniguchi, M., Puri, M. L. and Kondo, M. (1996). Nonparametric approach for non-Gaussian vector stationary processes. *Journal of Multivariate Analysis* **56** 259–283.

弱い共通サポート条件下での確率的介入に基づく因果効果の推定

山本 倫生 [1,2]
[1] 大阪大学大学院人間科学研究科, [2] 理化学研究所革新知能統合研究センター

1. はじめに　社会政策や個人の行動に関する介入など，決定論的な処置の割付が現実的でない場合に対応するための方法として確率的介入（stochastic interventions）があり，近年多くの研究が行われている．確率的介入では，仮想的な処置の割付は共変量に依存して確率的に決定されると仮定する．例えば，Díaz and van der Laan (2012) では，処置の割付メカニズムとして介入分布を想定し，処置の水準をシフトさせて確率的に処置を割り付ける方法を提案しており，これまでにいくつかの拡張がなされている．一般的な因果推論手法と同様に，Díaz and van der Laan (2012) の方法でも効果の識別や推定のためには正値性の仮定が重要となる．しかし，確率的介入で扱う多値や連続的な水準を持つ処置ではその成立を仮定するのは非現実的である．そこで本研究では Sufficient Dimension Reduction (SDR）に基づいて正値性に対する条件を緩和した推定量を提案する．

2. 提案方法　$\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^p$ を共変量ベクトル，$T \in \mathcal{T}$ を処置変数，$Y \in \mathbb{R}$ をアウトカムとし，$(\boldsymbol{X}, T, Y)$ は分布 P_0 に従うとする．また，独立同分布な観測データ $(\boldsymbol{X}_i, T_i, Y_i) \sim P_0$ $(i = 1, \ldots, n)$ が得られているとする．ノンパラメトリック構造方程式モデル $\boldsymbol{X} = f_X(E_{\boldsymbol{X}})$, $T = f_T(\boldsymbol{X}, E_T)$, $Y = f_Y(T, \boldsymbol{X}, E_Y)$ において，処置変数 T は介入分布 $P_\delta(g_0)(T = t|\boldsymbol{X}) = g_0(t - \delta(\boldsymbol{X})|\boldsymbol{X})$ に従って値が決定されると仮定する．ここで，$\delta(\boldsymbol{X})$ は処置の水準をシフトさせる量を表し，g_0 は真の（処置への）曝露メカニズム $g_0(t|\boldsymbol{X}) = P_0(T = t|\boldsymbol{X})$ を表す．介入分布を導入したもとでの潜在アウトカム $Y(P_\delta)$ の期待値 $\mathbb{E}[Y(P_\delta)]$ が興味のあるパラメータである．なお，介入により確率 1 で $T = t$ に固定された場合の潜在アウトカムが，一般的な潜在アウトカム $Y(t)$ に対応している．条件付き交換可能性 $Y(t) \perp\!\!\!\perp T \mid \boldsymbol{X}$, $\forall t \in \mathcal{T}$ および一致性の仮定 $T = t \Rightarrow Y = Y(t)$ のもとで，例えば $\boldsymbol{X}$ が離散の場合，興味のあるパラメータは確率分布 P の関数 $\Psi(P)$ により

$$\mathbb{E}[Y(P_\delta)] = \Psi(P) = \sum_{T \in \mathcal{T}} \sum_{\boldsymbol{X} \in \mathcal{X}} \mathbb{E}(Y|T, \boldsymbol{X}) P_\delta(g)(T|\boldsymbol{X}) P(\boldsymbol{X}) \tag{1}$$

のように表される．なお，パラメータの真値は真の分布 P_0 を用いて $\psi_0 = \Psi(P_0)$ と表される．

パラメータ (1) の推定において，Díaz and van der Laan (2012) では逆確率重み付け (IPTW) 推定量，拡張 IPTW (AIPTW) 推定量，および targeted maximum likelihood estimator (TMLE) が提案されている．それらの推定量の妥当性を保証するためには正値性条件が必要であり，そのためには強い共通サポート条件（Luo et al., 2017）

$$\Omega(\boldsymbol{X}) = \Omega(\boldsymbol{X} \mid T = t),\ t \in \mathcal{T}$$

が必要となる．ただし，$\Omega(\boldsymbol{X})$ および $\Omega(\boldsymbol{X} \mid T = t)$ はそれぞれ母集団全体での $\boldsymbol{X}$ の取りうる値，$T = t$ の集団での $\boldsymbol{X}$ の取りうる値を表す．この条件は各処置群での共変量 $\boldsymbol{X}$ の分布が完全に重なっていることを要求しているが，$\boldsymbol{X}$ が高次元の場合や経時測定データで時点ごとにこの条件が必要とされる場合など，その成立を仮定するのが難しい場合が多い．そこで，提案手法では SDR を用いてより弱い条件のもとでパラメータ (1) の推定を行う．

提案手法では，まず各 $t \in \mathcal{T}$ での潜在結果変数に対する Central Mean Subspace (CMS; Cook and Li, 2002) $\mathcal{S}_{\mathbb{E}(Y(t)|\boldsymbol{X})}$ を観測データから推定する必要がある．一般には，観測データから推定

可能な観測アウトカム Y における処置の各水準 t の部分集団での CMS $\mathcal{S}^{D}_{\mathbb{E}(Y|\boldsymbol{X},T=t)}$ と上記の CMS $\mathcal{S}_{\mathbb{E}(Y(t)|\boldsymbol{X})}$ が同じである保証はない．この問題に対して，提案手法の設定では，Luo et al. (2017) の Theorem 1 と同様に，条件付き交換可能性と一般の $\mathcal{T}$ に対する弱い共通サポート条件

$$\Omega(\mathbf{B}_t^\top \boldsymbol{X}) = \Omega(\mathbf{B}_t^\top \boldsymbol{X} \mid T = t),\ t \in \mathcal{T}$$

のもとで，二つの CMS が一致することが示される．よって，観測データから推定可能な CMS を，潜在結果変数に対する CMS と同一視して利用することができる．

提案手法では，まず各 $t \in \mathcal{T}$ に対する CMS $\mathcal{S}_{\mathbb{E}(Y(t)|\boldsymbol{X})}$ を minimum average variance estimator (Xia et al., 2002) を用いて推定し，その基底ベクトルからなる係数行列 $\mathbf{B}_t$ を推定する．具体的には，Luo et al. (2017) と同様に，各 $t \in \mathcal{T}$ に対して，

$$\sum_{i=1}^{n}\sum_{j=1}^{n} I(T_i = t)\{Y_i - a_t(\boldsymbol{X}_j, \mathbf{B}_t) - (\boldsymbol{X}_i - \boldsymbol{X}_j)^\top \mathbf{B}_t \boldsymbol{b}_t(\boldsymbol{X}_j)\}^2 K_{h_1, r(t)}\left\{\mathbf{B}_t^\top(\boldsymbol{X}_i - \boldsymbol{X}_j)\right\} \tag{2}$$

を最小にする $\mathbf{B}_t \in \mathbb{R}^{p\times r(t)}$ を求める．ここで，$K_{h,r}(\cdot) = h^r K(\cdot/h)$ はバンド幅 h を持つ $\mathbb{R}^r$ 上のカーネル関数である．(2) を最小にする $a_t(\boldsymbol{x}, \mathbf{B}_t)$，$\boldsymbol{b}_t(\boldsymbol{x})$，$\mathbf{B}_t$ は交互最小 2 乗法によって求めることができる．推定された係数行列を $\widehat{\mathbf{B}}_t$ とする．

次に，各 $t \in \mathcal{T}$ に対して，アウトカムの条件付き期待値 $\mathbb{E}(Y|T, \widehat{\mathbf{B}}_T^\top \boldsymbol{X})$ を

$$W_h(\widehat{\mathbf{B}}_t)\sum_{i=1}^{n} I(T_i = t)\left\{Y_i - a'_t(\boldsymbol{x}, \widehat{\mathbf{B}}_t) - (\boldsymbol{X}_i - \boldsymbol{x})^\top \widehat{\mathbf{B}}_t \boldsymbol{b}'_t(\boldsymbol{x})\right\}^2 K_{h_2, r(t)}\left\{\hat{\mathbf{B}}_t^\top(\boldsymbol{X}_i - \boldsymbol{x})\right\} \tag{3}$$

を最小にする $a'_t(\boldsymbol{x}, \hat{\mathbf{B}}_t)$ によって推定する．ただし，CMS の係数行列 $\mathbf{B}_t$ に対して $W_h(\mathbf{B}_t) = [\sum_{i=1}^{n} K_{h,r(t)}\{\mathbf{B}_t^\top(\boldsymbol{X}_i - \boldsymbol{x})\}]^{-1}$ である．なお，(3) を最小にする $a'_t(\boldsymbol{x}, \hat{\mathbf{B}})$ および $\boldsymbol{b}'_t(\boldsymbol{x}) \in \mathbb{R}^{r(t)}$ は，通常の局所線形回帰と同様に重み付き最小 2 乗推定量として求めることができる．$a'_t(\boldsymbol{x}, \widehat{\mathbf{B}}_t)$ の推定量を $\hat{a}'_t(\boldsymbol{x}, \widehat{\mathbf{B}}_t)$ とする．

最後に，興味のあるパラメータ (1) を

$$\frac{1}{n}\sum_{i=1}^{n} I(T_i + \delta(\boldsymbol{X}_i) \in \mathcal{T})\hat{a}_{T_i+\delta(\boldsymbol{X}_i)}(\boldsymbol{X}_i, \widehat{\mathbf{B}}_{T_i+\delta(\boldsymbol{X}_i)}) \tag{4}$$

によって推定する．

CMS の推定およびアウトカムの条件付き期待値の推定に関するいくつかの仮定のもとで推定量 (4) は一致性および漸近正規性を持つことが示される．シミュレーションにより，特に強い共通サポート条件が成立せず，アウトカムと共変量，および，処置と共変量の関係が非線形である場合に，既存の手法よりも提案手法が良い性質を持つことが確認された．

参考文献

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Annals of Statistics*, 30, 455–474.

Díaz, I., and van der Laan, M. J. (2012). Population intervention causal effects based on stochastic interventions. *Biometrics*, 68, 541—549.

Luo, W., Zhu, Y., and Ghosh, D. (2017). On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika*, 104, 51–65.

Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space (with discussion). *Journal of the Royal Statistical Society, Series B*, 64, 364–410.