科研費シンポジウム

「統計学と機械学習の数理と展開」

科学研究費・基盤研究（A）「大規模複雑データの理論と方法論の総合的研究」

研究代表者：青嶋 誠 (筑波大学). 課題番号：**15H01678**

日時： 2019 年 9 月 18 日 (水)〜19 日 (木)
場所： 東京工業大学 西 8 号館 E 棟 1 0 階 大会議室
開催責任者： 金森 敬文 (東京工業大学)

プログラム：

9/18 (水)

9:55 - 10:00 オープニング　青嶋 誠 (筑波大)

表現学習 (座長：川島)

10:00 ∼ 10:35 和田 裕一郎 (名大)
Robust Label Prediction via Label Propagation and Geodesic k-Nearest Neighbor in Online Semi-Supervised Learning

10:35 ∼ 11:10 高橋 昂 (東工大)
相関のある特徴量に対する半解析的 Stability Selection 法

特別講演 1 (座長：川島)

11:20 ∼ 12:05 竹内 一郎 (名工大/理研 AIP)
Selective Inference による教師なし学習結果の統計的信頼性評価

12:05〜13:30 ランチ・ブレイク

統計的学習 (座長：熊谷)

13:30 ∼ 14:05 稲津 佑 (理研 AIP)
入力がコストに応じたランダム性を持つ場合のレベルセット推定のための能動学習

14:05 ∼ 14:40 高梨 耕作 (理研 AIP)
Predictive properties of forecast combination, ensemble methods, and Bayesian synthesis

カーネル法 (座長：熊谷)

14:50 ∼ 15:25 牧草 夏実 (千葉大)
再生核ヒルベルト空間における射影のモーメントによる二標本検定

15:25 ∼ 16:00 松井 孝太 (理研 AIP)
Power Series Kernels に基づくノンパラメトリック学習のための変数選択法

高次元統計 **I** (座長：金森)

16:15 ∼ 16:50 Atina Husnaqilati (東北大)
Component Retention in PCA Applied to Microarray Datasets

16:50 ∼ 17:25 下野 寿之 (Digital Garage, Inc.)
超立方体に内接する超楕円球による線形回帰の解釈

特別講演 2 (座長：金森)

17:35 ∼ 18:20 後藤 振一郎 (統数研)
マスター方程式から厳密に得られる期待値変数の情報幾何学，接触幾何学およびその周辺

9/19 (木)

データ解析 (座長：松井)

10:00 ～ 10:35　金森 敬文 (東工大/理研 AIP)
Similarity Measures and Statistical Models in Recommendation Problems

10:35 ～ 11:10　新村 秀一 (成蹊大)
高次元遺伝子解析の呪いからの解放 3 -機械学習などの工学研究の問題点-

特別講演３ (座長：松井)

11:20 ～ 12:05　佐々木 博昭 (はこだて未来大)
確率密度関数のモード探索とその応用

12:05～13:30 ランチ・ブレイク

高次元統計 II (座長：高梨)

13:30 ～ 14:05　中山 優吾 (筑波大)
Asymptotic properties of kernel PCA with Gaussian kernel for high-dimensional data

14:05 ～ 14:40　梅津 佑太 (名工大)
超高次元スパース加法モデルにおける変数選択

14:40 ～ 15:15　川島 孝行 (東工大)
ガンマ・ダイバージェンス最小化に基づくロバストかつスパースな回帰

15:15 ～ 15:20　クロージング

Robust Label Prediction via Label Propagation and Geodesic k-Nearest Neighbor in Online Semi-Supervised Learning

Yuichiro WADA[1], Siqiang SU[2], Wataru KUMAGAI[3], and Takafumi KANAMORI[3,4]

[1]Nagoya University, Furocho, Chikusaku, Nagoya 464-8601 Japan
[2]The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[3]RIKEN AIP, Nihonbashi 1-chome Mitsui Building, 15th floor,1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027 Japan
[4]Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552 Japan

1 Introduction

We consider the case in which a few manually labeled data points are provided before the arrival of continuously streamed unlabeled data points. The goal is to predict the label of the newly arrived data point correctly and quickly under severe memory constraints. To handle above problems, online Semi-Supervised Learning (SSL) methods have been proposed [1, 2, 3]. These methods adopt a similar strategy: At each time, firstly the data adjacency graph is recompressed after incorporating a newly arrived data point, and secondly the label of new data point is predicted by using the graph and an offline semi-supervised algorithm. On one hand, the several graph compressing algorithms [1, 2, 3] are employed in the first step. On the other hand, the method employed in the second step is often fixed: Label Propagation (LP) [4]. However, LP is not robust against outliers but also computationally not efficient. The computational compleixity of LP is known as $O(n^3)$ where n is the number of data points. These drawbacks make the online methods underperform.

In this paper, we propose an offline SSL algorithm named *Robust Label Prediction* (RLP). RLP is not only more robust against outliers but also more computationally efficient than LP. This proposed method is intended to assist online graph-based SSL algorithms. The efficacy of RLP in creating new online algorithms is demonstrated in numerical experiments.

In the following, the labeled dataset and the unlabeled dataset are denoted by $L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ and $U = \{\mathbf{x}_i\}_{i=l+1}^{l+u}$ respectively, where $\mathbf{x}_i \in \mathbb{R}^p$ is the feature vector, and $y_i \in \{1, ..., S\}$ is its label. The total sample size $l + u$ is denoted by n as well. The set $L_{\mathbf{x}}$ means the set of feature vectors in L, i.e., $L_{\mathbf{x}} = \{\mathbf{x}_i\}_{i=1}^{l}$, and $\widehat{U}$ means the predicted U, i.e., $\widehat{U} = \{(\mathbf{x}_i, \hat{y}_i)\}_{i=l+1}^{l+u}$, where $\hat{y}_i$ is the predicted label of $\mathbf{x}_i$.

2 Proposed Method

We propose an offline SSL algorithm named RLP. This algorithm consists of three steps. On the basis of the neighbor graph, RLP first selects some unlabeled samples that represent the global structure of the data manifolds. The selected points are called *hub* points. The second step assigns labels to hub points by using a special type of LP. The third step predicts the labels on the remaining unlabeled samples by using Geodesic k-Nearest Neighbor (GkNN) [5]. The detailed procedure of RLP is as follows.

1. Construct the graph $G = (L_{\mathbf{x}} \cup U, E)$, where E is defined by k_1NN manner with the Euclidean metric.
2. Build the hub dataset H such that $|H| = h$, where the elements of H are top h highest degree nodes on G. Thereafter, define non-hub dataset $\bar{H}$ by $U \setminus H$.
3. Define the geodesic metric d_G by computing the graph shortest path distance on graph G.
4. For predicting labels of H, conduct a special type of LP on $L_{\mathbf{x}} \cup H$. In this LP, a different number of neighbors k_2 and the predefined geodesic metric d_G are employed. The predicted set is denoted by $\widehat{H}$.

Table 1: The averaged prediction accuracy with the standard deviation on unlabeled data in eight real-world data streams (Yale to USPS). In each dataset, l denotes the number of labeled data obtained before the arrival of each stream. The size of stream is denoted by T.

	(l, T)	Online QRLP	Online QLP
Yale	(75,85)	**0.528**(**0.046**)	0.471(0.048)
ORL	(80,220)	**0.722**(**0.029**)	0.656(0.058)
UMNIST	(60,240)	**0.637**(**0.028**)	0.544(0.059)
COIL	(80,120)	**0.660**(**0.034**)	0.592(0.062)
Vowel	(100,1300)	**0.969**(**0.004**)	0.966(0.002)
MNIST	(100,1900)	**0.679**(**0.021**)	0.663(0.019)
optdigits	(10,4990)	**0.971**(**0.002**)	0.961(0.031)
USPS	(100,1900)	**0.700**(**0.020**)	0.612(0.050)

5. Update the labeled dataset by $L \leftarrow L \cup \widehat{H}$. Then, by using L, predict the labels of $\bar{H}$ by conducting GkNN on G with $k = k_v$. Finally, output $\widehat{U} = \widehat{H} \cup \widehat{\bar{H}}$.

For the theoretical computational complexity of RLP, by summing up each step of above procedure, we can get $O\left(h^3 + pn^2 + \kappa(k_1 + \log n)n\right)$ where $\kappa = \max\{k_2, k_v\}$: see [5]. Since the number of hub data points h is the hyperparameter, by introducing the upper bound to h, we can make RLP faster than LP, i.e., the complexity of RLP is $O(n^2)$.

3 Numerical Experiment

In this experiment, we evaluate how helpful employing RLP is in online algorithms. The evaluation is based on the replacement of LP in the previous online SSL algorithm named online Qunatized LP (QLP) [1]. The replaced method is named online Quantized RLP (QRLP). In both online QLP and online QRLP, Doubling Algorithm [6] is employed for the graph compressing algorithm. For this evaluation, we applied both methods on eight real-world datasets. The result is shown in Table 1. Note that the hyperparameters in both methods were already tuned appropriately before handling the datasets. As we can see in this table, online QRLP outperforms online QLP in all datasets. Since, except label prediction part, both online methods are the same, we can conclude RLP tend to be more helpful in creating an online algorithm.

References

[1] M. Valko, B. Kveton, L. Huang, and D. Ting, "Online semi-supervised learning on quantized graphs," in *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, (Arlington, Virginia, United States), pp. 606–614, AUAI Press, 2010.

[2] Y. Tao, R. Triebel, and D. Cremers, "Semi-supervised online learning for efficient classification of objects in 3d data streams," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pp. 2904–2910, IEEE, 2015.

[3] T. Wagner, S. Guha, S. P. Kasiviswanathan, and N. Mishra, "Semi-supervised learning on data streams via temporal label propagation," in *International Conference on Machine Learning*, pp. 5082–5091, 2018.

[4] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," 2002.

[5] A. Moscovich, A. Jaffe, and B. Nadler, "Minimax-optimal semi-supervised regression on unknown manifolds," *arXiv preprint arXiv:1611.02221*, 2016.

[6] M. Charikar, C. Chekuri, T. Feder, and R. Motwani, "Incremental clustering and dynamic information retrieval," *SIAM Journal on Computing*, vol. 33, no. 6, pp. 1417–1440, 2004.

相関のある特徴量に対する 半解析的 Stability Selection 法

高橋 昂　樺島 祥介
東京工業大学　情報理工学院　数理・計算科学系
{takahashi, kaba}@sp.dis.titech.ac.jp

1　スパース線形回帰に対する stability selection（SS）法

線形回帰における変数選択の問題を考える。M 個の独立なデータ点 $(\boldsymbol{a}_\mu, y_\mu), \mu = 1, 2, ..., M$ からなるデータセット D があるとする。ただし、$\boldsymbol{a}_\mu \in \mathbb{R}^N$ と $y_\mu \in \mathbb{R}$ はそれぞれ特徴量と応答変数で、$y_\mu = \boldsymbol{a}_\mu^\top \boldsymbol{x}_0 + w_\mu$, $w_\mu \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$ の関係にある。$S(x) = \text{supp}(\boldsymbol{x}_0)$ とすると、変数選択の目的は D から $S(\boldsymbol{x}_0)$ を決定することである。このような問題は、高次元統計学などで頻出する。しかし、素朴に Lasso で $\boldsymbol{x}_0$ の推定値 $\hat{\boldsymbol{x}}$ を求め、$S(\hat{\boldsymbol{x}})$ を $S(\boldsymbol{x}_0)$ の推定値としても、特徴量に相関がある場合には交差検証によって選ばれた最適な正則化パラメータを用いても $S(\hat{\boldsymbol{x}}) \neq S(\boldsymbol{x}_0)$ である。

SS 法 [MB10] は、Lasso の変数選択性能を向上させるために提案された方法である。SS 法の基本的なアイディアは、正則化パラメータ $\boldsymbol{\lambda}$ について適当に設計した分布 $p(\boldsymbol{\lambda}) = \prod(\lambda_i)$ を考え、$D, \boldsymbol{\lambda}$ の両方が確率変数であるとして、Lasso 推定量 $\hat{\boldsymbol{x}}$ について $\Pi_i \equiv \text{Prob}[\hat{x}_i(D, \boldsymbol{\lambda}) \neq 0], i = 1, 2, ..., N$ を計算し、Π_i が大きな変数を選ぶことにある。ただし、真の D の分布は不明であるので、ブートストラップ法で近似計算する:

$$\Pi_i = \frac{1}{B} \# \left\{ b \mid \hat{x}_i(D, \boldsymbol{c}^{(b)}, \boldsymbol{\lambda}^{(b)}), b = 1, 2, ..., B \right\}, i = 1, 2, .., N, \tag{1}$$

$$\hat{\boldsymbol{x}}(D, \boldsymbol{c}^{(b)}, \boldsymbol{\lambda}^{(b)}) = \arg\min_{\boldsymbol{x} \in \mathbb{R}^N} \left[\frac{1}{2} \sum_{\mu=1}^{M} c_\mu^{(b)} (y_\mu - \boldsymbol{a}_\mu^\top \boldsymbol{x})^2 + \sum_{i=1}^{N} \lambda_i^{(b)} |x_i| \right], \tag{2}$$

$$c_\mu^{(b)} \sim_{\text{i.i.d.}} \text{Poisson}(c_\mu^{(b)}; 1/2), \quad \lambda_i^{(b)} \sim_{\text{i.i.d.}} (\delta(\lambda - 2\lambda_0) + \delta(\lambda - \lambda_0))/2. \tag{3}$$

ここで、B はブートストラップデータセットの数である。[MB10] らは、この $\boldsymbol{\Pi}$ を用いた変数選択によって、Lasso での変数選択よりも偽陽性確率を低減できることを示した。しかし、Π_i の計算には (2) 式の最適化問題を $B \gg 1$ 回繰り返し解く必要がある。Π_i を正確に評価するためには B を大きくとる必要があり、SS 法の実行には全体として大きな計算時間がかかることが問題である。

2　半解析的 stability selection 法

便利のため、SS 法の統計力学的な定式化を整理する。Boltzmann 分布 $p^{(\beta)}(\boldsymbol{x}|D, \boldsymbol{c}, \boldsymbol{\lambda})$ を以下のように定義する:

$$p^{(\beta)}(\boldsymbol{x} \mid D, \boldsymbol{c}, \boldsymbol{\lambda}) = \frac{1}{Z^{(\beta)}(D, \boldsymbol{c}, \boldsymbol{\lambda})} \prod_{\mu=1}^{M} e^{-\frac{\beta}{2} c_\mu (y_\mu - \boldsymbol{a}_\mu^\top \boldsymbol{x})^2} \prod_{i=1}^{N} e^{-\beta \lambda_i |x_i|}, \tag{4}$$

$\beta > 0$ は逆温度と呼ばれるパラメータ、Z は分配関数と呼ばれる規格化定数である。$\beta \to \infty$ で Boltzmann 分布は $\min_x [\sum_\mu c_\mu (y_\mu - \boldsymbol{a}_\mu^\top \boldsymbol{x})^2/2 + \sum_i \lambda_i |x_i|]$ を満たす $\boldsymbol{x}$ 上に集中するので、特定の $\boldsymbol{c}, \boldsymbol{\lambda}$ に対する Lasso 推定量は

$$\hat{x}_i(D, \boldsymbol{c}, \boldsymbol{\lambda}) = \lim_{\beta \to \infty} \mathbb{E} \left[x_i; p^{(\beta)}(\boldsymbol{x} \mid D, \boldsymbol{c}, \boldsymbol{\lambda}) \right], \tag{5}$$

と書ける。$\mathbb{E}_x[...; p(x)]$ は p に従う確率変数 x についての期待値である。最終目標は、指示関数 $\mathbb{1}(...)$ に対し、

$$\Pi_i = \mathbb{E}_{\boldsymbol{c}, \boldsymbol{\lambda}} \left[\mathbb{1}(\hat{x}_i(\boldsymbol{c}, \boldsymbol{\lambda}) \neq 0) \right] = \lim_{\beta \to \infty} \mathbb{E}_{\boldsymbol{c}, \boldsymbol{\lambda}} \left[\mathbb{1} \left(\mathbb{E} \left[x_i; p^{(\beta)}(\boldsymbol{x} \mid D, \boldsymbol{c}, \boldsymbol{\lambda}) \right] \neq 0 \right) \right], \tag{6}$$

を計算することである。確率モデルの期待値の極限として表すことで、確率的推論の近似技法を用い易くなる。

半解析的リサンプリング法 [MO03] はレプリカ法 [MPV87] と近似推論法を組み合わせることで、SS 法のようなリサンプリング法における、推定の繰り返しを行うことに伴う計算量を削減する方法である。(6) は Boltzmann

分布のマージナル分布のモーメントが $\boldsymbol{c}, \boldsymbol{\lambda}$ にどう依存するかを調べればよいが、一般に $r \in \mathbb{N}$ 次モーメントの $\boldsymbol{c}, \boldsymbol{\lambda}$ に対する期待値は、レプリカ法 [MPV87] を用いると

$$\mathbb{E}_{\boldsymbol{c},\boldsymbol{\lambda}}\left[\left(\mathbb{E}\left[x_i; p^{(\beta)}(\boldsymbol{x} \mid D, \boldsymbol{c}, \boldsymbol{\lambda})\right]\right)^r\right] = \lim_{n\to 0} \mathbb{E}_{x_i^{(1)}}\left[x_i^{(1)}; \tilde{p}^{(\beta)}(\{\boldsymbol{x}^{(\gamma)}\}_{\gamma=1}^n)\right], \tag{7}$$

$$\tilde{p}^{(\beta)}(\{\boldsymbol{x}^{(\gamma)}\}_{\gamma=1}^n) \doteq \mathbb{E}_{\boldsymbol{c},\boldsymbol{\lambda}}\left[\prod_{\gamma=1}^n \left\{\prod_{\mu=1}^M e^{-\frac{\beta}{2}c_\mu(y_\mu - \boldsymbol{a}_\mu^\top \boldsymbol{x}^{(\gamma)})^2} \prod_{i=1}^N e^{-\beta\lambda_i |x_i^{(\gamma)}|}\right\}\right], \tag{8}$$

と書ける。ただし、$\doteq$ は「規格化定数まで含めて一致」を表す記号である。また、(7) の $n \to 0$ の極限は、整数の n に対して期待値を評価し、$n \in \mathbb{R}$ に結果を外挿して極限をとることを表す。ここで、(8) は少なくとも $n = 1, 2, ...$ に対しては、$(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)})$ の置換について不変である**レプリカ対称性**と呼ばれる性質を持っている。このとき、de Finetti の表現定理 [HS55] から (8) は

$$\tilde{p}^{(\beta)}(\{\boldsymbol{x}^{(\gamma)}\}_{\gamma=1}^n) = \int \prod_{\gamma=1}^n p^{(\beta)}(\boldsymbol{x}^{(\gamma)} \mid \xi) p^{(\beta)}(\xi) d\xi = \mathbb{E}_\xi\left[\prod_{\gamma=1}^n p^{(\beta)}(\boldsymbol{x}^{(\gamma)} \mid \xi)\right], \tag{9}$$

と書き表せるはずである。ここで、ξ は $\boldsymbol{c}, \boldsymbol{\lambda}$ の効果を反映した確率変数である。この表現を用いて得た表式について $n \to 0$ の極限をとると、r 次モーメント (7) は非常に単純に表されることがわかる:

$$\lim_{\beta\to\infty} \int \left\{\mathbb{E}_{\boldsymbol{x}}\left[x_i; p^{(\beta)}(\boldsymbol{x} \mid \xi)\right]\right\}^r p^{(\beta)}(\xi) d\xi. \tag{10}$$

したがって、$p^{(\beta)}(\boldsymbol{x} \mid \xi), p^{(\beta)}(\xi)$ の $\beta \to \infty$ での近似分布を求めることができれば、繰り返しの推定をすることなく、任意の r 次モーメントを計算できる。それにとどまらず、$p^{(\beta)}(\boldsymbol{x} \mid \xi), p^{(\beta)}(\xi)$ の関数形から、$\hat{\boldsymbol{x}}$ が $\boldsymbol{c}, \boldsymbol{\lambda}$ にどのように依存しているのかを伺うこともできる。レプリカ法を用いて問題を書き直し、先にリサンプリングに関する平均をとってから近似推論を行うのが「半解析的リサンプリング法」である。

上述のレプリカ法を用いた後、近似推論を行う上では、

- $\boldsymbol{c}, \boldsymbol{\lambda}$ で平均をとった分布の近似分布を効率的に求められるか
- 変数選択の問題を考えているので、なるべく期待値一致近似法などの高精度な近似法を用いたいが、その近似分布を効率的に求められるのか

が問題になる。前者は、結局近似操作中に $\boldsymbol{c}, \boldsymbol{\lambda}$ の平均操作が大きな負荷を生み出さないかという点についての懸念、後者は近似推論技法自体の問題である。実際、[MO03] は高精度な近似分布を得るための反復アルゴリズムを提案し、ひとたび固定点が得られれば高精度な結果を出すことを示したが、その反復ダイナミクスは極めて不安定でしばしば発散するものであった。本研究では、情報理論のベクトル近似メッセージパッシング法 [RSF17] を用いて、前述の懸念両方を解決する近似推論法を提案した。講演では、研究のモチベーション、関連研究、近似推論法の枠組み、および疑似／実データを用いた実験による提案手法の性能評価結果を紹介する。

より詳細な説明についての文献としては、本講演の予稿、および [TK19] を参照されたい。

参考文献

[HS55] Edwin Hewitt and Leonard J Savage, *Symmetric measures on cartesian products*, Transactions of the American Mathematical Society **80** (1955), no. 2, 470–501.

[MB10] Nicolai Meinshausen and Peter Bühlmann, *Stability selection*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72** (2010), no. 4, 417–473.

[MO03] Dörthe Malzahn and Manfred Opper, *A statistical mechanics approach to approximate analytical bootstrap averages*, Advances in Neural Information Processing Systems 15 (S. Becker, S. Thrun, and K. Obermayer, eds.), MIT Press, 2003, pp. 343–350.

[MPV87] Marc Mézard, Giorgio Parisi, and Miguel Virasoro, *Spin glass theory and beyond: An introduction to the replica method and its applications*, vol. 9, World Scientific Publishing Company, 1987.

[RSF17] Sundeep Rangan, Philip Schniter, and Alyson K Fletcher, *Vector approximate message passing*, 2017 IEEE International Symposium on Information Theory (ISIT), IEEE, 2017, pp. 1588–1592.

[TK19] Takashi Takahashi and Yoshiyuki Kabashima, *Replicated vector approximate message passing for resampling problem*, arXiv preprint arXiv:1905.09545 (2019).

Selective Inferenceによる教師なし学習結果の統計的信頼性評価

井上 茂乗 [1]　谷崎 光佑 [1]　梅津 佑太 [1]　稲津 佑 [2]　坪田 庄真 [3]　橋本 典明 [3]　本谷 秀堅 [1]　○竹内 一郎 [1,2]
[1] 名古屋工業大学　[2] 理化学研究所　[3] 名古屋大学

1　はじめに

データ駆動型科学ではデータに基づいて仮説の選択が行われるため，仮説選択バイアスの補正が必要である．仮説選択バイアスの補正を行う方法として，selective inference と呼ばれるアプローチが注目を集めている [1, 2, 3]. Selective inference は主に特徴選択を行った後の線形モデルの統計的推測に用いられてきたが，本研究では selective inference を教師なし学習の結果の統計的推測に利用する．本講演では，K 平均法に基づくクラスタリングとグラフカットに基づく画像セグメンテーションの結果を selective inference の枠組で評価する方法を紹介する [4, 5].

2　K 平均法によるクラスタリングの信頼性評価

本節では，K 平均法によってクラスタを同定した後に，クラスタ間で各特徴が異なるかどうかを統計的仮説検定の枠組で判定する方法を紹介する．観測データとして，d 次元の特徴を持つ n 個の事例が行列 $X \in \mathbb{R}^{n\times d}$ として与えられているとする．K 平均法は各クラスタのクラスタ中心とクラスタ構成要素の更新を繰り返すアルゴリズムであり，n 個の事例が K 個のクラスタに分割される．クラスタリングのステップを $t = 1, \ldots, T$ と表し（T は K 平均法が収束するまでにかかったステップ数），各ステップ $t \in [T]$ におけるクラスタ $k \in [K]$ のクラスタ中心とクラスタ構成要素をそれぞれ，$\boldsymbol{m}_k^t \in \mathbb{R}^d$，$C_k^t \subseteq [n]$ と表す．2 つの異なるクラスタ a と b（$1 \le a \le b \le K$）において，各特徴がクラスタ a と b で異なるかどうかを考える．観測データ X の各要素を並べた $n \times d$ 次元のベクトルが $n \times d$ 次元正規分布に従っているものとし，その平均ベクトルは未知，分散共分散行列は既知とする．クラスタ a と b に属する事例の特徴 j の真の平均をそれぞれ，$\mu_{a,j}$，$\mu_{b,j}$ とし，統計的仮説検定

$$\mathrm{H}_0 : \mu_{a,j} = \mu_{b,j} \text{ v.s. } \mathrm{H}_1 : \mu_{a,j} \neq \mu_{b,j} \tag{1}$$

を考える．検定統計量として $\tau = m_{a,j}^T - m_{b,j}^T$ を用い，仮説がデータに基づいて選択されたことを踏まえても $|\tau|$ が十分に大きいかどうかを判定する．

式 (1) で与えられる仮説検定問題は，クラスタリングアルゴリズムによって選択されたものであるため，仮説選択バイアスが生じてしまう．仮説選択バイアスを補正するため，selective inference の枠組では，クラスタの選択に関するイベントに条件付けられた条件付き推論を行う．クラスタリング結果を得るためのイベントとして，各ステップのクラスタ構成要素 $\mathcal{E} := \{(C_a^t, C_b^t)\}_{t\in[T]}$ を考える．データ X に T ステップの K 平均法を適用してこれらのクラスタリングのイベントが得られる事象を $\mathcal{E} \leftarrow \mathcal{C}(X)$ と表記する．ここで，$\mathbb{R}^{n\times d}$ の部分領域として，$\mathcal{X} := \{X' \in \mathbb{R}^{n\times d} \mid \mathcal{E} \leftarrow \mathcal{C}(X')\}$ を考える．すなわち，この部分領域における任意の観測データ $X' \in \mathcal{X}$ に対して T ステップの K 平均法を適用すると，そのイベントが観測データの場合と同じく $\mathcal{E}$ となることを意味している．また，両側検定に関する計算の都合上，τ の符号に関するイベントも導入し，クラスタリングイベントと τ の符号が観測データにおけるものと同一になるようなデータ空間の部分領域を $\mathcal{X}^+ \subseteq \mathbb{R}^{n\times d}$ とする．詳細は割愛するが，selective inference の枠組を利用することで，式 (1) で与えられる仮説検定問題に対する selective p-value $p_j^{(a,b)}$ を計算することができる．Selective p-value $p_j^{(a,b)}$ は，$X \in \mathcal{X}^+$ の条件のもと $\mathbb{P}_{\mathrm{H}_0}(p_j^{(a,b)} \le \alpha \mid X \in \mathcal{X}^+) = \alpha \ \forall \alpha \in (0, 1)$. を満たすため，仮説選択の影響を排除したうえで，通常の p 値と同様に統計的信頼性の指標として用いることができる．なお，詳細は割愛するが上記のクラスタリングイベント $X \in \mathcal{X}^+$ が X に関する二次不等式として定式化できることを利用して，selective p-value の計算が行われている．

3　グラフカットによる画像セグメンテーション結果の信頼性評価

本説では，グラフカットによって画像をオブジェクト領域とバックグラウンド領域に分割するセグメンテーションを行った際に，2 つの領域の差が統計的に有意であるかを検証することによってセグメンテーション結果の信頼性評価を行う方法を紹介する．観測画像として，n 個のピクセルから成る画像が n 次元ベクトル $\boldsymbol{x} \in \mathbb{R}^n$ として与

えられているとする．各ピクセルをノードとし，隣接するピクセル間にエッジを持つグラフを考え，各エッジにピクセル値が似ているほど高い値を持つような重みを割り当てる．グラフカットによるセグメンテーションでは，重み付きグラフを分割することで画像をオブジェクト領域とバックグラウンド領域へ分割する詳細は割愛するが，グラフカットによるセグメンテーションは重み付き有向グラフの最大フロー問題として定式化され，効率的に解くことができる．グラフカットによって分割されたオブジェクト領域とバックグラウンド領域のピクセル ID の集合をそれぞれ $\mathcal{O} \in [n]$，$\mathcal{B} \in [n]$ とする．各画素を並べた n 次元ベクトル $\boldsymbol{x}$ が n 次元正規分布に従っているものとし，その平均ベクトルは未知，分散共分散行列は既知とする．

オブジェクト領域とバックグラウンド領域の差が十分に大きいかどうかを統計的に定量化するため，両者の画素値の真の平均をそれぞれ，$\mu_{\mathcal{O}}$，$\mu_{\mathcal{B}}$ とし，仮説検定問題

$$\mathrm{H}_0 : \mu_{\mathcal{O}} = \mu_{\mathcal{B}} \text{ v.s. } \mathrm{H}_1 : \mu_{\mathcal{O}} \neq \mu_{\mathcal{B}} \tag{2}$$

を考える．検定統計量として $\tau = m_{\mathcal{O}} - m_{\mathcal{B}}$ を用い，仮説がデータに基づいて選択されたことを踏まえても $|\tau|$ が十分に大きいかどうかを判定する．ただし，$m_{\mathcal{O}} = |\mathcal{O}|^{-1} \sum_{i \in \mathcal{O}} x_i$，$m_{\mathcal{B}} = |\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} x_i$ である．

式 (2) で与えられる仮説検定問題は，グラフカットセグメンテーションアルゴリズムによって選択されたものであるため，仮説選択バイアスが生じてしまう．仮説選択バイアスを補正するため，selective inference の枠組では，セグメンテーションに関するイベントに条件付けられた条件付き推論を行う．グラフカットアルゴリズムによってセグメンテーション結果を得るためのイベントを $\mathcal{E}$ とし，画像データ $\boldsymbol{x}$ にグラフカットを適用してイベント $\mathcal{E}$ が得られた事象を $\mathcal{E} \leftarrow \mathcal{S}(\boldsymbol{x})$ と表記する．ここで，$\mathbb{R}^n$ の部分領域として，$\mathcal{X} := \{\boldsymbol{x}' \in \mathbb{R}^n \mid \mathcal{E} \leftarrow \mathcal{S}(\boldsymbol{x}')\}$ を考える．すなわち，この部分領域における任意の観測データ $\boldsymbol{x}' \in \mathcal{X}$ に対してグラフカットを適用すると，そのイベントが観測データの場合と同じく $\mathcal{E}$ となることを意味している．詳細は割愛するが，selective inference の枠組を利用することで，式 (2) で与えられる仮説検定問題に対する selective p-value $p^{(\mathcal{O},\mathcal{B})}$ を計算することができる．Selective p-value $p^{(\mathcal{O},\mathcal{B})}$ は，$\boldsymbol{x} \in \mathcal{X}$ の条件のもと $\mathbb{P}_{\mathrm{H}_0}(p^{(\mathcal{O},\mathcal{B})} \leq \alpha \mid \boldsymbol{x} \in \mathcal{X}) = \alpha \ \forall \alpha \in (0,1)$ を満たすため，仮説選択の影響を排除したうえで，通常の p 値と同様に統計的信頼性の指標として用いることができる．なお，詳細は割愛するが上記のグラフカットのイベント $\boldsymbol{x} \in \mathcal{X}$ が，一部の非線形関数の二次スプライン近似を導入すれば，$\boldsymbol{x}$ に関する二次不等式として定式化できることを利用して，selective p-value の計算が行われている．

4　おわりに

本講演では，selective inference の基本的な考え方と，それぞれの問題における selective p-value の具体的計算法を説明する．クラスタリング結果の推論の問題ではシングルセル解析の分析に適用した例を，セグメンテーション結果の推論の問題では病理画像の分析に適用した結果を紹介する．

参考文献

[1] Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[2] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

[3] Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

[4] Inoue S., Umezu Y., Tsubota S., Suzuki K., duVerle DA., Kadomatsu K., Tsuda K., and Takeuchi I. Valid statistical method for cluster-specific differentially-expressed genes in single-cell rna-sequencing by post-clustering inference. *in preparation*, 2019.

[5] Tanizaki K., Inatsu Y., Hashimoto N., Hontani H., and Takeuchi I. Computing valid p-values for image segmentation by selective inference. *arXiv:1906.00629*, 2019.

入力がコストに応じたランダム性を持つ場合の レベルセット推定のための能動学習

稲津 佑 [1]　　竹内 一郎 [1,2,3]

[1] 理化学研究所革新知能統合研究センター　[2] 名古屋工業大学 情報工学専攻・情報科学フロンティア研究院

[3] 物質・材料研究機構 情報統合型物質・材料研究拠点

1. 概要

評価コストが高い black-box 関数が, ある閾値を上回る領域と下回る領域を同定する問題をレベルセット推定 (LSE) 問題という. LSE は環境モニタリングや許容可能なパラメータ空間のサブセットの同定等のタスクに用いられており ([1]), 実応用において重要なタスクとなっている. また, 実応用においては, 入力がコストに応じたランダム性を持つと想定できる (e.g., タンパク質のアミノ酸配列の改変実験) 場合がある. 本研究では, ガウス過程 (GP) とそれに基づく信用区間を用い, 更に, [2] によって提案された期待分類増加量に基づいた獲得関数を拡張することで, 入力がコストに応じたランダム性を持つ場合の LSE のための能動学習法を提案する.

2. 設定

関数 $f : D \to \mathbb{R}$ を, $D \subset \mathbb{R}^d$ で定義された評価コストが高い black-box 関数とする. 各入力 $\boldsymbol{x} \in D$ に対し, 関数 $f(\boldsymbol{x})$ の値は $f(\boldsymbol{x}) + \varepsilon$ として観測されるとする. ただし, ε は正規分布 $\mathcal{N}(0, \sigma^2)$ に従う, 独立なノイズである. このとき, D の有限部分集合 Ω 上での, f に対する Level set estimation (LSE) を考える:

Definition 2.1. 閾値を h とする. このとき, f に対する上位集合 H および下位集合 L を

$$H = \{\boldsymbol{x} \in \Omega \mid f(\boldsymbol{x}) > h\},\ L = \{\boldsymbol{x} \in \Omega \mid f(\boldsymbol{x}) \leq h\} \tag{2.1}$$

で定める.

更に, 本稿では, 入力に対しコストに依存した不確実性が伴う状況を考える. コスト $c_1, \ldots, c_k$ は, $0 < c_1 < c_2 < \cdots < c_k$ を満たすとする. 各コスト c_i, $i \in \{1, \ldots, k\} \equiv [k]$ と入力 $\boldsymbol{x} \in \Omega$ に対し, $\boldsymbol{x}$ を入力した際に実際に入力される値 $\boldsymbol{s}(\boldsymbol{x}, c_i)$ は確率変数 $\boldsymbol{S}(\boldsymbol{x}, c_i)$ からのランダム標本とする. ただし, $\boldsymbol{s}(\boldsymbol{x}, c_i) \in D$ かつ $\boldsymbol{S}(\boldsymbol{x}, c_i)$ は既知の密度関数 $g(\boldsymbol{s}|\boldsymbol{\theta}_{\boldsymbol{x}}^{(c_i)})$ を持つとする. この設定の下, 総コストをできるだけ小さくしつつ上位集合 H および下位集合 L を同定することを目標とする.

3. レベルセット推定

まず, 未知関数 f に対するモデリングとして, ガウス過程 (GP) を用いる. 関数 f に対する事前分布に GP, $\mathcal{GP}(0, k(\boldsymbol{s}, \boldsymbol{s}'))$ を仮定する. ここで, $k(\boldsymbol{s}, \boldsymbol{s}') : D \times D \to \mathbb{R}$ は正定値カーネルである. すなわち, 入力と観測の組 $\{(\boldsymbol{s}_j(\boldsymbol{x}_j, c_{i_j}), y_j\}_{j=1}^t$ が与えられたとき, f の事後分布は再び GP となり, $f(\boldsymbol{x})$ の事後平均 $\mu_t(\boldsymbol{x})$, 事後分散 $\sigma_t^2(\boldsymbol{x})$ および共分散 $k_t(\boldsymbol{x}, \boldsymbol{x}')$ はそれぞれ以下で与えられる:

$$\mu_t(\boldsymbol{x}) = \boldsymbol{k}_t(\boldsymbol{x})^\top \boldsymbol{C}_t^{-1} \boldsymbol{y}_t, \sigma_t^2(\boldsymbol{x}) = k_t(\boldsymbol{x}, \boldsymbol{x}), k_t(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{k}_t(\boldsymbol{x})^\top \boldsymbol{C}_t^{-1} \boldsymbol{k}_t(\boldsymbol{x}').$$

ここで, $\boldsymbol{k}_t(\boldsymbol{x}) = (k(\boldsymbol{s}_1(\boldsymbol{x}_1, c_{i_1}), \boldsymbol{x}), \ldots, k(\boldsymbol{s}_t(\boldsymbol{x}_t, c_{i_t}), \boldsymbol{x}))^\top$, $\boldsymbol{C}_t = (\boldsymbol{K}_t + \sigma^2 \boldsymbol{I}_t)$, $\boldsymbol{y}_t = (y_1, \ldots, y_t)^\top$ および $\boldsymbol{I}_t$ は t 次単位行列である.

次に, 各点 $\boldsymbol{x} \in \Omega$ に対し, 第 t 試行時における $f(\boldsymbol{x})$ に対する信用区間を $Q_t(\boldsymbol{x}) = [l_t(\boldsymbol{x}), u_t(\boldsymbol{x})]$ で定める. ただし, $l_t(\boldsymbol{x}) = \mu_t(\boldsymbol{x}) - \beta^{1/2}\sigma_t(\boldsymbol{x})$, $u_t(\boldsymbol{x}) = \mu_t(\boldsymbol{x}) + \beta^{1/2}\sigma_t(\boldsymbol{x})$ であり, $\beta^{1/2} \geq 0$ である. このとき, 精度パラメータ $\epsilon > 0$ を用いて, H および L を以下の H_t および L_t を用いて推定する:

$$H_t = \{\boldsymbol{x} \in \Omega \mid l_t(\boldsymbol{x}) > h - \epsilon\},\ L_t = \{\boldsymbol{x} \in \Omega \mid u_t(\boldsymbol{x}) < h + \epsilon\}. \tag{3.1}$$

更に, 未分類集合を $U_t = \Omega \setminus (H_t \cup L_t)$ と定義しておく.

4. 提案法

入力がコストに応じたランダム性を持つ場合の LSE のための能動学習法を提案する. そのために, 次に評価すべき入力点および, その入力点の観測に費やすコストを決定するための獲得関数を与える. 本研究では, [2] と似た考え方を採用する. [2] の考え方は, 新たな点が加わったときの分類数と現在の分類数の差が, 期待値的に最も大きくなる点を次の評価点とするという考え方である. 一方, 本研究の設定は, 入力がコストに依存した不確実性を持つため, 彼らの獲得関数を直接利用はできない. そこで, 期待分類増加量の入力の分布に関する期待値をコストで割ったものを考える. すなわち, 単位コストあたりの, 入力の不確実性を考慮した, 期待値的な分類増加量を考える. 更に, ϵ-greedy の考え方と組み合わせることで, アルゴリズムが確率 1 で収束することを保証できる.

はじめに, 単位コストあたりの入力の不確実性を考慮した期待値的な分類増加量を定義する. 点 $\boldsymbol{s}^*$ をあらたな入力点とし, $y^* = f(\boldsymbol{s}^*) + \varepsilon$ が得られたとする. 組 $(\boldsymbol{s}^*, y^*)$ が追加されたときの, H および L の推定集合を $H_t(\boldsymbol{s}^*, y^*)$, $L_t(\boldsymbol{s}^*, y^*)$ と書く. このとき, 入力点 $\boldsymbol{x} \in \Omega$ の観測にコスト c_i をかけたときの, 単位コストあたりの入力の不確実性を考慮した, 期待値的な分類増加量 $a_t(\boldsymbol{x}, c_i)$ は以下で与えられる:

$$a_t(\boldsymbol{x}, c_i) = c_i^{-1} \int \mathrm{E}_{y^*}[|H_t(\boldsymbol{s}^*, y^*) \cup L_t(\boldsymbol{s}^*, y^*)| - |H_t \cup L_t|] g(\boldsymbol{s}^* | \theta_{\boldsymbol{x}}^{(c_i)}) d\boldsymbol{s}^*. \tag{4.1}$$

なお, 式 (4.1) 中の期待値は解析形を持つが, ここでは省略する.

次に, ϵ-greedy 戦略に基づいた, 組 $(\boldsymbol{x}, c_i)$ を確率的に選ぶ方法について説明する. 集合 $\mathcal{C}$ を $\mathcal{C} = \{(\boldsymbol{x}, c_i) \mid \boldsymbol{x} \in \Omega,\ i \in [k]\}$ としておく. 確率変数 C_t を $\mathcal{C}$ の元に値を取る離散確率変数とし, その確率関数は以下で与えられるとする:

$$p_{C_t}(\boldsymbol{x}, c_i) = \mathrm{P}(C_t = (\boldsymbol{x}, c_i)). \tag{4.2}$$

このとき, コスト毎に一様な確率を持つ以下の確率関数を考える:

Definition 4.1. 各 $i \in [k]$ に対し, κ_i を $0 < \kappa_i < 1$ を満たし, かつ $\sum_{i=1}^{k} |\Omega| \kappa_i = 1$ を満たすとする. このとき, $p_{C_t}(\boldsymbol{x}, c_i)$ を以下で定める:

$$p_{C_t}(\boldsymbol{x}, c_i) = \kappa_i. \tag{4.3}$$

最後に, 式 (4.1) と (4.3) を用いた提案アルゴリズムについて述べる. 本稿で提案するアルゴリズムは, 各試行時において, 確率 $1 - p_t$, $0 \leq p_t \leq 1$ で (4.1) を最大とする組 $(\boldsymbol{x}, c_i)$ を選び, 確率 p_t で (4.3) に基づき組 $(\boldsymbol{x}, c_i)$ を選ぶものである.

本研究における, いくつかの理論結果および数値実験の結果については当日報告する.

謝辞

本研究の一部は，科学研究費（17H00758, 16H06538）, JST CREST (JPMJCR1302,JPMJCR1502), 理化学研究所革新知能統合研究センター, JST イノベーションハブ構築支援事業・情報統合型物質・材料開発イニシアティブの補助を受けて行われた.

参考文献

[1] Alkis Gotovos, Nathalie Casati, Gregory Hitz, and Andreas Krause. Active learning for level set estimation. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 1344–1350, 2013.

[2] Andrea Zanette, Junzi Zhang, and Mykel J Kochenderfer. Robust super-level set estimation using gaussian processes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 276–291. Springer, 2018.

Predictive properties of forecast combination, ensemble methods, and Bayesian synthesis

Kosaku Takanashi* & Kenichiro McAlinn†

Forecast combination has recently received a surge in interest across multiple fields due to an increase in usage of more complex models, added with the increasing availability of density forecasts. In statistics, especially in the Bayesian literature, Bayesian model averaging (BMA) has been a staple, with theoretical justification under certain conditions and recent developments to topical problems in statistics. In machine learning, ensemble methods, including boosting, bagging, and stacking, have been used extensively to mitigate overfitting, which machine learning algorithms tend to do and benefit from in certain contexts. In econometrics, the field is stimulated by the increased availability of formal forecasting models that yield full density forecasts and the need to improve information flows to policy and decision makers.

We contribute to the growing literature of forecast combination, model averaging, and ensemble learning by developing a novel strategy based on continuous time stochastic processes to evaluate and assess the theoretical predictive properties of classes of combination strategies. Our development is motivated by the recently introduced Bayesian predictive synthesis (BPS) (McAlinn and West 2017; McAlinn et al. 2017), which is a general Bayesian framework for forecast combination that encompasses other methods as special cases. The motivation is driven by the fact that a certain class of BPS, proposed in the papers, substantially improve forecasts over standard and advanced benchmarks in the literature. We show that this class of synthesis defines a broader class of what we call *non-linear synthesis*. We further show that this class has properties that we identify as the source of improved predictive performance; namely an extra term in the stochastic process that acts as a shrinkage term on the error process. Finally, we prove that this class of BPS outperforms any and all linear combination of forecasts, including popular methods such as Bayesian model averaging, equal weight averaging, etc. We note that our development of using continuous time stochastic processes to evaluate predictive performances opens up several avenues of research that goes beyond the analysis conducted in this paper and has further potential to be applied to other contexts.

*RIKEN AIP, Tokyo, Japan. Email: kosaku.takanashi@riken.jp
†Booth School of Business, University of Chicago, Chicago, IL 60637. Email: kenichiro.mcalinn@chicagobooth.edu

References

McAlinn, K., Aastveit, K. A., Nakajima, J., West, M., 2017. Multivariate bayesian predictive synthesis in macroeconomic forecasting. arXiv preprint arXiv:1711.01667.

McAlinn, K., West, M., 2017. Dynamic bayesian predictive synthesis in time series forecasting. Journal of Econometrics Forthcoming.

再生核ヒルベルト空間における射影のモーメントによる二標本検定

千葉大・融合理工学府　牧草 夏実

1　はじめに

P, Q をヒルベルト空間 $\mathcal{H}$ 上の確率分布とするとき，二標本 $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} Q$ に基づく検定

$$\text{帰無仮説 } H_0 : P = Q \quad \text{vs.} \quad \text{対立仮説 } H_1 : P \neq Q$$

を考える．ユークリッド空間での二標本検定はすでに様々な検定方法が議論されているが，ヒルベルト空間に値をとる確率変数に対する二標本検定を考えることで，高次元データに対する二標本検定の議論を与える．高次元データに対するアプローチとして，カーネル法という方法がある．カーネル法を用いた二標本検定として，Maximum Mean Discrepancy(MMD) に基づく二標本検定が [1] によりすでに議論されているが，この MMD と同様の考え方により，2 次のカーネルの MMD による二標本検定を考える．

2　再生核ヒルベルト空間でのモーメントの定義

確率変数 $X \sim P$, $Y \sim Q$ を正定値カーネル k によって，この k に対応する再生核ヒルベルト空間 $H(k)$ 上に，それぞれ $k(\cdot, X)$，$k(\cdot, Y)$ により変換を行う．このとき，この $k(\cdot, X)$，$k(\cdot Y)$ の平均まわりの 2 次モーメント $\Sigma_k(P)$，$\Sigma_k(Q)$ は，それぞれヒルベルト空間 $H(k)^{\otimes 2} = H(k) \otimes H(k)$ での期待値 $\Sigma_k(P) = \mathbb{E}_{X \sim P}[(k(\cdot, X) - \mu(P))^{\otimes 2}]$，$\Sigma_k(Q) = \mathbb{E}_{Y \sim Q}[(k(\cdot, Y) - \mu(Q))^{\otimes 2}]$ によって定められている．ここで，$\mu(P)$，$\mu(Q)$ は $k(\cdot, X)$ の 1 次モーメント $\mu(P) = \mathbb{E}_{X \sim P}[k(\cdot, X)]$, $\mu(Q) = \mathbb{E}_{Y \sim Q}[k(\cdot, Y)]$ であり，$\otimes$ はテンソル積を表しており，任意の $f \in H(k)$ に対し，$f^{\otimes 2} = f \otimes f = \langle f, \cdot \rangle_{H(k)} f$ である．

3　検定統計量の構築

この $k(\cdot, X)$ と $k(\cdot, Y)$ の $f \in H(k)$ への射影のモーメント差

$$\sup_{||f||_{H(k)}=1} | \langle f, \mu(P) - \mu(Q) \rangle_{H(k)} | = ||\mu(P) - \mu(Q)||_{H(k)}$$

により，2 つの分布の違いを測るのが，Maximum Mean Discrepancy (MMD) と呼ばれるものである．同様の考え方により，$(k(\cdot, X) - \mu(P))^{\otimes 2}$ と $(k(\cdot, Y) - \mu(Q))^{\otimes 2}$ の $A \in H(k)^{\otimes 2}$ への射影のモーメント差

$$\sup_{||A||_{H(k)^{\otimes 2}}=1} | \langle A, \Sigma_k(P) - \Sigma_k(Q) \rangle_{H(k)^{\otimes 2}} | = ||\Sigma_k(P) - \Sigma_k(Q)||_{H(k)^{\otimes 2}}$$

により 2 つの分布の違いを測る．これは，MMD のようなある種の分布の違いを測っている．この違い $\|\Sigma_k(P) - \Sigma_k(Q)\|^2_{H(k)^{\otimes 2}}$ は

$$\widehat{T}^2 = \left\|\widehat{\Sigma}_k(P) - \widehat{\Sigma}_k(Q)\right\|^2_{H(k)^{\otimes 2}}$$

によって推定することができる．ただし，

$$\widehat{\Sigma}_k(P) = \frac{1}{n}\sum_{i=1}^{n}(k(\cdot, X_i) - \widehat{\mu}(P))^{\otimes 2}, \quad \widehat{\mu}(P) = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i),$$

$$\widehat{\Sigma}_k(Q) = \frac{1}{m}\sum_{i=1}^{m}(k(\cdot, Y_i) - \widehat{\mu}(Q))^{\otimes 2}, \quad \widehat{\mu}(Q) = \frac{1}{m}\sum_{i=1}^{m} k(\cdot, Y_i)$$

である．

本発表では，この検定統計量 $\widehat{T}^2$ の漸近挙動について報告を行った．特に，帰無仮説 $H_0 : P = Q$ のもとで，退化 V 統計量の結果に帰着させることで，$\widehat{T}^2$ の漸近分布が，独立な自由度 1 の χ^2 分布の重み付きつき無限和の形で得られること ([2] 参照)，その重みをデータに基づいて推定する方法について報告を行った．また，対立仮説 $H_1 : P \neq Q$ のもとでの $\widehat{T}^2$ の漸近分布が平均 0 の正規分布になっていることについて報告を行った．

参考文献

[1] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2007). A kernel method for the two sample problem. *Advances in Neural Information Procesing Systems*, **19** of *MIT Press*, *Cambridge*.

[2] Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. Wiley, New York.

Power Series Kernelsに基づく ノンパラメトリック学習のための変数選択法

松井孝太[1], 熊谷亘[1], 金森研太[2], 錦見満顕[3], 金森敬文[4,1]

[1] 理化学研究所 革新知能統合研究センター, [2] 名古屋工業大学, [3] 名古屋大学, [4] 東京工業大学

1 導入

機械学習の様々な問題において, 変数選択は, 学習したモデルのパフォーマンスの向上, 重要な特徴の選択, 解釈性などの観点から非常に重要なタスクである [4]. 線形モデルを含むパラメトリック学習の枠組みでは, 多くの変数選択のための方法が提案されている一方で, ノンパラメトリック学習における変数選択法の研究は, その重要性に比較して限定的である [1, 3]. 本研究では, カーネル法によるノンパラメトリック学習 (カーネルリッジ回帰, カーネル密度, 密度比推定などを含む) における変数選択法を提案する. 具体的には, power series kernel というクラスのカーネルを用いてモデリングし, non-negative garrote 型の罰則によって変数選択を行う. 理論的には, 提案法はゆるい条件の下で変数選択の一致性を持つことが示される. この性質は, オリジナルの (線形モデルに対する) non-negative garrote の変数選択の一致性の, カーネル法に基づく推定量 (非線形モデル) への拡張と位置づけられる. なお, 計算機実験の結果は当日報告する.

2 問題設定

$\mathcal{X}$ を d 次元サンプル空間, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ をカーネル関数とする. k に対応する再生核ヒルベルト空間を $\mathcal{H}$ と書き, その上の内積を $\langle f, g \rangle$, ノルムを $\|f\| = \sqrt{\langle f, f \rangle}$ と書く ($f, g \in \mathcal{H}$).

多くの学習アルゴリズムでは, ターゲットとする関数の推定は正則化付き経験損失最小化問題として定式化できる :

$$\min_{f \in \mathcal{H}} \hat{L}(f) + \lambda R(f) \tag{1}$$

本研究では, 特に損失として 2 次形式 $\hat{L}(f) = \frac{1}{2}\langle f, \hat{C} f \rangle - \langle \hat{g}, f \rangle$, 2 乗正則化項 $R(f) = \frac{1}{2}\|f\|^2$ を考える. カーネルリッジ回帰, カーネル密度比推定, カーネル密度推定, density-ridge 推定などのノンパラメトリック学習はこのクラスに属する [5]. カーネル法によるターゲット関数の推定量は $\hat{f}$ は通常全ての変数 $\boldsymbol{x} = (x_1, ..., x_d)$ に依存するが, ここでは真のターゲット関数 f^* は $s < d$ 個の変数のみに依存しているとし, これら f^* に寄与する変数を特定することを目的とする.

3 Adaptive scalingとpower series kernelによる変数選択

3.1 Kernel法における Adaptive Scaling

変数選択を実現するために, Breiman の non-negative garrote (NNG) による adaptive scaling を RKHS モデルに適用する [2]. $f \in \mathcal{H}$ と, $\boldsymbol{\xi} \in \mathbb{R}^d$ に対して, $f_{\boldsymbol{\xi}}(\boldsymbol{x}) := f(\boldsymbol{\xi} \circ \boldsymbol{x})$ と定める ($\circ$ は成分毎の積). $f_{\boldsymbol{\xi}}$ の推定量 $\hat{f}_{\hat{\boldsymbol{\xi}}}$ を導出するためのアリゴリズムを Algorithm 1 に示す.

Algorithm 1 Two-stage kernel-based estimator with NNG.

Input: Training samples, and regularization parameters, λ and η.

Step 1: Find the kernel-based estimator $\hat{f}$ by solving (1).

Step 2: Let us define $\hat{f}_{\boldsymbol{\xi}}(\boldsymbol{z}) = \hat{f}(\boldsymbol{\xi} \circ \boldsymbol{z})$. Find the optimal garrote parameter $\hat{\boldsymbol{\xi}}$ by solving

$$\min_{\boldsymbol{\xi}} \widehat{L}(\widehat{f_{\boldsymbol{\xi}}}) + \eta \|\boldsymbol{\xi}\|_1, \quad \text{s.t.} \ \ \boldsymbol{\xi} \in [0,1]^d.$$

Output: The estimator $\hat{f}_{\hat{\boldsymbol{\xi}}}(\boldsymbol{x})$.

3.2 Power Series Kernelとその不変性

提案法における統計モデルは, 一般には $\widetilde{\mathcal{H}} = \cup_{\boldsymbol{\xi}\in[0,1]^d} \mathcal{H}_{\boldsymbol{\xi}}$, $\mathcal{H}_{\boldsymbol{\xi}} = \{f_{\boldsymbol{\xi}}(\boldsymbol{x}) \mid f \in \mathcal{H}\}$ なる多重カーネルモデルとなる. もし, 任意の $\boldsymbol{\xi}$ に対して $\mathcal{H}_{\boldsymbol{\xi}}$ が不変性 ($\mathcal{H}_{\boldsymbol{\xi}} \subset \mathcal{H}$) を持つならば, $\widetilde{\mathcal{H}} = \mathcal{H}$ が成立し, 多重カーネルの計算を回避することができる. Power series kernel

$$k(\boldsymbol{x}, \boldsymbol{y}) := \sum_{\boldsymbol{\alpha}\in\mathbb{N}_0^d} \frac{w_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}} \boldsymbol{y}^{\boldsymbol{\alpha}}}{(\boldsymbol{\alpha}!)^2}, \quad \boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \cdots x_d^{\alpha_d}, \quad \boldsymbol{\alpha}! = \alpha_1! \cdots \alpha_d!$$

は上述の不変性を持つカーネルのクラスであり, 多項式カーネルや指数カーネルはこのクラスに含まれる. 一方で, ガウスカーネルはこのクラスには含まれず, 一般に不変性を満たさない.

4 変数選択の一致性

ターゲット関数 f^* は $s < d$ 個の変数のみに依存するので, 最適な NNG パラメータは $\boldsymbol{\xi}^* = (\boldsymbol{\xi}_1^*, \boldsymbol{\xi}_0^*) = (\mathbf{1}, \mathbf{0}) \in \mathbb{R}^s \times \mathbb{R}^{d-s}$ と書けることに注目する.

Theorem 1 ([5], Assumption 1∼3, Theorem 1)**.** いくつかの仮定の下で, 提案法による推定量 $\hat{\boldsymbol{\xi}}$ は以下の性質を持つ.

1. $\hat{\boldsymbol{\xi}}_1$ は $\boldsymbol{\xi}_1^*$ に確率収束する.
2. サンプルサイズが十分大きいとき, $\hat{\boldsymbol{\xi}}_0 = \boldsymbol{\xi}_0^*$ が高い確率で成り立つ.

参考文献

[1] Genevera I Allen. Automatic feature selection via weighted kernels and regularization. *Journal of Computational and Graphical Statistics*, 22(2):284–299, 2013.

[2] Leo Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[3] Zaili Fang, Inyoung Kim, and Patrick Schaumont. Flexible variable selection for recovering sparsity in nonadditive nonparametric models. *Biometrics*, 72(4):1155–1163, 2016.

[4] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[5] Kota Matsui, Wataru Kumagai, Kenta Kanamori, Mitsuaki Nishikimi, and Takafumi Kanamori. Variable selection for nonparametric learning with power series kernels. *Neural computation*, 31(8):1718–1750, 2019.

Component Retention in PCA Applied to Microarray Datasets

Atina Husnaqilati[1], Yohji Akama[2]

[1] *The Mathematical Institute, Tohoku University, Japan, husqila@gmail.com*
[2]*The Mathematical Institute, Tohoku University, Japan, yoji.akama.e8@tohoku.ac.jp*

Principal component analysis (PCA) is a multivariate statistical technique introduced by Karl Pearson [5]. Many fields, such as ecology, biology, economics, psychology, and zoology, employ PCA.

Suppose we have a $p \times n$ data matrix $X = [x_1, \ldots, x_n]$ where $x_j = [x_{1j}, \ldots, x_{pj}]^T$, $j = 1, \ldots, n$, are independent and identically distributed as a p-dimensional distribution with mean zero and positive definite covariance matrix. The sample covariance matrix is $S = n^{-1}XX^T$. The eigenvalues of sample covariance matrix are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

PCA finds new components to illustrate a great part of the variance in the data by using the sample covariance matrix S of that data. The new components are called *principal components* (PC). A principal component is a linear combination of the original variables. It corresponds to an eigenvector u of S. Then the variance of the PC is the eigenvalue λ corresponding to u. The greater λ is, the more "significant" the PC is. When reducing the dimension of data from p to q, we can choose q principal components corresponding to the q largest eigenvalues of S. PCA has been useful to various data with $p < n$.

Recently PCA is employed to the analysis of microarray datasets [1]. The microarray dataset contains the relative activity of thousand of genes from a single sample. Size n of observations in microarray datasets is less than the number p of variables. The microarray datasets have large number of variable because the single gene has many components to measure.

We consider the $n \times n$ dual sample covariance matrix $S_D = n^{-1}X^TX$ with $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0$ being the eigenvalues. Note

$$\lambda_i = \begin{cases} \mu_i & (1 \leq i \leq \min(p,n)) \\ 0 & (\min(p,n) < i \leq p). \end{cases} \tag{1}$$

The representation of large dimension p of microarray datasets is difficult, so we reduce the dimension p with PCA to q by taking q principal components.

The stopping rules are methods to determine the number q of important principal components. We concern two types of stopping rules, the first based on empirical results and the second based on statistical distributions. *Guttman-Kaiser*

criterion [7] and *Jolliffe's rules* [4] are the first types, while a *broken stick model* [1, 2] of the second type.

The question of stopping rules is to balance the accuracy (or fit) of the model with ease of analysis and the potential loss of information. In this paper, we apply the three methods to 16 publicly available microarray datasets [6] and summarize the results. Then, we have

$$\text{(Guttman-Kaiser)} > \text{(Broken stick model)} > \text{(Jolliffe's rule)}. \tag{2}$$

Our results contrast with result for $p < n$ datasets. For simulated datasets with $p < n$, Jackson [3] gave

$$\text{(Guttman-Kaiser)} > \text{(Broken stick model)}, \tag{3}$$

while Cangelosi and Goriely [1] did

$$\text{(Broken stick model)} < \text{another Jolliffe's rule } \min\left\{ i \;\middle|\; \lambda_i > \frac{0.7}{p}\sum_{i=1}^{p}\lambda_i \right\}.$$

In view of the inequality (2) and the inequality (3), broken stick model may deserve further study for general setting.

Acknowledgment

The first author gratefully acknowledge the scholarship for her master course study from the MEXT (Ministry of Education, Culture, Sports, Science and Technology) Japan.

References

[1] Cangelosi, R. & Goriely, A. (2007). Component Retention in Principal Component Analysis with Application to cDNA Microarray Data. *Biology Direct*. **2**:2.

[2] Frontier, S. (1976). Étude de la décroissance des valeurs propres dans une analyse en composantes principales: comparaison avec le modèle du bâton brisé. *Biol Ecol*. **25**. 67–75.

[3] Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*. **74**:8. 2204–2214.

[4] Jolliffe, I. (2002). *Principal Component Analysis*. New York: Springer.

[5] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Phil Mag*. **2**. 559–572.

[6] Ramey, J. (2013). *Datamicroarray*. R package version 1.14.4.

[7] Yeomans, K. A. & Golder, P. A. (1982). The Guttman-Kaiser criterion as a predictor of the number of common factor. *Journal of the Royal Statistical Society*. 3D **31**. 221–229.

超立方体に内接する超楕円球による線形回帰の解釈

下野寿之 (株式会社デジタルガレージ)

1　序論

p 個の $X_i\,(i=1,2,..,p)$ を独立変数、Y を従属変数とする。線形回帰モデル

$$\hat{Y}=\sum_{i=1}^{p}\hat{a}_iX_i+\hat{b}$$

は広く用いられるが、$p\geq 2$ となった途端に、重回帰をすると、回帰係数 $\hat{a}_i$ の値の正負と大小が直感に反したり、その値の意味の解釈が難しくなったり、多重共線性が発生したりするなどの困難が知られている。$i=1,..,p$ に対して ρ_i を X_i と Y のピアソンの積率相関係数とすると、重相関係数は下限 $\max\{|\rho_1|,|\rho_2|,..,|\rho_p|\}$ と上限 1 の間に値を取るが、どのような場合にその下限または上限に一致したり近かったりするかを、計算機による算出無しに直感的に判断するのは難しい場合は多い。

これらの問題は、機械学習で線形モデリングを含んだ手法を実行した場合にも同様に現れるが、下記に示すような幾何学的な作図を用いた 3 個の新しい鳥瞰的な定理による新しい観点により、解決の可能性があると考えられる。ユークリッド幾何学的な発想を用いて、次ページに示唆されるように様々な有用な命題を導くことが期待できるからである。

2　新記号の定義

(1) $X_i^{\wedge}$ とは、$(X_1,..,X_{i-1},X_{i+1},..,X_p)$ すなわち X の持つ p 個の変数の内、X_i 以外の $p-1$ 個の変数を集めたものとする。(2) $R_{A\vdash B}$ とは、変数の集まり A で変数 B を重回帰した場合の重相関係数とする。(3) $A=(A_i)_{i=1}^{I},B=(B_j)_{j=1}^{J}$ がそれぞれ変数の並びの場合 ($I=1$ または $J=1$ の場合もありうる)、$R_{A\times B}$ とは、I 行 J 列の行列で、第 (i,j) 要素は変数 A_i と B_j のピアソンの積率相関係数とする (通常、$A=B$ の場合に「相関行列」と呼ばれる)。

3　既知の結果

全ての X_i と Y を標準化 (平均が 0 で分散が 1) したとする。$R_{X\times X}^{-1}$ の存在を仮定する。すると、(1) X_i に対応した偏回帰係数 $\hat{a}_i$ は p 次元ベクトル $R_{X\times X}^{-1}R_{X\times Y}$ の i 番目の成分に等しい [1, 68 ページと 78 ページ]。(2) 重相関係数 $R_{X\vdash Y}$ は $(R_{Y\times X}R_{X\times X}^{-1}R_{X\times Y})^{1/2}$ に等しい [1, 73 ページ]。(3) $X_i^{\wedge}$ の影響を除いた X_i と Y の間の偏相関係数は、$p\neq 1$ の場合に $(R_{X_i\times Y}-R_{X_i\times X_i^{\wedge}}R_{X_i^{\wedge}\times X_i^{\wedge}}^{-1}R_{X_i^{\wedge}\times Y})/\{(1-R_{X_i^{\wedge}\vdash X_i}^2)(1-R_{X_i^{\wedge}\vdash Y}^2)\}^{1/2}$[1, 77 ページ] に等しい。(C. R. Rao [2] や T. W. Anderson [3] の本を含め多数の文献でこれらの式を調べたが、偏相関係数の式については上記の式に容易に変形できるものは [1] の他には見つからなかった。)

4　定理の構成

上記から、以下に述べるユークリッド幾何学的な定理を導くことが出来る [4]。

> **補題:** M を $p\times p$ の正定値対称行列で対角成分は全て 1 とする。空間 $\mathbb{R}^p$ に直交座標系を導入し x_1 軸, x_2 軸, .., x_p 軸があるとする。そして、$x^{\top}M^{-1}x=1$ を満たす x の集合を考える。それは、$x_i=\pm 1\,(i=1,..,p)$ で表される $2p$ 枚の超平面で囲まれた超立方

体 S に内側から接する超楕円球面 E である。M を p 次元の縦ベクトル p 本に分解して $(R_1|\ldots|R_p)$ とする。S と E の接点の集合 $S\cap E$ は $2p$ 個の点の集合 $\{\pm R_1, \pm R_2, .., \pm R_p\}$ と等しい。

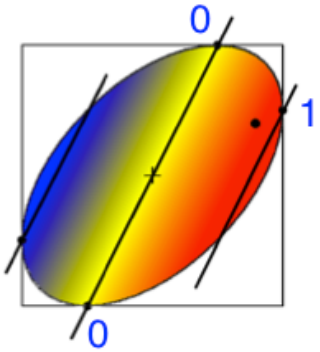
この補題の M に $R_{X\times X}$ を代入し、まず、超立方体 S と超楕円球面 E とそれらの接点 $+R_1, .., +R_p$, $-R_1, .., -R_p$ を得る。そして新たな点 P の座標を $R_{X\times Y}(=(\rho_1, .., \rho_p))$ とする。

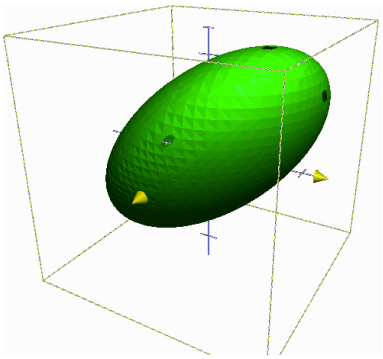
定理 1 (重相関係数): 原点 O と点 P を通る直線と E (超楕円球面) との交点を P' とすると、重相関係数 $R_{X\vdash Y}$ は長さの比 $|OP|/|OP'|$ に等しくなる。

定理 2 (偏回帰係数): 各 $i=1, .., p$ に対して、線形関数 $f_i:\mathbb{R}^p\to\mathbb{R}$ を、$\{O\}\cup(S\cap E)\backslash\{\pm R_i\}$ の任意の点では値が 0 で、点 R_i では値が 1 を与えるようにする。すると $f_i(P)$ は X_i と Y を標準化した上での偏回帰係数 $\hat{a}_i$ に等しい。

定理 3 (偏相関係数): 各 $i=1, .., p$ に対して、 $P_i^-, P_i^+\in S$ は、線分 $P_i^-P_i^+$ が x_i 軸と平行かつ同じ向きで P を通るとする。アフィン関数 $g_i:\mathbb{R}\to\mathbb{R}^p$ が $g_i(\pm 1)=P_i^\pm$ (復号同順) を満たすようにする。変数 $X_1, .., X_{i-1}, X_{i+1}, .., X_p$ の影響を除いた X_i と Y の間の偏相関係数は $g_i^{-1}(P)$ に等しい。

上記のことは、次元 p が 2 または 3 の場合は、下にある様な図で考えることが出来る。正方形と立方体が S に相当し、楕円と楕円球が E に相当する。S と E は $2p$ 個の点 $\pm R_i\,(i=1, .., p)$ で接している。E の内部に点 P が打点されている ($R_{(X,Y)\times(X,Y)}$ は半正定値であることによる)。P が原点 O に近いか E に近いかで重相関係数が決まったり、関数 f_i に 0 や ± 1 の値を与えさせる超平面の考察から標準化回帰係数が ± 1 の範囲をどんな場合にどの程度離れるかを幾何的に導いたり、線分 $P_i^-P_i^+$ を物差し (定規) と見なし -1 から $+1$ までの目盛りを入れると偏相関係数を読んだりすることが出来る。





参考文献

[1] 竹内啓, 柳井晴夫 (1972). **多変量解析の基礎**, 東洋経済新報社.

[2] Calyampudi Radhakrishna Rao (1973). *Linear Statistical Inference and Its Applications*, Wiley Series in Probability and Statistics.

[3] Theodore Wilbur Anderson (2003). *An Introduction to Multivariate Statistical Analysis*, Wiley Series in Probability and Statistics.

[4] Toshiyuki Shimono (2019). Interpreting Multiple Regression via an Ellipse Inscribed in a Square Extensible to Any Finite Dimensionality, *In Proceedings of Data Science, Statistics & Visualisation 2019 (DSSV 2019) in Kyoto, Japan, page 119.*

マスター方程式から厳密に得られる期待値変数の情報幾何学, 接触幾何学およびその周辺

後藤 振一郎, 日野 英逸

情報 · システム研究機構　統計数理研究所

概 要

本報告では, 時間連続マスター方程式から厳密に得られる期待値変数に対する力学系を導入し, その力学系の微分幾何学を用いた記述を行う. 特にその微分幾何学として平衡系に対しては情報幾何学, 非平衡系に対しては接触幾何学を用い, かつ期待値変数に対する力学系は接触幾何学における接触ハミルトン力学系を用いる.

導入

情報幾何学は数理統計学の微分幾何学化として知られており, その微分幾何学的側面や応用の探察や拡張が行われている. その例は様々である. 本報告では特に接触幾何学と呼ばれる奇数次元版シンプレクティック幾何学と情報幾何学を有機的に統合する研究に着目する. この文脈において, 疑リーマン計量等を導入した場合, パラ接触計量多様体の一つであるパラ佐々木多様体が熱力学相空間を記述するのに適していることが判明している. これらを鑑みると, パラ接触計量多様体がどのように非平衡熱統計力学を記述するかが興味の対象になる. 特にマスター方程式は数学的に非自明でありながら単純で, 非平衡熱力学で頻繁に用いられているので, これに着目する. またマスター方程式はモンテカルロ法の基礎ともなるため, 数理工学への応用も期待できる.

本研究では連続時間マスター方程式から期待値変数を記述する力学系を導出する. そしてその非平衡熱力学過程をパラ接触計量多様体上で構成する. 本稿の大部分は文献 [1, 2] をもとにしている. 本稿で用いる用語の定義はそれらの文献や予稿に記述があるので必要に応じて参照されたい.

可解マスター方程式

この節ではマルコフ核を選ぶことにより, あるクラスのマスター方程式を導入する. そして, その可解性を示す. 集合 Γ を離散状態を要素にもつ有限集合, $t \in \mathbb{R}$ を時間, そして $p(j,t)\,\mathrm{d}t$ を時刻 t と $t+\mathrm{d}t$ の間に状態 $j \in \Gamma$ を見つける確率とする. 第一目標は, 与えられた分布関数 $p_\theta^{\mathrm{eq}}(j) = \frac{\pi_\theta(j)}{Z(\theta)}$, $Z(\theta) := \sum_{j\in\Gamma}\pi_\theta(j)$, を実現することである. ただし, Θ はパラメーター $\theta = \{\theta^1, \ldots, \theta^n\}$ を要素にもつ集合で $Z : \Theta \to \mathbb{R}$ はいわゆる分配関数で p_θ^{eq} が規格化するために用いられる: $\sum_{j\in\Gamma} p_\theta^{\mathrm{eq}}(j) = 1$.

以下ではあるクラスのマスター方程式に着目する. 時間依存する分布関数を $p : \Gamma \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ とする. そして, マスター方程式とよばれる次を考える

$$\frac{\partial}{\partial t}p(j,t) = \sum_{j'(\neq j)} \left[\, w(j|j')\,p(j',t) - w(j'|j)\,p(j,t) \,\right], \tag{1}$$

ここで $w : \Gamma \times \Gamma \to I$, $(I := [0,1] \subset \mathbb{R})$ は $w(j|j')$ が状態が j' から j へジャンプする確率を表す. 式 (1) と仮定 $w_\theta(j|j') = p_\theta^{\mathrm{eq}}(j)$, および, $p_\theta^{\mathrm{eq}}(j) \neq 0$, $\forall j \in \Gamma$, により, 可解マスター方程式

$$\frac{\partial}{\partial t}p(j,t) = p_\theta^{\mathrm{eq}}(j) - p(j,t). \tag{2}$$

を導出することができる. 解 $p(j,t)$ の陽な表式は (2) を解くことによって得られる. そして解 p は θ に依存することに気がつく. これを考慮し, $p(j,t)$ は $p(j,t;\theta)$ と書くことにする. また,(2) により平衡状態が時間に関して漸近極限で実現されていることが確認できる.

観測量の時間発展とその幾何学

本節では観測量の時間発展を記述する微分方程式をある仮定のもと可解マスター方程式から導出する. そして, それらの観測量の漸近極限について述べる. ここで本稿における 観測量とは状態 (や確率変数) に依らない関数として定義しておく. それゆえ, 分布関数に関する期待値変数は観測量である.

$\mathcal{O}_a : \Gamma \to \mathbb{R}$ $(a \in \{1, \dots, n\})$ を関数, そして $p : \Gamma \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ を可解マスター方程式 (2) に従う分布関数とする. すると $\langle \mathcal{O}_a \rangle_\theta(t) := \sum_{j\in\Gamma} \mathcal{O}_a(j)\, p(j,t;\theta)$, および $\langle \mathcal{O}_a \rangle_\theta^{\mathrm{eq}} := \sum_{j\in\Gamma} \mathcal{O}_a(j)\, p_\theta^{\mathrm{eq}}(j)$, は $\mathcal{O}_a$ の期待値と呼ばれる.

もし平衡分布関数が指数分布族に属していれば, 以下で指定される関数 $\Psi^{\mathrm{eq}} : \Theta \to \mathbb{R}$ と $\Psi : \Theta \times \mathbb{R} \to \mathbb{R}$

$$\Psi^{\mathrm{eq}}(\theta) := \ln\left(\sum_{j\in\Gamma} \mathrm{e}^{\theta^b \mathcal{O}_b(j)}\right), \qquad \Psi(\theta,t) := \left(\frac{1}{J^0}\sum_{j\in\Gamma}\frac{p(j,t;\theta)}{p_\theta^{\mathrm{eq}}(j)}\right)\Psi^{\mathrm{eq}}(\theta), \quad \text{where} \quad J^0 := \sum_{j'\in\Gamma} 1. \tag{3}$$

は様々な役割を果たす. これ以降では (3) が存在することを仮定する. 情報幾何学の文脈ではこの関数は θ-ポテンシャルとして引用される. 本稿では離散分布関数が考察の対象であり, 従って分布関数は指数分布族に属する. 従って,(3) における Ψ^{eq} は本稿でも重要な役割を果たす. 値 $\Psi^{\mathrm{eq}}(\theta)$ は負の無次元自由エネルギーである. 期待値と $\{\langle \mathcal{O}_a \rangle_\theta\}$ と Ψ に対する微分方程式は以下のように導出される.

命題 1. *([1]).* θ を時間依存しないパラメーターの組でこれにより分布 p_θ^{eq} をパラメタライズする. すると, $\{\langle \mathcal{O}_a \rangle_\theta\}$ および Ψ は $\mathbb{R}^{2n+1}$ 上の以下の微分方程式の解となる:

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^a = 0, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\langle \mathcal{O}_a \rangle_\theta = -\langle \mathcal{O}_a \rangle_\theta + \frac{\partial \Psi^{\mathrm{eq}}}{\partial \theta^a}, \quad \textit{and} \quad \frac{\mathrm{d}}{\mathrm{d}t}\Psi = -\Psi + \Psi^{\mathrm{eq}},$$

この力学系をモーメント力学系と名付ける. モーメント力学系は接触ハミルトン力学系である :

命題 2. *([1]).* 命題 *1* での力学系は接触幾何学で知られる接触ハミルトン力学系として記述される.

そしてこの命題をパラ接触計量多様体の上で考察すると, 以下が示される :

定理 1. *([2]).* 命題 *2* における接触ハミルトンベクトル場を X_h, その接触ハミルトニアンを h とする. またパラ接触計量多様体における $(1,1)$-型テンソル場を ϕ とすると, $\phi(X_h)$ や $\phi^2(X_h)$ に沿って h は保存する $\mathcal{L}_{\phi(X_h)}h = \mathcal{L}_{\phi^2(X_h)}h = 0$.

非平衡統計力学では状態がどれほど平衡状態に近いか, しばしば着目される. 一般にそのような距離を幾何学を用いて定義するには曲線の長さを使うことができる. リーマン幾何学において, ベクトル場の積分曲線で結ばれた多様体上の 2 点間の距離を計算する手法が確立されている. 以下の本稿での主張が得られる.

定理 2. *(*期待値変数の幾何学的記述とその収束*, [2]).* 可解マスター方程式から導出されたモーメント力学系はパラ接触計量多様体の上で記述され*, Mrugala* 計量テンソル場と呼ばれる計量に付随した収束は指数型である.

謝辞

本研究の実施にあたり, 著者 S.G. は日本学術振興会から JSPS 科研費 19K03635 の部分的助成を受けました. また, 著者 H.H. は同じく日本学術振興会から JSPS 科研費 17H01793 の部分的助成を得ました. これらに加え, 両著者は JST CREST JPMJCR1761 の部分的助成を受けました.

参考文献

[1] S. Goto and H. Hino, *Information and contact geometric description of expecation variables exactly derived from master equations*, arXiv:1805.10592v2, Submitted.

[2] S. Goto and H. Hino, *Expectationo variables on a para-contact metric manifold exactly derived from master equations*, GSI 19 proceedings, Lecture Notes in Computer Science, to appear, (2019).

Similarity Measures and Statistical Models in Recommendation Problems

Kotaro Sudo (NS Solutions Co.), Naoya Osugi (Recruit Technologies Co.,Ltd.), and Takafumi Kanamori (Tokyo Institute of Technology)

In this paper, we study recommendation problems, in particular, the *reciprocal recommendation*. The reciprocal recommendation is regraded as a kind of edge prediction problem of random graphs. For example, the job recruiting service provides preferable matches between companies and job seekers. The corresponding graph is a bipartite graph, and nodes are categorized into two groups; one is job seekers and the other is companies. Directed edges from one group to the other one mean the expression of user's interests. The job recruiting service recommends unobserved potential matches between users and companies. Another common example is online dating services. Again, the corresponding graph is expressed as a bipartite graph with two groups, i.e., males and females. The directed edges mean the preference expressions among users. Then, the recommendation system provides potentially preferable partners to each user. The quality of such services totally depends on the prediction accuracy of unobserved or newly added edges. The edge prediction has been widely studied as a class of important problems in social networks [15, 10, 8, 11, 1].

In recommendation problems, it is often assumed that *the similar people like or dislike similar items, people, etc.* Based on this assumption, researchers have proposed many similarity measures. The similarity is basically defined through the topological structure of the graph that represents the relationship among users or items. Neighbor-based metrics, path-based metrics, and random walk based metrics are commonly used. Then, a similarity matrix defined from the similarity measure is used for the recommendation. As the other approach, statistical models such as stochastic block models [13] are used in order to estimate network structures such as clusters or edge distributions. The learning methods using statistical models often achieve high prediction accuracy in comparison to similarity-based methods. Details are reported in [12] and references therein.

The main purpose of this paper is to investigate the relationship between similarity-based methods and statistical models. We show that a class of widely applied similarity-based methods can be derived from the *Bernoulli mixture models* or from *the stochastic block models* in general. More precisely, the Bernoulli mixture model with the Expectation-Maximization (EM) algorithm [5] naturally derives a completely positive matrix [2] as the similarity matrix. The class of completely positive matrices is a subset of doubly nonnegative matrices, i.e., positive semidefinite and element-wise nonnegative matrices [3]. Also, we provide an interpretation of completely positive matrices as a statistical model satisfying exchangiability [6, 16, 7, 4]. Based on the above argument, we connect the similarity measures using completely positive matrices to statistical models. First, we prove that most of commonly

used similarity measures yield completely positive matrices as the similarity matrix. Then, we propose an algorithm that transforms the similarity matrix to the Bernoulli mixture model. As a result, we obtain a statistical interpretation of similarity-based methods through Bernoulli mixture models. We conduct numerical experiments using synthetic data and real-world data provided from an online dating site, and report the efficiency of the recommendation method based on Bernoulli mixture models.

This work is based on [9, 14].

References

[1] Deepak K. Agarwal and Bee-Chung Chen. *Statistical Methods for Recommender Systems*. Cambridge University Press, 2016.

[2] A. Berman and N. Shaked-Monderer. *Completely Positive Matrices*. World Scientific Publishing Company Pte Limited, 2003.

[3] Samuel Burer, Kurt M. Anstreicher, and Mirjam Dür. The difference between 5×5 doubly nonnegative and completely positive matrices. *Linear Algebra and its Applications*, 431(9):1539–1552, 2009.

[4] Bruno de Finetti. *Theory of Probability*. Wiley, 1970.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[6] P. Diaconis. Finite forms of de finetti's theorem on exchangeability. *Synthese: An International Journal for Epistemology, Methodology and Philosophy of Science*, 36:271–281, 1977.

[7] P. Diaconis and D. Freedman. Finite exchangeable sequences. *The Annals of Probability*, 8(4):745–764, 1980.

[8] Mohammad Al Hasan and Mohammed J. Zaki. A survey of link prediction in social networks. In *Social Network Data Analytics*, pages 243–275. 2011.

[9] Takafumi Kanamori and Naoya Osugi. Model description of similarity-based recommendation systems. *Entropy*, 21(7), 2019. 702.

[10] David Liben-nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 2007.

[11] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A*, 390(6):1150–1170, 2011.

[12] Wang Mengdi. Vanishing price of decentralization in large coordinative nonconvex optimization. *SIAM Journal on Optimization*, 27(3):1977–2009, 2017.

[13] Natalie Stanley, Thomas Bonacci, Roland Kwitt, Marc Niethammer, and Peter J. Mucha. Stochastic block models with multiple continuous attributes, 2018. arXiv:1803.02726.

[14] Kotaro Sudo, Naoya Osugi, and Takafumi Kanamori. Numerical study of reciprocal recommendation with domain matching. *Japanese Journal of Statistics and Data Science*, 2(1):221–240, 2019.

[15] Peng Wang, BaoWen Xu, YuRong Wu, and XiaoYu Zhou. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, 58(1):1–38, 2015.

[16] G. R. Wood. Binomial mixtures and finite exchangeability. *The Annals of Probability*, 20(3):1167–1173, 1992.

高次元遺伝子解析の呪いからの解放

成蹊大学名誉教授　　新村秀一

「高次元 Microarray データを用いて癌遺伝子の特定と癌の亜種を見つける研究」が 1970 年頃から行われてきた(Golub, 1999)。これらの研究で用いられたデータが公開されているので、統計に限らず機械学習（AI)、パターン認識、Bio 工学の新テーマとして研究されてきたが、いずれの研究も成功していない(判別分析の **Problem5**)。しかし、仮に症例数が n=100 で遺伝子数が p=10,000 の発現量データとすれば、2 群判別が最も適した手法である。筆者は 2015 年 10 月 28 日から 12 月 20 日の僅か 54 日間で、簡単にこの問題を解決した。用いたデータは、1999 年から 2004 年の間に米国の 6 医学研究プロジェクトが論文を発表し、研究に用いた公開データである。これらは癌と健常、あるいは 4 種の異なった癌の 2 クラスである。

結果は非常に単純である。6 種のデータは線形分離可能なデータ(Linearly Separable Data, **LSD**)である（判別分析の **Fact3**)。この重要な事実であり信号が、これまでの研究で誰も指摘していない。さらに、筆者が開発した Matryoshka Feature Selection Method **(Method2)**で簡単に線形分離可能な含まれる遺伝子数が n 個以下の遺伝子の k 組の部分空間(SMall Matryoshka, **SM**)と最小誤分類数(Minimum Number of MisclassificationSM**NM**)が 1 以上の遺伝子の雑音部分空間に分割できた(**Fact4**)。各 SM は統計分析が容易な小標本であるが、ロジスティック回帰以外の統計手法（一元配置の分散分析、 t 検定、相関分析、クラスター分析、PCA）で線形分離可能な事実が示されなかった(**Problem6**)。そこで、MNM 基準による改定 IP-OLDF（**RIP**）の判別スコア(RIP Discriminant Score, **RipDS**)を変数とし、n 症例*k 次元の**信号データ**(k<=n)を作成した。これを上記の統計手法で分析し「癌の遺伝子診断の統計分析法を世界で初めて提案」できた。今年 5 月に Springer から **Springer2** を出版した。

以上の研究が簡単にできたのは、大学卒業以来行ってきた判別分析の新理論(**Springer1**）が 2015 年に完成し、新理論がその応用問題として 1970 年頃から未解決の Microarray を用いた癌の遺伝子解析（**Problem5**）を簡単に解決できた。本来であれば癌の遺伝子研究の専門家でない筆者が「癌の遺伝子診断」までを行うことは適していない。しかし、癌は遺伝子の病気であり、高次元の Microarray 空間で 2 群が完全に分かれていて、さらに MNM ＝ 0 である k 組の SM に分割できる。そして RipDS で信号データを作ることで、上記の統計手法で線形分離可能であり有効と考えられる Malignancy Indexes が数多く発見できた。しかし、これ等のどの Malignancy Indexes が医学的に役に立つか否かは医学専門家の検証が必要である。

残念ながら Golub らの研究後に、「**NIH が乳がん以外の癌に関して** Microarray による**研究は成果が出ないと判断し、医学研究が終わったようである**」。このため、いかに医学専門家の検証につなげるかを 2016 年から模索している。しかし統計や工学研究者は、NIH の報告を知らずに研究を続けているのは一般的に問題であろう。また、データが LSD であるのに、そのデータを学習標本に用いた AI 研究が LSD の事実を指摘しない点だけが、まだ説明できていない。

大学卒業以来の研究テーマである判別分析の新理論を確立し、その応用として「高次元 Microarray の癌の遺伝子解析と診断」にはじめて成功した。そこでこれまでの研究を見直した結果、LSD である高次元データは、ケース数 n 個以下の遺伝子の k 組の小標本に必ず分割できるという事実が統計にとって一番重要と考えた。すなわち、我々は「**高次元データの呪いから数理計画法(MP)の LP と IP で定式化した LDF で解放される**（2 次計画法 QP で定式化した SVM ではできない)**。そして、分割された SM を統計分析するとこれまで見えてこなかった新しい癌の遺伝子診断の世界が広がる**」。

以上の重要なテーマを以下の 4 回のシンポジュームで報告したい。

1)　九州大学：「高次元遺伝子解析の呪いからの解放 1　-統計が 1970 年からこの問題を解決できなかった理由-」

2) 新潟大学：「高次元遺伝子解析の呪いからの解放 2　-癌の遺伝子診断-」

3) 東京工業大学：「高次元遺伝子解析の呪いからの解放 3　-機械学習などの工学研究の問題点-」

4) 秋田大学：「高次元遺伝子解析の呪いからの解放 4　-高次元データの分割法の最新結果-」

予稿集は、以下の 6 章からなり、1 章と 2 章は共通の基礎知識、3 章から 6 章は 4 大学での発表に対応していて、各大学の予稿はその章しか含んでいないで、必要であれば他大学の予稿を参考にしてほしい。

1 章では、Springer1 と「新村(2010).　最適線形判別関数．日科議連」の中から、統計的判別関数がなぜ癌の遺伝子解析に役に立たなかったかの理由を報告する。すなわち LSD 判別は、MNM 基準による改定 IP‐OLDF(**RIP**)とハードマージン最大化 SVM（**H-SVM**）でしか理論的に正しくできないことが原因である。回帰係数や判別係数を 0 にすることで、Problem5 に対応できると考える LASSO 研究の間違いを指摘する

2 章では、1970 年頃から解決できなかった Problem5 の結果を紹介する。

3 章では、高次元の Microarray がなぜ n 個以下の MNM＝0 である k 組の SM に簡単に分割できるかを数理計画法(MP)の基礎知識（新村(2010).　数理計画法による問題解決法．日科議連」）と連立方程式の解の基礎知識の簡単な組み合わせで説明する。そして、統計的判別関数と 2 次計画法 QP で定式化された H－SVM が高次元の LSD である Microarray を k 組の SM に分割できない理由を説明する。

4 章では、本研究における 8 種類の LDF の役割を概観する。Microarray が LSD である Fact3 は、RIP、H-SVM、Revised LP-OLDF と SVM4(C=10^4)で発見できる。そして RIP と Revised LP-OLDF が高次元の呪いから研究者を解放し、SVM ができない理由を示す。高次元 Microarray は k 組の n 個以下の遺伝子の部分空間の SM に分割できる。これらは小標本であり統計手法で簡単に分析し癌の遺伝子診断が行えると考えた。しかし、ロジスティック回帰だけが全ての SM がＮＭ＝0 で、LSD であることが分かる。しかし他の統計手法で LSD の事実が得られなかった（**Problem6**）。試行錯誤の末、信号データを作成してこれを解決し、癌の遺伝子診断が可能になった。しかし LSD である Microarray データを学習に用いているのに、なぜ AI 研究は LSD の事実を発見できないかを参加者と議論したい。6 月開催の IEEE の機械学習の国際会議で良い情報が得られればそれも報告する。

5 章では、世界で初めて成功した癌の遺伝子診断の結果の概略を Springer２から説明する。

6 章では、Method２が Microarray データだけでなく、6 変数の普通車と小型車の 2 群判別にも適用できることを示す。即ち本研究は、LSD は必ずより小さい SM や最小次元の Basic Gene Sets(**BGS**)に分割できることが今後の統計にとって重要なことを示す。そして Method2 が求めた SM をさらに BGS（iPS 研究の山中 4 因子と同じ概念）に分割すると、多くの場合に 2 個の BGS が含まれることが分かった。

筆者の方法によって、今後高次元の他の LSD であっても、容易に統計分析の研究対象になる。本研究は、質が高く、2 群が LSD であるという検証しやすいデータを用いたことで、LINGO[2]と JMP[1]の組み合わせで初めて役に立つ研究を退官後に完成できたことは医学データを研究対象としたことが幸運であったと考える。

1　新村秀一　(2004).『JMP活用　統計学とっておき勉強法』．講談社.

2　新村秀一 (2010).『最適線形判別関数』．日科技連出版.

3　新村秀一 (2011).『数理計画法による問題解決法』．日科技連出版.

4　Shinmura S (2000b). Optimal Linear Discriminant Function using Mathematical Programming. Disertation, Okayama Univ.

5　Springer1: Shinmura S (2016). The New Theory of Discriminant Analysis after R Fisher, Springer. DOI: 10.1007/978-981-10-2164-0

6　Springer2: Shinmura S (2019a) High Dimensional Microarray Data Analysis – Cancer Gene Diagnosis and Malignancy Indexes by Microarray. Springer.

確率密度関数のモード探索とその応用

はこだて未来大学複雑系知能学科　佐々木 博昭

1　はじめに

統計的データ解析において，確率密度関数のモード（極大点）を推定・探索することは重要な研究課題である．例えば，モード探索クラスタリングは，推定した確率密度関数のモード点へ向けてデータ点を更新し，同じモード点へ収束したデータ点に対して，同じクラスタリングラベルを割り当てる [1]．確率密度関数のモードを利用した他のデータ解析手法は，文献 [2, 3] を参照せよ．

確率密度関数のモードを推定・探索する上で重要なタスクの 1 つが，確率密度関数の微分を推定することである．最も単純なアプローチは，最初にデータの確率密度関数を推定し，次にその密度関数の推定結果の微分を計算する 2 段階推定であろう．しかしながら，この 2 段階推定は，確率密度関数の微分を推定する上で適切なアプローチではない．何故ならば，良い確率密度関数の推定結果が，必ずしも良い確率密度関数の微分推定結果をもたらすとは限らないからである．

より適切なアプローチは，確率密度関数の推定を実行せず，直接的に確率密度関数の微分を推定することであろう．ここでは，この直接推定の考えに基づく微分推定法とそのモード推定への応用を紹介する [4, 5, 6]．微分推定は再生核ヒルベルト空間の理論に基づき実践的な方法が構築され，そして，その推定法を用いたモード探索のための逐次的な更新式を紹介する．

2　対数密度微分の直接推定 [4, 6]

確率密度関数 $p(\boldsymbol{z})$ より生成された n 個のデータ点 $\{\boldsymbol{z}_i = (z_i^{(1)}, z_i^{(2)}, \ldots, z_i^{(d_z)})^\top\}_{i=1}^n$ が与えられているとする．$^\top$ は転置を意味する．基本的なアプローチは，モデル $g^{(j)}(\boldsymbol{z})$ を真の対数密度微分へ $\partial_j \log p(\boldsymbol{z})$ を直接適合することである．

$$
\begin{aligned}
J_j(g^{(j)}) &= \frac{1}{2}\int \{g^{(j)}(\boldsymbol{z}) - \partial_j \log p(\boldsymbol{z})\}^2 p(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \\
&= \frac{1}{2}\int \{g^{(j)}(\boldsymbol{z})\}^2 p(\boldsymbol{z})\mathrm{d}\boldsymbol{z} + \int \left\{\partial_j g^{(j)}(\boldsymbol{z})\right\} p(\boldsymbol{z})\mathrm{d}\boldsymbol{z} + \frac{1}{2}\int \{\partial_j \log p(\boldsymbol{z})\}^2 p(\boldsymbol{z})\mathrm{d}\boldsymbol{z}.
\end{aligned}
$$

上式は，$\lim_{|z^{(j)}|\to\infty} |g^{(j)}(\boldsymbol{z})p(\boldsymbol{z})| \to 0$ という仮定下で部分積分を実行し，導出されている．$\{\boldsymbol{z}_i\}_{i=1}^n$ を用いて，$g^{(j)}$ に無関係な右辺第 3 項を除いた J_j は次のように推定できる．

$$
\widehat{J}_j(g^{(j)}) = \frac{1}{2n}\sum_{i=1}^n \left[g^{(j)}(\boldsymbol{z}_i)^2 + 2\partial_j g^{(j)}(\boldsymbol{z}_i)\right].
$$

次に，$g^{(j)}$ を再生核ヒルベルト空間 $\mathcal{H}$ に制限し，推定量の導出を行う．k を $\mathcal{H}$ におけるカーネル関数とすると，微分に関する表現定理 [7] により，g_j の最適な形式は次で与えられる．

$$
g^{(j)}(\boldsymbol{z}) = \sum_{i=1}^n \alpha_i^{(j)} k(\boldsymbol{z}, \boldsymbol{z}_i) + \beta_i^{(j)} \partial'_j k(\boldsymbol{z}, \boldsymbol{z}')\big|_{\boldsymbol{z}'=\boldsymbol{z}_i} = \boldsymbol{\alpha}^{(j)\top}\boldsymbol{k}(\boldsymbol{z}) + \boldsymbol{\beta}^{(j)\top}\partial'_j \boldsymbol{k}_j(\boldsymbol{z}). \tag{1}
$$

式 (1) における $\partial'_j k(\boldsymbol{z}, \boldsymbol{z}_i) := \partial'_j k(\boldsymbol{z}, \boldsymbol{z}')\big|_{\boldsymbol{z}'=\boldsymbol{z}_i}$，$\partial'_j$ は k の 2 番目の変数 $\boldsymbol{z}'$ の j 番目の要素についての微分，$\boldsymbol{k}(\boldsymbol{z}) := (k(\boldsymbol{z}, \boldsymbol{z}_1, \ldots, k(\boldsymbol{z}, \boldsymbol{z}_n))^\top$, $\partial'_j \boldsymbol{k}(\boldsymbol{z}) = (\partial'_j k(\boldsymbol{z}, \boldsymbol{z}_1), \ldots, \partial'_j k(\boldsymbol{z}, \boldsymbol{z}_n))^\top$．そして，係数ベクトル

$\boldsymbol{\alpha}^{(j)} := (\alpha_1^{(j)}, \ldots, \alpha_n^{(j)})^\top$ と $\boldsymbol{\beta}^{(j)} = (\beta_1^{(j)}, \ldots, \beta_n^{(j)})^\top$ を次の最小化問題を解くことで決定できる．

$$\widehat{\boldsymbol{\alpha}}^{(j)}, \widehat{\boldsymbol{\beta}}^{(j)} := \underset{\boldsymbol{\alpha}^{(j)}, \boldsymbol{\beta}^{(j)}}{\operatorname{argmin}} \left[\widehat{J}_j(g^{(j)}) + \frac{\lambda_j}{2} \|g^{(j)}\|_{\mathcal{H}}^2 \right].$$

上式の $\|\cdot\|_{\mathcal{H}}$ は $\mathcal{H}$ におけるノルム，λ_j は非負の正則化パラメータである．また，$\boldsymbol{\alpha}^{(j)}$ と $\boldsymbol{\beta}^{(j)}$ のノルムを正則化項として用いる場合もある．詳細は割愛するが，$\widehat{\boldsymbol{\alpha}}^{(j)}$ と $\widehat{\boldsymbol{\beta}}^{(j)}$ は，解析的に計算可能である．最終的に, 対数密度微分の推定量は

$$\widehat{g}^{(j)}(\boldsymbol{z}) = \sum_{i=1}^{n} \widehat{\alpha}_i^{(j)} k(\boldsymbol{z}, \boldsymbol{z}_i) + \widehat{\beta}_i^{(j)} \partial_j' k(\boldsymbol{z}, \boldsymbol{z}_i)$$

で与えられる．

3　モード探索法 [6]

まず，カーネル関数を次の形式に制限する．

$$k(\boldsymbol{z}, \boldsymbol{z}_i) = \phi\left(\frac{\|\boldsymbol{z} - \boldsymbol{z}_i\|^2}{2\sigma^2} \right).$$

$\sigma > 0$ はカーネル関数の幅を表すパラメータ，ϕ は非負，単調非増加，微分可能な凸関数である．このカーネル関数を用いて，微分の推定量は次のように表現できる．

$$\widehat{g}^{(j)}(\boldsymbol{z}) = \sum_{i=1}^{n} \left[\widehat{\alpha}_i^{(j)} \phi\left(\frac{\|\boldsymbol{z} - \boldsymbol{z}_i\|^2}{2\sigma^2} \right) + \widetilde{\beta}_i^{(j)} \frac{z_i^{(j)} - z^{(j)}}{\sigma^2} \varphi\left(\frac{\|\boldsymbol{z} - \boldsymbol{z}_i\|^2}{2\sigma^2} \right) \right]. \tag{2}$$

式 (2) において，$\widetilde{\beta}_i^{(j)} = -\widehat{\beta}_i^{(j)}$，$\varphi(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\phi(t)$ である．

不動点法に基づき，$\widehat{g}^{(j)}(\boldsymbol{z}) = 0$ とすることで，次の更新式を得る．

$$z^{(j,\tau+1)} = \frac{\sum_{i=1}^{n} \left[\sigma^2 \widehat{\alpha}_i^{(j)} \phi\left(\frac{\|\boldsymbol{z}^\tau - \boldsymbol{z}_i\|^2}{2\sigma^2} \right) + \widetilde{\beta}_i^{(j)} z_i^{(j)} \varphi\left(\frac{\|\boldsymbol{z}^\tau - \boldsymbol{z}_i\|^2}{2\sigma^2} \right) \right]}{\sum_{i=1}^{n} \widetilde{\beta}_i^{(j)} \varphi\left(\frac{\|\boldsymbol{z}^\tau - \boldsymbol{z}_i\|^2}{2\sigma^2} \right)}. \tag{3}$$

式 (3) 中の $\boldsymbol{z}^\tau$ は適当な初期値からの τ 回目の更新結果，$z^{(j,\tau)}$ は $\boldsymbol{z}^\tau$ の j 番目の要素を意味する．ある条件下で更新式 (3) は勾配上昇法と等価であるため，モード探索へ応用可能である．

参考文献

[1] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

[2] Y.-C. Chen. Modal regression using kernel density estimation: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(4):e1431, 2018.

[3] J.E. Chacón. The modal age of statistics. *arXiv preprint arXiv:1807.02789*, 2018.

[4] D. D. Cox. A penalty method for nonparametric estimation of the logarithmic derivative of a density function. *Annals of the Institute of Statistical Mathematics*, 37(1):271–288, 1985.

[5] H. Sasaki, Y. Ono, and M. Sugiyama. Modal regression via direct log-density gradient estimation. In *Proceedings of the 23th International Conference on Neural Information Processing (ICONIP)*, volume 9948, pages 108–116. Springer, 2016.

[6] H. Sasaki, T. Kanamori, A. Hyvärinen, G. Niu, and M. Sugiyama. Mode-seeking clustering and density ridge estimation via direct estimation of density-derivative-ratios. *Journal of machine learing research*, 18(180), 2018.

[7] D.X. Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1-2):456–463, 2008.

Asymptotic properties of kernel PCA with Gaussian kernel for high-dimensional data

Yugo Nakayama [1], Kazuyoshi Yata[2] and Makoto Aoshima[2]
[1] Graduate School of Pure and Applied Sciences, University of Tsukuba
[2]Institute of Mathematics, University of Tsukuba

1 Introduction

Suppose we have independent and d-variate two populations, $\Pi_i,\ i = 1, 2$, having an unknown mean vector $\boldsymbol{\mu}_i$ and unknown covariance matrix $\boldsymbol{\Sigma}_i\ (\geq \boldsymbol{O})$. We assume that $\limsup_{d\to\infty} \|\boldsymbol{\mu}_i\|^2\ /d < \infty$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i)/d \in (0, \infty)$ as $d \to \infty$ for $i = 1, 2$, where $\|\cdot\|$ denotes the Euclidean norm. Here, for a function, $f(\cdot)$, "$f(d) \in (0, \infty)$ as $d \to \infty$" implies $\liminf_{d\to\infty} f(d) > 0$ and $\limsup_{d\to\infty} f(d) < \infty$. Suppose we have a $d \times n$ data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, where $\boldsymbol{x}_j$s are independently taken from Π_1 or Π_2. Let

$$n_i = \#\{j | \boldsymbol{x}_j \in \Pi_i \ \text{ for } j = 1, ..., n\},$$

where $\#A$ denotes the number of elements in a set A. Note that $n = n_1 + n_2$. We assume that n and n_is are independent of d, and $n_i \geq 1$ for $i = 1, 2$. For the sake of simplicity, we assume that $\mathrm{tr}(\boldsymbol{\Sigma}_1) \leq \mathrm{tr}(\boldsymbol{\Sigma}_2)$ and

$$\boldsymbol{x}_j \in \Pi_1 \ \text{ for } j = 1, ..., n_1 \ \text{ and } \ \boldsymbol{x}_j \in \Pi_2 \ \text{ for } j = n_1 + 1, ..., n. \tag{1}$$

In this talk, we studied asymptotic properties of the kernel PCA in the HDLSS context that $d \to \infty$ while n is fixed. Yata and Aoshima (2015) showed that the linear PCA enjoys geometric consistency properties for the PC scores in high-dimensional mixture models. Let $\boldsymbol{K}$ be an $n \times n$ gram matrix with the (j, j') element $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \phi(\boldsymbol{x}_j)^T\phi(\boldsymbol{x}_{j'})$, where $\phi(\cdot)$ is a feature map. Let $\boldsymbol{P}_n = \boldsymbol{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^T$, where $\boldsymbol{I}_n$ denotes the n-square identity matrix and $\mathbf{1}_n = (1, ..., 1)^T$. We define the (centroid) gram matrix by

$$\boldsymbol{K}_0 = \boldsymbol{P}_n\boldsymbol{K}\boldsymbol{P}_n.$$

Note that $\mathrm{rank}(\boldsymbol{K}_0) \leq n - 1$. Let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{n-1}$ be the eigenvalues of $\boldsymbol{K}_0$. Then, we define the eigen-decomposition of $\boldsymbol{K}_0$ by

$$\boldsymbol{K}_0 = \sum_{i=1}^{n-1} \hat{\lambda}_i\hat{\boldsymbol{u}}_i\hat{\boldsymbol{u}}_i^T, \ \text{where } \hat{\boldsymbol{u}}_i = (\hat{u}_{i1}, ..., \hat{u}_{in})^T,\ \|\hat{\boldsymbol{u}}_i\| = 1 \text{ and } \hat{\boldsymbol{u}}_i^T\hat{\boldsymbol{u}}_{i'} = 0 \text{ for all } i \neq i'.$$

The ith (normalized) PC score of $\boldsymbol{x}_j$ is given by

$$\sqrt{n}\hat{u}_{ij} \ \ \text{(hereafter called } s_{ij}\text{)}.$$

We note that $\sum_{j=1}^n s_{ij}^2/n = 1$ for all i. Also, note that $\sum_{j=1}^n s_{ij} = 0$ when $\hat{\lambda}_i > 0$ from the facts that $\mathbf{1}_n^T\hat{\boldsymbol{u}}_i = \sum_{j=1}^n s_{ij}/\sqrt{n}$ and $\mathbf{1}_n^T\boldsymbol{K}_0\mathbf{1}_n = 0$. Since the sign of an eigenvector is arbitrary, we assume that $(\mathbf{1}_{n_1}^T, -\mathbf{1}_{n_2}^T)\hat{\boldsymbol{u}}_1 \geq 0$ without loss of generality.

In this talk, we considered the following two typical kernels:

(I) The linear kernel: $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \boldsymbol{x}_j^T \boldsymbol{x}_{j'}$; and

(II) The Gaussian (radial basis function) kernel: $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \exp(-\|\boldsymbol{x}_j - \boldsymbol{x}_{j'}\|^2/\gamma)$, where $\gamma > 0$.

2 Kernel PCA with the linear kernel (I)

In this section, we consider the KPCA with (I). We assume the following condition:

(A-i) $\mathrm{Var}(\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 | \boldsymbol{x} \in \Pi_i) = O\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}$ as $d \to \infty$ for $i = 1, 2$.

If Π_is are Gaussian, it holds that $\mathrm{Var}(\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 | \boldsymbol{x} \in \Pi_i) = 2\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ for $i = 1, 2$, so that (A-i) naturally holds. Let $\boldsymbol{K}_{0(I)}$ and $s_{ij(I)}$ denote $\boldsymbol{K}_0$ and s_{ij} given by using the kernel functions (I), respectively. Let $\Delta_\mu = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ and $\Delta_\Sigma = |\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)|$. We assume

(A-ii) $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\Delta_\mu^2 = o(1)$ as $d \to \infty$ for $i = 1, 2$.

Then, we have the following result.

Theorem 1. *Assume (A-i) and (A-ii). Assume also*

(A-iii) $\limsup_{d\to\infty} \dfrac{\Delta_\Sigma}{n_1 \Delta_\mu} < 1$ *when* $n_2 \geq 2$.

Then, it holds that as $d \to \infty$

$$s_{1j(I)} = \begin{cases} \sqrt{n_2/n_1} + o_P(1) & \text{when } j = 1, ..., n_1, \\ -\sqrt{n_1/n_2} + o_P(1) & \text{when } j = n_1 + 1, ..., n. \end{cases} \tag{2}$$

Remark 1. Yata and Aoshima (2015) gave the results similar to Theorem 1 under $\Delta_\Sigma/\Delta_\mu = o(1)$ as $d \to \infty$. Note that (A-iii) is milder than $\Delta_\Sigma/\Delta_\mu = o(1)$. When $n_2 = 1$, (2) holds under (A-i) and (A-ii).

Remark 2. If $\mathrm{tr}(\boldsymbol{\Sigma}_1) > \mathrm{tr}(\boldsymbol{\Sigma}_2)$, (2) holds as $d \to \infty$ under (A-i), (A-ii) and the following condition:

$$\limsup_{d\to\infty} \Delta_\Sigma/(n_2 \Delta_\mu) < 1 \quad \text{when } n_1 \geq 2.$$

From Theorem 1, under (A-iii), one can classify $\boldsymbol{x}_j$s into two groups, effectively, by the first PC scores. If Δ_μ/Δ_Σ is small, we do not recommend to use the linear PCA.

In this talk, we investigated asymptotic properties of KPCA with kernel function (II). We showed that the Gaussian kernel gives better performances than the linear kernel in HDLSS settings. Finally, we checked performances of KPCA by numerical simulations and actual data analysis.

References

[1] Yata, K. and Aoshima M. (2015). Principal component analysis based clustering for high-dimension, low-sample-size data. *arXiv:1503.04525.*

超高次元スパース加法モデルにおける変数選択

名古屋工業大学　梅津 佑太

1　はじめに

観測データを $(y_i, \boldsymbol{x}_i) \in \mathbb{R} \times \mathbb{R}^d$ $(i = 1, \dots, n)$ とする非線形回帰モデルを考える. 背後にある真の回帰構造は加法的かつスパースであるとする: つまり, $f^*(\boldsymbol{x}) = \mathbb{E}[y \mid \boldsymbol{x}] = \sum_{j=1}^d f_j^*(x_j)$ かつ $S^* = \{j \mid f_j^* \not\equiv 0\}$ に対して $s = |S^*| \ll d$ である. Fan et al. (2011) は B-スプライン基底を用いて周辺回帰関数を変数ごとに推定し, 推定した関数 $\hat{f}_j$ に対して $S^{\mathrm{NIS}} = \{j \mid \|\hat{f}_j\|_n^2 \geq \lambda_n\}$ によって変数選択することを提案した (NIS). ここで, $\|\cdot\|_n$ は経験 ℓ_2 ノルム, λ は適当な定数である. NIS は sure screening property, つまり, 適当な条件のもと, 1 に収束する確率で $S^* \subset S^{\mathrm{NIS}}$ を満足することが知られている.

本講演では, 線形回帰モデルに対する変数選択基準として提案された sure independence screening (SIS: Fan and Lv, 2008) を自然に非線形回帰モデルへと拡張することで, 再生核ヒルベルト空間における回帰モデルに基づく新たな変数選択基準を導出する. また, この変数選択基準が sure screening property を満たすことを示す.

2　KRIS: Kernel Regression based Independence Screening

説明変数行列の列ベクトルを $\boldsymbol{z}_1, \dots, \boldsymbol{z}_d \in \mathbb{R}^n$ としたとき, SIS は $|\boldsymbol{z}_j^\top \boldsymbol{y}|/n$ をスクリーニングのためのスコアとして利用し, $S^{\mathrm{SIS}} = \{j \mid |\boldsymbol{z}_j^\top \boldsymbol{y}|/n \geq \lambda_n\}$ なる変数を選択する. ここで, $\boldsymbol{y}$ は目的変数, λ_n は調整パラメータである. SIS のスコアを “説明変数と目的変数の相関”, あるいは “周辺回帰の推定量” として解釈することで, SIS を非線形へ拡張できる (例えば, Fan et al., 2011; Li et al., 2012; Balasubramanian et al., 2013; Zambom et al., 2018) が, ここでは, 次のように正則化推定量として解釈する.

補題 1. SIS によって選択される変数を $S^{\mathrm{SIS}} = \{j \mid |\boldsymbol{z}_j^\top \boldsymbol{y}/n| \geq \lambda_n\}$ とする. このとき,

$$\hat{\beta}_j^{\mathrm{ML}} = \operatorname*{arg\,min}_{\beta_j \in \mathbb{R}} \frac{1}{2n} \sum_{i=1}^n (y_i - x_{ij}\beta_j)^2 + \lambda_n |\beta_j|, \qquad j = 1, \dots, d \tag{1}$$

とすれば, $S^{\mathrm{ML}} := \{j \mid \hat{\beta}_j^{\mathrm{ML}} \neq 0\} = S^{\mathrm{SIS}}$ が成り立つ.

補題 1より, (1) の目的関数を非線形化することで, 次の最適化問題を考えることができる.

$$\hat{f}_j = \operatorname*{arg\,min}_{f_j \in \mathcal{H}_j} \frac{1}{2n} \sum_{i=1}^n (y_i - f_j(x_{ij}))^2 + \lambda_n \|f_j\|_{\mathcal{H}_j}, \qquad j = 1, \dots, d. \tag{2}$$

ただし, $\mathcal{H}_j$ はカーネル k_j を持つ RKHS である. 表現定理 (Kimeldorf and Wahba, 1971) より, (2) の最適解は $f_j = \sum_{i=1}^n \alpha_{ij} k_j(\cdot, x_{ij})$ とかけるから, (2) は次の最適化問題を解くことと等価である.

$$\hat{\boldsymbol{\alpha}}_j = \operatorname*{arg\,min}_{\boldsymbol{\alpha}_j = (\alpha_{1j}, \dots, \alpha_{nj})^\top \in \mathbb{R}^n} \frac{1}{2} \|\boldsymbol{y} - K_j \boldsymbol{\alpha}_j\|_n^2 + \lambda_n \sqrt{\boldsymbol{\alpha}_j^\top K_j \boldsymbol{\alpha}_j}, \qquad j = 1, \dots, d. \tag{3}$$

ただし, $K_j = (k_j(x_{pj}, x_{qj}))_{p,q=1,\ldots,n}$ は, k_j に基づくグラム行列である. KKT 条件より $\hat{\boldsymbol{\alpha}}_j \neq \mathbf{0}$ (したがって, $\hat{f}_j \not\equiv 0$) であることと, $\|\boldsymbol{y}\|_{K_j}/n \geq \lambda_n$ は等価である. 以上より, カーネル回帰に基づく変選択基準として

$$S^{\mathrm{KRIS}} = \{j \mid \|\boldsymbol{y}\|_{K_j}/n \geq \lambda_n\}. \tag{4}$$

を考え, これを KRIS と呼ぶことにする.

適当な条件のもと, 次の定理を示すことができる.

定理 1. $\lambda_n = 2\xi n^{(1-\kappa)/(4+6\kappa)}$ このとき, 任意の $\zeta \in (0,1)$ に対して, 適当な条件のもと,

$$\mathrm{P}\left(\max_{1\leq j\leq d} \|\hat{f}_j - f_j^{\mathrm{M}}\|_{\mathcal{H}_j} \gtrsim \lambda_n\right) \lesssim d\Bigg[\exp(-n^{\zeta}) + \exp\left\{-Cn^{(3+2\kappa)/(2+3\kappa)}\right\} + n^{2/(2+3\kappa)}\exp\left\{-\frac{3}{10}n^{2/(2+3\kappa)}\right\} + \exp\left\{-\frac{3}{8}n^{(3+2\kappa)/(2+3\kappa)}\right\}\Bigg]$$

が成り立つ. ただし, κ と ξ はそれぞれ, カーネル関数の固有値および, ノイズに関連する定数であり, C は universal constant である. さらに, $\min_{j\in S^*} \|f_j^{\mathrm{M}}\|_{\mathcal{H}_j} \geq 2\lambda_n$ ならば,

$$\mathrm{P}(S^* \subset S^{\mathrm{KRIS}}) \gtrsim 1 - s\Bigg[\exp(-n^{\zeta}) + \exp\left\{-Cn^{(3+2\kappa)/(2+3\kappa)}\right\} + n^{2/(2+3\kappa)}\exp\left\{-\frac{3}{10}n^{2/(2+3\kappa)}\right\} + \exp\left\{-\frac{3}{8}n^{(3+2\kappa)/(2+3\kappa)}\right\}\Bigg]$$

が成り立つ.

定理 1より, n の指数部分の大小関係に注意すれば, KRIS による変数選択は,

$$\log d = \mathrm{o}(n^{\zeta} + n^{(3+2\kappa)/(2+3\kappa)})$$

なる超高次元でも効率的に変数選択を行えることが期待出来る. また, κ の極限を考えることで, カーネル k_j の固有値が速く減少するほど, より小さなシグナルを検出できることもわかる.

参考文献

Balasubramanian, K., Sriperumbudur, B., and Lebanon, G. (2013) "Ultrahigh dimensional feature screening via RKHS embeddings," in *Artificial Intelligence and Statistics*, pp. 126–134.

Fan, J., Feng, Y., and Song, R. (2011) "Nonparametric independence screening in sparse ultra-high-dimensional additive models," *Journal of the American Statistical Association*, Vol. 106, pp. 544–557.

Fan, J. and Lv, J. (2008) "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B*, Vol. 70, pp. 849–911.

Kimeldorf, G. and Wahba, G. (1971) "Some results on Tchebycheffian spline functions," *Journal of mathematical analysis and applications*, Vol. 33, pp. 82–95.

Li, R., Zhong, W., and Zhu, L. (2012) "Feature screening via distance correlation learning," *Journal of the American Statistical Association*, Vol. 107, pp. 1129–1139.

Zambom, A. Z., Akritas, M. G. et al. (2018) "Hypothesis testing sure independence screening for nonparametric regression," *Electronic Journal of Statistics*, Vol. 12, pp. 767–792.

ガンマ・ダイバージェンス最小化に基づくロバストかつスパースな回帰

東京工業大学　川島　孝行
統計数理研究所　藤澤　洋徳

1　はじめに

KL ダイバージェンスに基づく回帰は，次のように定義できる，

$$D_{KL}(g(y|x), f(y|x;\theta); g(x)) = \int D_{KL}(g(y|x), f(y|x;\theta))g(x)dx,$$

ただし，$g(y|x)$ および $g(x)$ は，データを生成する分布で，$f(y|x;\theta)$ はパラメトリックモデルである．$g(y|x)$ および $g(x)$ を，それぞれ経験密度関数 $\bar{g}(y|x)$ と $\bar{g}(x)$ で置き換えると，上記の KL ダイバージェンスに基づく回帰の経験推定は最尤推定に基づく回帰に一致する．また，適切なパラメトリックモデルを選択することで，線形，ロジスティックおよびポアソン回帰といった主要な回帰モデルの多くを一般化線形回帰の枠組みで捉えることができる (Nelder and Wedderburn, 1972)．しかし，KL ダイバージェンスは，外れ値に弱い．そこで，我々は，ロバストなダイバージェンスとして知られている，ガンマ・ダイバージェンス (Fujisawa and Eguchi, 2008) に基づく回帰を考える．また，これにスパース正則化を組み合わせることで，ロバストかつスパースな回帰を達成する．

2　ガンマ・ダイバージェンスに基づくロバストかつスパースな回帰

まずは，ガンマ・ダイバージェンスを回帰問題の設定に対応させるために，条件付き分布からなる回帰用のガンマ・ダイバージェンスへと拡張を行う．回帰用のガンマ・ダイバージェンスは以下の式で表される．

$$\begin{aligned}
&D_\gamma(g(y|x), f(y|x); g(x)) \\
&= \frac{1}{\gamma}\log\int\left(\int g(y|x)^{1+\gamma}dy\right)^{\frac{1}{1+\gamma}} g(x)dx \\
&- \frac{1}{\gamma}\log\int\int \frac{f(y|x)^\gamma}{\left(\int f(y|x)^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}} g(x,y)dydx.
\end{aligned}$$

これの経験推定を考え，スパース正則化を加えたものを最小化することで，以下の式で表されるロバストかつスパースな回帰の枠組みが達成される．

$$\begin{aligned}
&\arg\min_\theta D_\gamma(\bar{g}(y|x), f(y|x;\theta); \bar{g}(x)) + \lambda P(\theta) \\
&= \arg\min_\theta -\log\left\{\frac{1}{n}\sum_{i=1}^n \frac{f(y_i|x_i;\theta)^\gamma}{\left(\int f(y|x_i;\theta)^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}}\right\} + \lambda P(\theta). \qquad (1)
\end{aligned}$$

3　推定アルゴリズム

上記の最適化は，通常の方法では，推定アルゴリズムの導出自体が困難となる．そこで，補助関数法の一つである，MM アルゴリズム (Hunter and Lange, 2004) の考え方を用いることで，目的関

数の値に関して単調性をもつ推定アルゴリズムの導出に成功した．目的関数 (1) に対応する MM アルゴリズムの考えに基づく補助関数は以下の式で表される．

$$\begin{aligned}&\arg\min_{\theta} h_{MM}\left(\theta|\theta^{(m)}\right)\\&=\arg\min_{\theta}-\frac{1}{\gamma}\sum_{i=1}^{n}\alpha_i\left(\theta^{(m)}\right)\log\left\{\frac{f\left(y_i|x_i;\theta\right)^{\gamma}}{\left(\int f\left(y|x_i;\theta\right)^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}}\right\}+\lambda P(\theta),\end{aligned}$$

ただし，

$$\alpha_i\left(\theta^{(m)}\right)=\left\{\frac{f\left(y_i|x_i;\theta^{(m)}\right)^{\gamma}}{\left(\int f\left(y|x_i;\theta^{(m)}\right)^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}}\right\}\Bigg/\left\{\sum_{l=1}^{n}\frac{f\left(y_l|x_l;\theta^{(m)}\right)^{\gamma}}{\left(\int f\left(y|x_l;\theta^{(m)}\right)^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}}\right\}.$$

具体例として，線形回帰+L1 正則化 (Tibshirani ,1996) を対象とした場合の MM アルゴリズムの考えに基づく推定アルゴリズムは以下の式で表される．

$$\arg\min_{\beta_0,\beta,\sigma^2}\frac{\gamma}{2(1+\gamma)}\log\sigma^2+\frac{\gamma}{2}\sum_{i=1}^{n}\alpha_i^{(m)}\frac{(y_i-\beta_0-x_i^T\beta)^2}{\sigma^2}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

座標降下法に基づく，より具体的な各パラメータの更新式は以下で表される．

$$\begin{aligned}\beta_0^{(m+1)}&=\frac{1}{n}\sum_{i=1}^{n}\alpha_i^{(m)}(y_i-{x_i}^T\beta^{(m)}).\\\beta_j^{(m+1)}&=S\left(\sum_{i=1}^{n}\alpha_i^{(m)}(y_i-r_i^{[-j]})x_{ij},\frac{{\sigma^2}^{(m)}}{\gamma}\lambda\right)\Bigg/\left(\sum_{i=1}^{n}\alpha_i^{(m)}x_{ij}^2\right).\\{\sigma^2}^{(m+1)}&=(1+\gamma)\sum_{i=1}^{n}\alpha_i^{(m)}(y_i-\beta_0^{(m+1)}-x_i^T\beta^{(m+1)})^2.\end{aligned}$$

ただし，$S(a,\lambda)=\mathrm{sign}(a)(|a|-\lambda)_+$，$r_i^{[-j]}=\sum_{k\neq j}x_{ik}\beta_k$．発表当日は，条件付き分布へと拡張したガンマ・ダイバージェンスのロバスト性・数値実験の結果についても報告を行う．

参考文献

Nelder, J. A. and Wedderburn, R. W. M. (1972). *Generalized Linear Models. Journal of the Royal Statistical Society: Series A*, Vol. 135, No. 3, 370-384.

Fujisawa, H. and Eguchi, S. (2008). *Robust parameter estimation with a small bias against heavy contamination. Journal of Multivariate Analysis*, Vol.99, 2053-2081

Hunter, D. R. and Lange, K. (2004) *A tutorial on MM algorithms. The American Statistician*, Vol. 57, No.1, 30-37

Tibshirani, R. (1996). *Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B*, Vol. 58, 267-288.