2022 年度科学研究費シンポジウム

大規模複雑データの理論と方法論〜新たな発展と関連分野への応用〜

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），学術研究助成基金助成金 挑戦的研究 (萌芽) 22K19769「テンソル構造をもつ巨大データの統計的圧縮技術の開発」（研究代表者: 青嶋誠）によるシンポジウムを下記のように催しますので，ご案内申し上げます.

青嶋　誠　(筑波大学)
矢田和善　(筑波大学)
中山優吾　(京都大学)

記

日時：2022 年 11 月 4 日 (金)〜5 日 (土)
場所：つくば国際会議場 中会議室 202 (https://www.epochal.or.jp/ja/)
〒305-0032 茨城県つくば市竹園 2-20-3

プログラム

11 月 4 日 (金)

13:00〜13:10　開会

13:10〜13:50　藤森　洸　（信州大・経法学部）
後藤　佑一　（九州大・大学院数理学研究院）
劉　言　（早稲田大・理工学術院）
谷口　正信　（早稲田大・理工学術院）
高次元・定常時系列に対するスパース主成分分析

13:55〜14:35　赤間　陽二　（東北大・理学研究科）
Unbounded largest eigenvalue of large sample correlation matrices: Asymptotics and applications

14:40〜15:20　永井　勇　（中京大・教養教育研究院）
GMANOVA モデルにおける新たな推定方法と解釈

15:35〜16:15　中川　智之　（東京理科大・理工学部）
ロバストダイバージェンスを用いたベイズ推論について

16:20〜17:00　寺田　吉壱　（大阪大・大学院基礎工学研究科）
山本　倫生　（大阪大・大学院人間科学研究科）
代表点を用いた大規模クラスタリングの近似法とその性質

17:05〜17:45　竹内　努　　（名古屋大・理学研究科）
Recent progress in the application of high-dimensional statistics to astrophysics and cosmology

11 月 5 日 (土)

9:00〜9:30　岡崎　彰良　（電気通信大・大学院情報理工学研究科）
川野　秀一　（九州大・大学院数理学研究院）
凸クラスタリングによるマルチタスク学習

9:35〜10:05　書川　侑子　（電気通信大・大学院情報理工学研究科）
川野　秀一　（九州大・大学院数理学研究院）
馬蹄事前分布に基づく連結 lasso ロジスティック回帰モデリング

10:10〜10:50　仲北　祥悟　（東京大・大学院総合文化研究科）
凸最適化によるエルゴード的拡散過程のオンライン推定

10:55〜11:35　今泉　允聡　（東京大・大学院総合文化研究科）
非スパースな高次元漸近論の理論と応用

11:35〜13:00　　昼食

13:00〜13:40　橋本　真太郎（広島大・大学院先進理工系科学研究科）
羽村　靖之　（京都大・大学院経済学研究科）
鬼塚　貴広　（広島大・大学院先進理工系科学研究科）
菅澤　翔之助（東京大・空間情報科学研究センター）
Sparse Bayesian inference on gamma-distributed observation

13:45〜14:25　矢野　恵佑　（統計数理研究所）
清　智也　　（東京大・大学院情報理工学系研究科）
Minimum information dependence modeling for mixed domain data

14:30〜15:10　鈴木　大慈　（東京大・大学院情報理工学系研究科）
高次元データ学習における特徴学習の優位性

15:25〜16:05　鹿野　豊　　（群馬大・大学院理工学府, Institute for Quantum Studies, Chapman University）
量子乱数プロトコルを用いた量子計算機の安定性評価

16:10〜16:50　福地　純一郎（学習院大・経済学部）
Bootstrap for selecting a subset which contains all populations better than a standard

16:50〜17:00　閉会

高次元・定常時系列に対するスパース主成分分析

藤森洸[1]　後藤佑一[2]　劉言[3,4]　谷口正信[3]

[1] 信州大学　経法学部
[2] 九州大学　数理学研究院
[3] 早稲田大学　基幹理工学部
[4] 早稲田大学　理工学術院総合研究所　重点研究領域「数理科学研究所」

1　概要

主成分分析は多変量データの共分散構造の解析に際し，中心的な役割を果たす重要な手法である．近年は，データの次元 p がサンプル数 n と比較して大きい場合，すなわち高次元の設定における主成分分析の研究が盛んである．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いる．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．本研究は，高次元の設定におけるスパースな第一主成分の推定問題を，定常時系列データに対して適用し，得られる推定量の誤差評価を行うことを目的としている．具体的には，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．ここで，添え字集合 S を $S = \{j : \beta_j^0 \neq 0\}$ で定義する．特に，$\boldsymbol{\beta}^0$ は s_0-*sparse*, つまり，$|S| = s_0$ であると仮定する．さて，$\boldsymbol{\beta}^0$ の推定量として，次を定義する (van de Geer (2016))：

Definition 2.1.

$$\hat{\boldsymbol{\beta}}_n := \arg\min_{\beta \in \mathcal{B}} \left\{ \frac{1}{4} \|\hat{\Sigma}_n - \boldsymbol{\beta}\boldsymbol{\beta}^\top\|_F^2 + \lambda \,\mathrm{pen}(\boldsymbol{\beta}) \right\}, \quad \mathcal{B} := \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_2 \leq \eta\},$$

ただし，$\lambda \geq 0$ は tuning parameter, $\mathrm{pen}(\cdot)$ は罰則関数，$\eta > 0$ は適当な定数である．特に，本研究では次の Lasso 型推定量に焦点を当てる．

$$\hat{\boldsymbol{\beta}}_n^1 := \arg\min_{\beta \in \mathcal{B}} \left\{ \frac{1}{4} \|\hat{\Sigma}_n - \boldsymbol{\beta}\boldsymbol{\beta}^\top\|_F^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \right\}, \tag{2.1}$$

$$\hat{\boldsymbol{\beta}}_n^0 := \arg\min_{\beta \in \mathcal{B}} \left\{ \frac{1}{4} \|\hat{\Sigma}_n - \boldsymbol{\beta}\boldsymbol{\beta}^\top\|_F^2 + \lambda_0 \|\boldsymbol{\beta}\|_0 \right\}. \tag{2.2}$$

Lasso 型推定量を用いて，最大固有値 $\phi_{\max}^2$ 及び対応する規格化された固有ベクトル $\boldsymbol{q}^0$ の推定量を次のように構成することができる：

$$\hat{\phi}_{\max}^2 := \|\hat{\boldsymbol{\beta}}_n^1\|_2^2, \quad \hat{\boldsymbol{q}}_n := \frac{\hat{\boldsymbol{\beta}}_n^1}{\|\hat{\boldsymbol{\beta}}_n^1\|_2}.$$

これらの推定量について，$n \to \infty$ のとき，mixing 条件や，スパイク性に関する条件をはじめとする適当な正則条件の下で，α-mixing Gaussian process に対する推定量について次が成立することが示される：

$$\|\hat{\boldsymbol{\beta}}_n^1 - \boldsymbol{\beta}^0\|_1 = O_p\left(s_0^{3/2} \phi_{\max} \sqrt{\frac{\log p}{n}} \right), \quad \|\hat{\boldsymbol{\beta}}_n^1 - \boldsymbol{\beta}^0\|_2^2 = O_p\left(s_0^3 \phi_{\max}^2 \frac{\log p}{n} \right),$$

$$|\hat{\phi}_{\max} - \phi_{\max}| = O_p\left(s_0^{3/2} \phi_{\max} \sqrt{\frac{\log p}{n}} \right), \quad \|\hat{\boldsymbol{q}}_n - \boldsymbol{q}^0\|_2 = O_p\left(s_0^{3/2} \sqrt{\frac{\log p}{n}} \right).$$

講演内では，Gaussian process よりも重い裾を持ちうる sub-Wiebull process に対する Lasso 型推定量の誤差評価と，Gaussian process に対する l_0 正則化推定量の誤差評価についても述べ，i.i.d. case との比較を行った．さらに，上記の推定量の漸近的な性質を議論する際の設定について詳細な議論を行った．

References

van de Geer, S. A. (2016). *Estimation and Testing under Sparsity*. Springer.

報告書

赤間陽二

2022 年 11 月 9 日

“Unbounded largest eigenvalue of large sample correlation matrices: Asymptotics and applications”(大規模標本相関行列の非有界な最大固有値の漸近論と応用) の題目で講演し, 標本相関行列や標本分散行列の最大固有値の極限での振る舞いについて議論した. ただし, 母集団分布は多次元正規分布だが一定の値 ρ があって異なる変数の組みはどれも相関係数が $\rho \geq 0$ になっていることを要求した. この母集団を equi-correlated normal population (ENP) ということにした. ENP は, *intraclass covariance matrix* [3] を母分散行列とする正規母集団である. 講演を次のとおり行なった:

1. ENP の標本相関行列と標本分散行列の極限固有値分布は, $\rho = 0$ の場合の極限固有値分布が $1 - \rho$ 倍縮小されるという, 講演者たちの結果 [1] を説明した.

2. 講演者たちの結果 [1] の拡張として, 一般の ENP に対して, Guttman-Kaiser criterion が削減する次元の割合は, 次元 p と標本の大きさ n が共に大きいときに, さらに, それらの比 p/n が非常に小さいときは, ρ が 0 ならば 1/2 に収束し, 同相関係数が正ならば 0 に収束し, 母分散は任意であることを注意した. これは, 自由確率論でとりあえず示せることより強い結果であった.

3. 講演者の結果として, 標本相関行列 $\mathbf{R}$ の最大固有値 $\lambda_1(\mathbf{R})$ を $\mathbf{R}$ のサイズ p で割ったものが ρ の強一致推定量であることを紹介した.

4. 経済学のデータとして S&P500 の株のリターンの平均相関係数の時系列を, 相関構造が equicorrelation である GJR GARCH [2] で計算し, その時間平均 $\bar{\rho}$ が $\lambda_1(\mathbf{R})/p$ よりやや小さいが, 両者の間の重回帰係数が高いため, 高速に計算できる $\lambda_1(\mathbf{R})/p$ が有用であることを指摘した.

5. 分子生物学における binary multiple sequence alignment のデータが, 我々の数学的結果 [1] と講演者の結果に比較的合うことを説明した.

6. DNA マイクロアレイデータの相関行列のヒートマップを提示し, それらさまざまな構造を持つにも拘らず, ガットマン・ケイザー基準が示唆する次元と標本の大きさとの積は, ほぼ 1 であることを指摘した。

7. 将来の研究方向として, A. Quadeer (香港科学技術大学), D. Morales-Jimenez (Queen's University Belfast), M. McKay (University of Melbourne) のグループとの議論で彼らが示唆していた相関行列のモデル, すなわち, 生物学や経済学に現れる相関行列で, ブロック構造とバックグラウンドの同相関があり, ただし, ブロックサイズはそんなに大きくならないけれど, ブロックの個数が増えていくモデルを示した.

ENP の標本分散行列の最大固有値の漸近正規性について, 最大固有値のデータ次元に関するオーダーについて注意を喚起され, 青嶋誠氏と矢田和善氏から関連文献 [4, 5, 6] および母集団分布が ENP であるかの検定手順として文献 [3] の付録を紹介された.

最大固有値の次元に関するオーダーについての矢田・青嶋の一連の論文により, 2022 年の秋の数学会 (北海道大学) での講演者の発表に対する, 確率論研究者からの質問に答える見込みを得た. すなわち, $\rho > 0$ である正規母集団の標本分散行列の最大固有値の漸近正規性についての着想を得た.

高次元 PCA を観測天文学のデータに適用する天文学者に, 高次元の探索的因子分析の可能性について質疑して, 面識を得た.

参考文献

[1] Y. Akama and A. Husnaqilati. A dichotomous behavior of Guttman-Kaiser criterion from equi-correlated normal population. *J. Indones. Math. Soc.*, 2022. To appear. arXiv:2210.12580v2.

[2] R. Engle and B. Kelly. Dynamic equicorrelation. *J. Bus. Econ. Stat.*, 30(2):212–228, 2012.

[3] Aki Ishii, Kazuyoshi Yata, and Makoto Aoshima. Hypothesis tests for high-dimensional covariance structures. *Annals of Institute of Statistical Mathematics*, 2021.

[4] Kazuyoshi Yata and Makoto Aoshima. PCA consistency for non-gaussian data in high dimension, low sample size context. *Communications in Statistics - Theory and Methods*, 38(16), 2009.

[5] Kazuyoshi Yata and Makoto Aoshima. Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. *Journal of Multivariate Analysis*, 101(9):2060–2077, 2010.

[6] Kazuyoshi Yata and Makoto Aoshima. Effective PCA consistency for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis*, 2012.

GMANOVA モデルにおける新たな推定方法と解釈 (報告書)

中京大学 教養教育研究院 永井 勇

本講演では, n 個の各個体に対して, 全ての個体で測定時点を揃えて p 回測定して得られる経時測定データの分析を考えた. このようなデータはバランス型経時測定データと呼ばれ, 各個体で測定時点が揃っていないものはアンバランス型経時測定データと呼ばれる. これらのデータの分析の目的は, データの裏に潜む経時変動を上手く捉えることある. 本講演では, バランス型経時測定データの分析について考えた.

バランス型経時測定データの分析の際には, Pothoff and Roy (1964) で提案された次の一般化多変量分散分析 (Generalized Multivariate Analysis of Variance; GMANOVA) モデルがよく使われるため, 本講演でもこのモデルを用いた;

$$\boldsymbol{Y} = \mathbf{1}_n \boldsymbol{\mu}' \boldsymbol{X}' + \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}' + \boldsymbol{\mathcal{E}}, \tag{1}$$

ここで, $\mathbf{1}_n$ は n 次元の全てが 1 からなるベクトル, $\mathbf{0}_r$ は r 次元の全てが 0 からなるベクトル, $\boldsymbol{Y}$ は各行が各個体で測定して得られる経時測定データからなる $n \times p$ 行列, $\boldsymbol{A}$ は各個体の特徴を表す測定時点に無関係な k 個の変数からなる $\mathrm{rank}(\boldsymbol{A}) = k$ の $n \times k$ 行列とし, $\boldsymbol{A}'\mathbf{1}_n = \mathbf{0}_k$ (各説明変数で中心化されている) を満たしているとし, $\boldsymbol{X}$ は後述のように各行が測定時点の関数からなる $p \times q$ 行列であり, これらの $\boldsymbol{Y}$, $\boldsymbol{A}$, $\boldsymbol{X}$ は既知である. また, $\boldsymbol{\mu}$ は q 次元未知ベクトル, $\boldsymbol{\Xi}$ は $k \times q$ 未知行列であり, $\boldsymbol{\mathcal{E}}$ は $E[\boldsymbol{\mathcal{E}}] = \mathbf{0}_n\mathbf{0}_p'$, $\mathrm{Cov}[\mathrm{vec}(\boldsymbol{\mathcal{E}})] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$ の $n \times p$ 誤差行列とし, $\boldsymbol{\Sigma}$ は正則な $p \times p$ 未知行列とする. このモデルにおいて, $E[\boldsymbol{Y}] = \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' + \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}'$ の部分が, 経時測定データの分析で目的としている経時変動に対応していることを報告した. ここで測定時点を $t_1 < t_2 < \cdots < t_p$ として, $\boldsymbol{X}$ の i 列目を $(t_i^0, t_i^1, \ldots, t_i^{q-1})$ とすることは, 経時変動を測定時点の $(q-1)$ 次多項式で推定することに対応することを報告した.

このモデル (1) において, 未知の $\boldsymbol{\mu}$ と $\boldsymbol{\Xi}$ の推定としてよく使われる推定量は, 次のリスクを最小にする $\boldsymbol{\mu}$ と $\boldsymbol{\Xi}$ を求めることで得られることを報告した;

$$R(\boldsymbol{\mu}, \boldsymbol{\Xi}|\boldsymbol{\Sigma}) = \mathrm{tr}\left\{\left(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' - \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}'\right)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' - \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}'\right)'\right\}. \tag{2}$$

実際にこのリスクを最小にする $\hat{\boldsymbol{\mu}}_{\boldsymbol{\Sigma}}$ と $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\Sigma}}$ を求めると, $(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})\hat{\boldsymbol{\mu}}_{\boldsymbol{\Sigma}} = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}'\mathbf{1}_n/n$, $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\Sigma}}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ の解となる. これらの方程式を解き $\hat{\boldsymbol{\mu}}_{\boldsymbol{\Sigma}}$ や $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\Sigma}}$ を得るために, モデル (1) において $\mathrm{rank}(\boldsymbol{X}) = q$ を仮定することが多い. この仮定は**経時変動の推定に用いる関数を制限している**とも考えられることを報告した.

もし $\mathrm{rank}(\boldsymbol{X}) = q$ ならば $(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$ が存在するので, $\hat{\boldsymbol{\mu}}_{\boldsymbol{\Sigma}} = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}'\mathbf{1}_n(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}/n$ であり, $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\Sigma}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$ となる. 実際に経時変動を推定する際は $\boldsymbol{\Sigma}$ が未知のため, その不偏推定量 $\boldsymbol{S} = \boldsymbol{Y}'\{\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n'/n - \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\}\boldsymbol{Y}/(n-k-1)$ が代わりに使われる. このとき, $E[\hat{\boldsymbol{\mu}}_{\boldsymbol{S}}] = \boldsymbol{\mu}$, $E[\hat{\boldsymbol{\Xi}}_{\boldsymbol{S}}] = \boldsymbol{\Xi}$ であることを報告した.

一方で, $\mathrm{rank}(\boldsymbol{X}) < q$ の場合, $(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$ が存在しないため, これらの推定量は得られない. そこで本講演では, 永井 (2021) のアイデアと同様に, $q_1 + \cdots + q_r = q$ となる正の整数を q_i とし, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_r)'$ ($\boldsymbol{\mu}_i$; q_i 次元ベクトル), $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_r)$ ($\boldsymbol{X}_i$; $\mathrm{rank}(\boldsymbol{X}_i) = q_i$ の $p \times q_i$ 行列), $\boldsymbol{\Xi} = (\boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r)$ ($\boldsymbol{\Xi}_i$; $k \times q_i$ 行列) として, モデル (1) を次のように書き換えた;

$$\boldsymbol{Y} = \sum_{i=1}^{r} \mathbf{1}_n \boldsymbol{\mu}_i' \boldsymbol{X}_i' + \sum_{i=1}^{r} \boldsymbol{A}\boldsymbol{\Xi}_i \boldsymbol{X}_i' + \boldsymbol{\mathcal{E}}. \tag{3}$$

ここで $\boldsymbol{X}_i$ の各列が $\mathbf{0}_p$ でない際に, $r = q$ (つまり 1 列ずつ分ける) とすると, $\mathrm{rank}(\boldsymbol{X}_i) = q_i = 1$ となる. つまり, $\boldsymbol{X}_i$ の各列が $\mathbf{0}_p$ でない場合まで考えることができることを報告した.

このように書き換えたモデルにおいても, リスク (2) と同様のリスクを最小にするような $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_r$ や $\boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r$ を求めて並べれば, $\mathrm{rank}(\boldsymbol{X}) < q$ でも $\boldsymbol{\mu}$ および $\boldsymbol{\Xi}$ の推定量が得られるというのが, 本講演のアイデアであった. 実際, リスク (2) の $\boldsymbol{\mu}$ と $\boldsymbol{\Xi}$ と $\boldsymbol{X}$ に分割した形のものを代入すれば, モデル (3) に対応したリスク R' が得られる. R' を最小にする $\hat{\boldsymbol{\mu}}_{i\cdot\boldsymbol{\Sigma}}$ と $\hat{\boldsymbol{\Xi}}_{i\cdot\boldsymbol{\Sigma}}$ $(i = 1, \ldots, r)$ として, それぞれの推定量がどのように求まるかを報告した. ここで, これらを求めて $(\hat{\boldsymbol{\mu}}_{1\cdot\boldsymbol{S}}, \ldots, \hat{\boldsymbol{\mu}}_{r\cdot\boldsymbol{S}})$ や $(\hat{\boldsymbol{\Xi}}_{1\cdot\boldsymbol{S}}, \ldots, \hat{\boldsymbol{\Xi}}_{r\cdot\boldsymbol{S}})$ とすると, $\boldsymbol{\mu}$ と $\boldsymbol{\Xi}$ の推定量が得られ経時変動の推定ができる. しかしながら, そのまま展開するとそれぞれ求まりそうにないため, 本講演では永井 (2021) と同様に, $\boldsymbol{A}'\mathbf{1}_n = \mathbf{0}_k$ に注意しつつ, 展開を工夫した.

その結果, $R' = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_r$に関連する項$) + (\boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r$に関連する項) と分割できることを報告した. したがって, 第一項を最小にする $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_r$ を求め, 第二項を最小にする $\boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r$ を別々に求めればよい. 実際, 各 ℓ でそれぞれ求めると以下のようになることを報告した;

$$\hat{\boldsymbol{\mu}}_{\ell\cdot\boldsymbol{\Sigma}} = (\boldsymbol{X}_\ell'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_\ell)^{-1}\boldsymbol{X}_\ell'\boldsymbol{\Sigma}^{-1}\left(\frac{\boldsymbol{Y}'\mathbf{1}_n}{n} - \sum_{j<\ell}\boldsymbol{X}_j\hat{\boldsymbol{\mu}}_{j\cdot\boldsymbol{\Sigma}}\right),$$

$$\hat{\boldsymbol{\Xi}}_{\ell\cdot\boldsymbol{\Sigma}} = \left((\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y} - \sum_{j<\ell}\hat{\boldsymbol{\Xi}}_{j\cdot\boldsymbol{\Sigma}}\boldsymbol{X}_j'\right)\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_\ell(\boldsymbol{X}_\ell'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_\ell)^{-1},$$

ここで, $\sum_{j<1}\boldsymbol{X}_j\hat{\boldsymbol{\mu}}_{j\cdot\boldsymbol{\Sigma}} = \mathbf{0}_p$, $\sum_{j<1}\hat{\boldsymbol{\Xi}}_{j\cdot\boldsymbol{\Sigma}}\boldsymbol{X}_j' = \mathbf{0}_k\mathbf{0}_p'$ である. これは, $\ell = 1$ の場合の両方が陽に求まり, それを代入することで $\ell = 2$ のときの推定量が得られることを示していることを報告した. これらにおいて $\boldsymbol{\Sigma}$ の部分を $\boldsymbol{S}$ に置き換えることで推定量が構築できる. この推定手法は永井 (2021) と同様に, 何らかの罰則などを用いていないため**最適化のための反復計算が不要**であることを報告した. さらに, $\mathrm{rank}(\boldsymbol{X})$ に関する制約なしで推定が可能となるので, **経時変動の推定に用いる関数への制約がほぼ全てなくなる**ことを表していることを報告した. また, これらの推定量を用いるとモデル (3) へ入れることで, 経時変動が推定できる.

これらの推定量の解釈などについて, 当日の講演で報告した. また, この推定量の問題点や永井 (2022) と同様に別の形の推定量についても当日の講演で報告した.

引用文献:

[1] Pothoff, R. F. & Roy, S. N. (1964) A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313–326.

[2] Nagai, I. (2011) Modified C_p criterion for optimizing ridge and smooth parameters in the MGR Estimator for the nonparametric GMANOVA model. *Open Journal of Statistics*, **1**, 1–14.

[3] 永井 勇 (2021) 高次元小標本における多変量線形回帰モデルでの推定法, 2021 年度統計関連学会連合大会

[4] 永井 勇 (2022) 説明変数がランク落ちしている状況での多変量線形回帰における不偏推定量, 2022 年度統計関連学会連合大会

ロバストダイバージェンスを用いたベイズ推論について

中川 智之 [1]

[1] 東京理科大学理工学部情報科学科

ベイズ統計において, 通常の事後分布は外れ値に対しては頑健ではないため, 外れ値の問題は古くから研究されてきた. 近年では正則変動よりさらに裾の重い対数正則変動 (log-Regularly varying) するモデルに関するベイズ推定も提案され, 頑健性の 1 つである外れ値が含まれたデータを用いた事後分布と外れ値を除いたデータを用いた事後分布が漸近的に一致すること (Posterior robustness) が示されている (Desgagné, 2015, など).

一方で, 裾の重い分布の構成は位置尺度母数モデルよりも一般的なモデルへの拡張は容易ではなく, それぞれのモデルやデータ構造によって構成が必要になる. そこで, 本研究では, 一般のモデルに拡張が容易であるロバストなダイバージェンスを用いたベイズ推論を考える. ロバストなダイバージェンスを用いたベイズ推定は, Hooker and Vidyashankar (2014) や Ghosh and Basu (2016), Nakagawa and Hashimoto (2020) などによって提案されている.

本講演では, 以下のダイバージェンス基準 $D(y,\theta)$ を用いた一般化事後分布について紹介した. $f_\theta(y)$ を仮定した統計モデルとし, $y_1,\ldots,y_n$ を $g(y)$ からの標本とする. このとき, 一般化事後分布は,

$$\pi(\theta \mid Y) := \frac{\pi(\theta)\exp(\sum_{i=1}^n D(y_i;\theta))}{\int_\Theta \pi(\theta)\exp(\sum_{i=1}^n D(y_i;\theta))\mathrm{d}\theta}$$

と定義される. 以下は, それぞれのダイバージェンスに基づく $D(y,\theta)$ である.

- Kullback-Leibler divergence: $D_{KL}(y;\theta) = \log f_\theta(y)$.
- Density power divergence (Basu et al. (1998)):

$$D_\alpha(y;\theta) = \frac{1}{\alpha} f_\theta(y)^\alpha - \frac{1}{\alpha+1}\int f_\theta(x)^{\alpha+1}\mathrm{d}x \quad (\alpha > 0).$$

- γ-divergence (Jones et al., 2001; Fujisawa and Eguchi, 2008):

$$D_\gamma(y;\theta) = \frac{1}{\gamma} f_\theta(y)^\gamma \left\{\int f_\theta(x)^{\gamma+1}\mathrm{d}x\right\}^{-\gamma/(\gamma+1)} - \frac{1}{\gamma} \quad (\gamma > 0).$$

Density power divergence, γ-divergence に関してはそれぞれロバストのダイバージェンスとして知られており, 様々な観点から外れ値に対する頑健性が示されている. 一方で, Posterior robustness との関係は示されていない. 本研究では, Posterior robustness を満たすためのダイバージェンス基準の条件として以下を与えた.

$$\lim_{|z|\to\infty} \frac{\exp\{D(z;\theta)\}}{h(z)} = 1 \ (\forall\theta \in \Theta)$$

この条件は, $\lim_{|z|\to\infty} f_\theta(z) = 0 (\forall \theta \in \Theta)$ を満たすとき, γ-divergence を用いた一般化事後分布は常に成り立つことが簡単にわかる. 一方で, 通常の事後分布や Density power divergence を用いた一般化事後分布に関しては, 十分条件を満たさないことが分かった. 図 1 と図 2 はそれぞれの事後分布のヒストグラムを描いたものである. 図 1 を見てわかる通り, 通常の事後分布は外れ値を含んだ事後分布と外れ値を除いた事後分布は乖離しており, 外れ値に影響を受けることがわかる. 一方で, 図 2 を見ると γ-divergence を用いた一般化事後分布は外れ値を含んだ事後分布と外れ値を除いた事後分布がほぼ同じである.

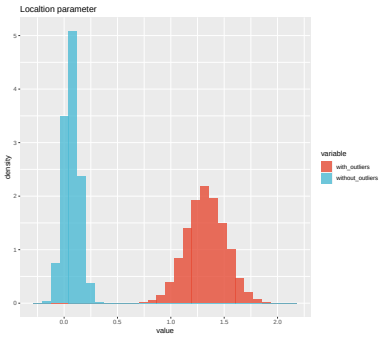

図 1: 通常の事後分布

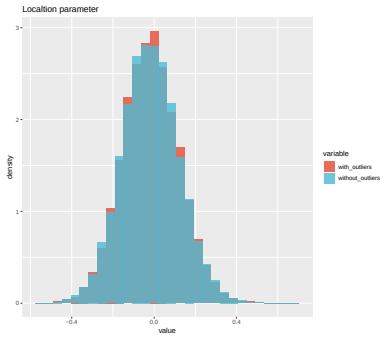

図 2: γ-divergence を用いた一般化事後分布

講演では応用例として, 正規分布の場合と順序付き回帰の場合についての数値実験の詳細を紹介した. 順序付き回帰については, さまざまなリンク関数を柔軟に用いることができ, ロバストなベイズ推定を可能にするためには, ロバストなダイバージェンスを用いる必要があることを報告した.

参考文献

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

Desgagné, A. (2015). Robustness to outliers in location–scale parameter model using log-regularly varying distributions. *The Annals of Statistics*, 43(4):1568–1595.

Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081.

Ghosh, A. and Basu, A. (2016). Robust bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437.

Hooker, G. and Vidyashankar, A. N. (2014). Bayesian model robustness via disparities. *Test*, 23(3):556–584.

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

Nakagawa, T. and Hashimoto, S. (2020). Robust bayesian inference via γ-divergence. *Communications in Statistics-Theory and Methods*, 49(2):343–360.

代表点を用いた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おり, 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, これらの複雑な方法は, 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 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, 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 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．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．また, KASP のアルゴリズムとは本質的に異なるが, 代表点として subsample を用いる spectral clustering の近似法が提案されている (Mohan and Monteleoni, 2017)．しかし, Terada and Yamamoto (2019) の理論から, この方法は subsample に対する normalized cut と等価であり, 汎用的な近似方法ではない．

本発表では, KASP の問題点を明らかにし, その問題点を解決した汎用的な大規模クラスタリングの近似法を提案する．

2. KASP の問題点と提案手法

KASP では, データの代表点として, k-means 法のクラスタ中心を用いる．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．具体的には, 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．

この問題点の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．このことから, 代表点に適切な重みを与えることで, KASP の問題点を解消することができる．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, 新しい代表点の生成方法である Density-Preserving Vector Quantization (DPVQ) を提案する．DPVQ は重み付き k-means 法の一種であり, 容易に代表点を生成できる．また, 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

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**

が示せる. そのため, DPVQ による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, 高次元データに対しては不安定となる.

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も提案する.

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

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. 提案手法を用いた場合の spectral clustering の近似方法, 詳細な理論的性質, DPVQ や $r < 2$ としたベクトル量子化器を用いた提案手法と KASP や subsample を用いた既存の近似手法の数値実験による比較は当日報告する.

参考文献

Mohan, M. and Monteleoni, C. (2017). Beyond the Nystrom Approximation: Speeding up Spectral Clustering using Uniform Sampling and Weighted Kernel k-means. *In Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2494–2500.

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.

Recent progress in the Application of High-Dimensional Statistics to Astrophysics and Cosmology

竹内 努[1,2], COORAY, Suchetha[1], 矢田 和善[3], 青嶋 誠[3], 石井 晶[4], 江頭 健斗[5],
吉川耕司[6], 中西 康一郎[7], 河野 海[1], 河野 孝太郎[8]
1. 名古屋大学理学研究科素粒子宇宙物理学専攻,
2. 統計数理研究所統計的機械学習研究センター,
3. 筑波大学数理物質系,
4. 東京理科大学理工学部情報科学科,
5. 筑波大学 理工情報生命学術院,
6. 筑波大学計算科学研究センター,
7. 国立天文台 ALMA プロジェクト,
8. 東京大学天文学教育研究センター

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

実験室実験の不可能な遠方宇宙を対象とする天文学において，天体の分子の物理情報を得る唯一の手段が分光観測である．しかし詳細な分光観測は一般的に非常に時間がかかり，天体をマッピングして独立な観測点を多数得ることは容易ではない．また，対照が非常に稀な天体であることも多い．どちらのケースも，データの持つ波長方向の次元 d に対してサンプル数 n が非常に小さい($n \ll d$)という状況が生じる．伝統的な天文学では，このような状態は**不良設定問題**とされ，物理的考察のためにはデータ d の情報を大幅に捨て，n より小さくして解析する他ないと考えられていた．しかし当然，このような無駄は避け，情報量をフルに活かせる方法が望ましい．$n \ll d$ となるデータは高次元小標本(HDLSS)データと呼ばれる．天文学以外の分野，たとえばゲノム解析では遺伝子の標本数 $n \sim 100$ に対し塩基配列の次元数が $d \sim 10^5$ といった問題は特殊ではない．HDLSS の問題を解決するためにここ 10 年で発展した統計学の方法が**高次元統計学**であり，現在も続々と新しい知見が得られている(e.g., Aoshima 2018).

本研究では，1) 電波分光マップ観測，および 2) 希少天体の分光観測の 2 つの場合について，高次元統計の方法を適用し，従来の天文学では不可能とされてきた解析を行う．1)では，爆発的星形成銀河 NGC253 の各点をチリの大型ミリ波サブミリ波干渉計 ALMA によって分光観測した探査(マッピング)型について解析する．分光マッピングには一般的に大変な観測的コストがかかり，ALMA の場合現状では最大級でも $n \sim 200$ 程度の観測点が得られる程度であるのに対して波長(振動数)分散方向には $d \sim 2000$ の情報がある．よって，ALMA の分光マッピングは典型的な

HDLSS データである．従来の天文学では，分光データから物理の第一原理から有用であるとわかっているいくつかの輝線のみを抽出し，それを用いて分類を行っていた．しかし，ALMA の観測によって，たとえ隣り合った星間分子雲の領域であっても分子輝線は大幅に異なった様相を示すことが判明し，古典的方法の限界が明らかになっている．巨大すぎる情報量のため，分子雲の進化と星形成の関係が全く見えないのである．本研究では NGC253 の ALMA マップに高次元主成分解析(PCA)を適用した．この結果，高次元 PCA は物理的仮定をおかずに NGC253 中心部のガスの回転運動や星形成領域からのガス流(アウトフロー)を自動的に検出できることを発見した．これは，複雑に見える分光データはわずか数個の固有スペクトルによって表現できることを意味する．Yata & Aoshima (2022)の方法によって，それぞれの固有スペクトルをコントロールするのは HCN (青酸)および HNC 分子であることも見出した．今回の解析は比較的振動数分解能の小さいデータについて行ったが，よりよいデータがすでに得られており，星形成の物理に肉薄できる期待が持てる．

さらに 2)のタイプとして，クェーサーの原子水素吸収線系(HI forest)による「宇宙再電離期」の検証を進めている．宇宙誕生から 1 億年以内の初期，宇宙は原子状態の水素で満たされていた状態から電離水素がほとんどの体積を占める状態への相転移を経験した．これが宇宙再電離と呼ばれている現象である．この研究に有効なのが，宇宙初期に形成されたクェーサーと呼ばれる天体である．クェーサーは巨大なブラックホールに物質が落下することにより，重力エネルギーが解放されて巨大なエネルギーの放射が生じている天体で，あらゆる電磁波の波長できわめて明るい．我々は長波長電波(波長数 cm－数 m)の放射に注目した．クェーサーの手前に水素ガスが存在すると，クェーサーの明るい電波放射に吸収線という「影」として見える．この吸収線の数は膨大であり，天体形成および宇宙論について豊富な情報を持っていることが知られている．しかし，宇宙再電離期にすでに形成され，活動していたクェーサーの数はかなり少ないことが明らかになっており，現在建設中の次世代超大型長波長電波干渉計 Square Kilometre Array (SKA)をもってしても数十個しか検出されないと予想されている．すなわち，クェーサーの HI forest の観測も典型的な HDLSS データである．我々は HI forest から空間 2 点相関関数およびその分散，そして宇宙年齢依存性を高次元統計の方法で解析する方法を開発した．宇宙論シミュレーション Illustris TNG のデータを用いた検証から，この新しい手法によって HI forest を通じた宇宙再電離期の物質の空間分布および進化を検出できることが見出された．今後の発展が期待される．

凸クラスタリングによるマルチタスク学習

電気通信大学 大学院情報理工学研究科　岡崎　彰良
九州大学 大学院数理学研究院　川野　秀一

1　はじめに

マルチタスク学習とは，タスクと呼ばれる複数のデータ集合が得られている際にタスク毎にモデルを設定し，タスク間の関係性を考慮しながらモデルの推定を行う方法論である．タスク毎に設定された各モデルが他のモデルの情報を取り入れることで，モデル全体の推定精度が改善される．ネットワーク lasso 正則化 (Hallac et al., 2015) を用いた手法 (Yamada et al. (2017); He et al. (2019)) では，モデル毎に設定された回帰係数ベクトルの値を縮小して推定することにより，クラスタリングに基づくマルチタスク学習を実行している．しかし，これらの手法の問題点として，ネットワーク lasso 正則化は推定量としてバイアスを持ち，異なるクラスタに属するタスクの情報に影響を受けやすい問題が存在する．そこで本報告では，ネットワーク lasso 正則化に基づくマルチタスク学習の枠組みにおいて，クラスタの重心を表す新たなパラメータを各タスクに対して導入したマルチタスク学習手法を提案する．クラスタの重心に関するパラメータを導入したことにより，縮小によってバイアスを受ける推定量と回帰を行うための推定量が分離され，推定のバイアスが軽減される．提案手法は k 平均法に基づくマルチタスク学習手法の Zhou et al. (2011) とは異なる形での凸緩和手法として見なすことができ，解釈が容易である．

提案手法の有効性をモンテカルロ・シミュレーションおよび実データに対する適用を通して検証する．

2　グループ連結正則化に基づくマルチタスク学習

n_m 個の説明変数に関する p 次元データ $\{\boldsymbol{x}_{mi}; i = 1, \ldots, n_m\}$ および n_m 個の目的変数に関するデータ $\{y_{mi}; i = 1, \ldots, n_m\}$ に対して，これらの組 $\{(y_{mi}; \boldsymbol{x}_{mi}), i = 1, \ldots, n_m\}$ が独立に与えられているとし，目的変数ベクトル $\boldsymbol{y}_m = (y_{m1}, \ldots, y_{mn_m})^\top \in \mathbb{R}^{n_m}$ および計画行列 $X_m = (\boldsymbol{x}_{m1}, \ldots, \boldsymbol{x}_{mn_m})^\top \in \mathbb{R}^{n_m \times p}$ として表す．これらの組を一つのタスクとし，さらに T 個の関連するタスク $\{(\boldsymbol{y}_m, X_m); m = 1, \ldots, T\}$ が与えられているとする．T 個のタスクに対して，目的変数と説明変数の関係を次の T 個の線形回帰モデル

$$\boldsymbol{y}_m = \boldsymbol{X}_m \boldsymbol{w}_m + \boldsymbol{\epsilon}_m, \qquad m = 1, \ldots, T \tag{1}$$

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

$$\min_{\boldsymbol{w}_m \in \mathbb{R}^p,\ m=1,\ldots,T} \left\{ \sum_{m=1}^{T} \frac{1}{2n_m} \|\boldsymbol{y}_m - \boldsymbol{X}_m \boldsymbol{w}_m\|_2^2 + \lambda \sum_{(m,l)\in\mathcal{E}} r_{m,l} \|\boldsymbol{w}_m - \boldsymbol{w}_l\|_q \right\} \tag{2}$$

を考える．ここで，$r_{m,l}$ は非負の値を取る m 番目と l 番目のタスクの関係の大きさを表す重みであり，$\mathcal{E}$ はタスクの組の集合を表す．また，λ は非負の値を取る正則化パラメータである． 第二項目は m 番目と l 番目のタスクに関する回帰係数ベクトル $\boldsymbol{w}_m, \boldsymbol{w}_l$ の併合を誘引する L_q ノルムグループ連結正則化項である．(2) 式は

$q \geq 1$ である場合，凸最適化問題であることが知られており，大域的最適解が得られる．また，$q = 2$ のとき，ネットワーク lasso 正則化 (Hallac et al., 2015) の枠組みとして捉えることができる．

3 提案手法: 凸クラスタリングに基づくマルチタスク学習

最小化問題 (2) においてクラスタリングを実行する第二項目の正則化項は，異なるクラスタに属する回帰係数ベクトル間で縮小を行い，推定精度が低下する問題を含んでいる．この問題を軽減するため，Yamada et al. (2017) や He et al. (2019) では凸クラスタリングと同様に k-近傍法によって重み $r_{m,l}$ の値を与えている．また， Zhou and Zhao (2016) や Shimamura and Kawano (2021) では $r_{m,l}$ をモデルに含まれる潜在変数として考え，その計算アルゴリズムを提案している．しかし，最小化問題 (2) における重み $r_{m,l}$ を潜在変数として推定する方法は目的関数の非凸性を誘引し，大域的最適解を得ることが困難になる問題が生じる．

この回帰係数ベクトル間の縮小に関する問題を克服するために，本報告では以下の最小化問題を考える．

$$\min_{\boldsymbol{w}_m, \boldsymbol{u}_m \in \mathbb{R}^p,\, m=1,\ldots,T} \left\{ \sum_{m=1}^{T} \frac{1}{2n_m} \|\boldsymbol{y}_m - \boldsymbol{X}_m \boldsymbol{w}_m\|_2^2 + \frac{\lambda_1}{2} \sum_{m=1}^{T} \|\boldsymbol{w}_m - \boldsymbol{u}_m\|_2^2 + \lambda_2 \sum_{(m,l)\in\mathcal{E}} r_{m,l} \|\boldsymbol{u}_m - \boldsymbol{u}_l\|_2 \right\}. \tag{3}$$

ここで，$\boldsymbol{u}_m \in \mathbb{R}^p$ は m 番目のタスクが属するクラスタの重心に関するパラメータ，λ_1, λ_2 は正の値を取る正則化パラメータである．本報告では，この手法を MTLCVX (Multi-Task Learning via ConVeX clustering) と呼ぶ.

最小化問題 (3) の第二項目および第三項目は $\boldsymbol{w}_m$ に関して凸クラスタリングを実行する項である．最小化問題 (2) とは異なり，回帰係数 $\boldsymbol{w}_m$ に関しては直接縮小を行わず，$\boldsymbol{u}_m$ に関して縮小を行う．そして，第二項目により，縮小推定された $\boldsymbol{u}_m$ の周りで $\boldsymbol{w}_m$ を推定することでマルチタスク学習が実行される．これにより，$\boldsymbol{w}_m$ に関して生じる縮小に関するバイアスが軽減される．また，最小化問題 (3) は $\boldsymbol{u}_m, \boldsymbol{w}_m$ に関して同時凸関数であるため，その解は大域的最適解となる．モンテカルロ・シミュレーションおよび実データ解析の結果については当日報告する。

参考文献

Hallac, D., Leskovec, J., and Boyd, S. (2015). Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 387–396.

He, X., Alesiani, F., and Shaker, A. (2019). Efficient and scalable multi-task regression on massive number of tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**(01), 3763–3770.

Shimamura, K. and Kawano, S. (2021). A bayesian approach to multi-task learning with network lasso. Preprint, arXiv:1402.6455.

Yamada, M., Koh, T., Iwata, T., Shawe-Taylor, J., and Kaski, S. (2017). Localized Lasso for High-Dimensional Regression. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, **54**, 325–333.

Zhou, J., Chen, J., and Ye, J. (2011). Clustered multi-task learning via alternating structure optimization. *Advances in Neural Information Processing Systems*, **24**, 702–710.

Zhou, Q. and Zhao, Q. (2016). Flexible clustered multi-task learning by learning representative tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**(2), 266–278.

馬蹄事前分布に基づく連結 lasso ロジスティック回帰モデリング

電気通信大学大学院情報理工学研究科　書川　侑子
九州大学大学院数理学研究院　川野　秀一

1　はじめに

二値データの解析手法として，ロジスティック回帰モデルに正則化項を取り入れることでデータのスパースな構造を予測に反映させる手法は近年注目を集めている．ベイズ統計学の観点に立つと，正則化は回帰係数に縮小事前分布を仮定することとして捉え直すことができる．中でもベイジアン連結 lasso (Kyung et al., 2010) は，回帰係数と隣接する回帰係数の差にラプラス事前分布を仮定することで，変数選択および変数の併合を可能にする．しかし，ラプラス事前分布には本来非ゼロと推定されるべき箇所を過剰にゼロに縮小するという問題点がある．

この問題点を解決するため，本報告では，馬蹄事前分布 (Carvalho et al., 2010) を連続する回帰係数の差に仮定したロジスティック回帰モデルを提案した．推定方法としては Polson et al. (2013) による尤度の階層表現を用いることでギブスサンプリングを可能にする．また，数値実験を通して提案手法の有効性を検証した．

2　馬蹄事前分布による連結 lasso を用いたロジスティック回帰モデル

二値を取る目的変数 $y_1, y_2, \cdots, y_n$ と p 次元説明変数ベクトル $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ に対し, 次の二値ロジスティック回帰モデルを考える．

$$\Pr(y_i = 1 \mid \boldsymbol{x}_i) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta})}}, \quad i = 1, \ldots, n. \tag{1}$$

ここで，$\Pr(y_i = 1 \mid \boldsymbol{x}_i)$ は $\boldsymbol{x}_i$ が与えられたもとで $y_i = 1$ となる確率，β_0 は切片項，$\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ は回帰係数ベクトルである．

(1) 式のロジスティック回帰モデルにおいて，回帰係数にラプラス分布を仮定し，回帰係数の差に馬蹄事前分布を仮定することを考える．Pólya-Gamma 分布を用いた尤度の階層表現 (Polson et al., 2013) を用いて，提案モデルは次のように表される．

$$\begin{aligned}
&y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \beta_0, w_i \sim \mathrm{Binom}\left(1, \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta})}}\right), \\
&w_i \sim \mathrm{PG}(1, 0), \\
&\beta_j \mid \tilde{\lambda}_1 \sim \mathrm{Laplace}\left(0, \tilde{\lambda}_1\right), \\
&\beta_j - \beta_{j-1} \mid \lambda_j^2, \tilde{\tau}^2 \sim \mathrm{N}\left(0, \lambda_j^2 \tilde{\tau}^2\right), \\
&\lambda_j \sim \mathrm{C}^+(0, 1), \\
&\tilde{\tau} \sim \mathrm{C}^+(0, 1).
\end{aligned} \tag{2}$$

ここで，パラメータ $a > 0, b$ を持つ Pólya-Gamma 分布は次の確率密度関数で表される．

$$\mathrm{PG}(x \mid a, b) = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - \frac{1}{2})^2 + \frac{c^2}{(4\pi^2)}}$$

g_k はガンマ分布 $\mathrm{Ga}(a, 1)$ に従う独立な確率変数である．また，$\mathrm{Laplace}(x \mid 0, \lambda)$ は，確率密度関数 $\mathrm{Laplace}(x \mid 0, \lambda) = \dfrac{\lambda}{2} \exp\left(-\lambda|x|\right)$ で表されるラプラス分布であり，$\mathrm{C}^+(x \mid 0, 1)$ は確率密度関数 $\mathrm{C}^+(x \mid 0, 1) = \dfrac{2}{\pi\left(x^2 + 1\right)}$ で表される半コーシー分布である．馬蹄事前分布を連続する回帰係数の差に仮定することで，回帰係数の差のうち非 0 に推定されるべき箇所が過度に縮小されることを抑えた推定が行える．

正規尺度混合 (Andrews and Mallows, 1974) と半コーシー分布の階層表現 (Wand et al., 2011) を用いると (2) 式のモデルは次のように表せる．

$$\begin{aligned}
&y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \beta_0, w_i \sim \mathrm{Binom}\left(1, \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta})}}\right), \\
&w_i \sim \mathrm{PG}(1, 0), \\
&\beta_j \mid \tau_j^2 \sim \mathrm{N}(0, \tau_j^2), \\
&\tau_j^2 \sim \mathrm{EXP}\left(\frac{\tilde{\lambda}_1^2}{2}\right), \\
&\lambda_j^2 \mid \nu_j \sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{\nu_j}\right), \\
&\tau^2 \mid \xi \sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{\xi}\right), \\
&\nu_1, \ldots, \nu_p, \xi \sim \mathrm{IG}\left(\frac{1}{2}, 1\right).
\end{aligned} \tag{3}$$

(3) 式の階層表現により，回帰係数の推定方法としてギブス・サンプリングが構成できる．

参考文献

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, **36**(1), 99–102.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**(2), 465–480.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, **5**(2), 369–411.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, **108**(504), 1339–1349.

Wand, M. P., Ormerod, J. T., Padoan, S. A., and Frühwirth, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis*, **6**(4), 847–900.

凸最適化によるエルゴード的拡散過程のオンライン推定

ONLINE ESTIMATION OF ERGODIC DIFFUSION PROCESSES WITH CONVEX OPTIMIZATION

仲北祥悟（東京大学大学院総合文化研究科）

Let us consider the parametric estimation of the following d-dimensional stochastic differential equation (SDE):

$$\mathrm{d}X_t^{a,b} = b\left(X_t^{a,b}\right)\mathrm{d}t + a\left(X_t^{a,b}\right)\mathrm{d}w_t,\ X_0 = x \in \mathbf{R}^d,\ t \geq 0.$$

SDEs describe dynamics with randomness and allow for flexible model structures under mild conditions. Therefore, they are used to model phenomena in broad disciplines such as finance, biology, epidemiology, physics, meteorology, and machine learning. In this study, we propose an online parametric estimation method of b based on discrete observations $\{X_{ih_n}^{a,b}\}_{i=0,\ldots,n}$ with $h_n > 0$.

Online estimation, where the estimator is updated as data are acquired, is a typical and significant concern in time series data analysis because it is quite useful for real-time decision making. However, most studies on the online parametric estimation of SDEs depend on the setting of continuous observations $\{X_t^{a,b}\}_{t\geq 0}$ [3, 1, 2], which is restrictive in real data analysis. Hence, we aim to propose online estimation methods for SDEs with discrete observations.

We provide uniform risk bounds for the parametric estimation of both diffusion and drift coefficients of SDEs with discrete observations and model misspecification via online gradient descent with convex loss functions and their convex approximations. Those bounds give theoretical convergence guarantees of the proposed online estimation method for SDEs with discrete observations, which are the main contribution of our study. To derive the bounds, we combine the three theoretical discussions: (i) model-wise non-asymptotic risk bound for the stochastic mirror descent (SMD) with dependent and biased subgradients; (ii) simultaneous ergodicity and uniform moment bounds for a class of SDEs; and (iii) the proposal of loss functions for the online parametric estimation.

Selecting drift estimation as an example in this section, we set the convex and compact parameter space $\Theta \subset \mathbf{R}^p$ and the triple of measurable functions (b^{m}, M, J) such that $b^{\mathrm{m}}(x,\theta)$ is the possibly misspecified parametric model, $M(x)$ is a positive semi-definite weight function, and $J(\theta)$ is the regularization term. We set the function

$$\phi(x,y,\theta) := \frac{1}{2}M(x)\left[\left(y - b^{\mathrm{m}}(x,\theta)\right)^{\otimes 2}\right] + J(\theta).$$

Assume that $\phi(x,y,\theta)$ is convex in θ for all $x, y \in \mathbf{R}^d$ and has measurable elements in the subdifferential for all x, y, and θ. $\{\theta_i; i = 1,\ldots,n+1\}$ defined by the following online gradient descent algorithm

$$\theta_{i+1} := \mathrm{Proj}_\Theta\left(\theta_i - \frac{h_n}{\sqrt{i}}\partial_\theta\phi\left(X_{(i-1)h_n}^{a,b}, \frac{1}{h_n}\Delta_i X^{a,b}, \theta_i\right)\right),$$

with an arbitrary initial value $\theta_1 \in \Theta$ and a sequence of discrete observations $\{X_{ih_n}^{a,b}; i = 0,\ldots,n\}$, is then well-defined as a sequence of random variables by choosing measurable subgradients, where $\Delta_i X^{a,b} = X_{ih_n}^{a,b} - X_{(i-1)h_n}^{a,b}$ and $h_n > 0$ is the discretization step. Note that the learning rate chosen here does not lead to the best convergence but is simple and approximately the best in our study. Our contributions (i) and (iii) provide the following risk bound for the

This work was partially supported by JSPS KAKENHI Grant Number JP21K20318 and JST CREST Grant Number JPMJCR21D2 and JPMJCR2115.

estimator $\bar{\theta}_n := \frac{1}{n}\sum_{i=1}^n \theta_i$ with a fixed (a,b): for some $c>0$,

$$\sup_{\theta\in\Theta}\left(\mathbf{E}_x^{a,b}\left[f^{a,b}\left(\bar{\theta}_n\right)\right]-f^{a,b}\left(\theta\right)\right)\leq c\left(\frac{\log nh_n^2}{\sqrt{nh_n^2}}+h_n^{\beta/2}\right),$$

where $\beta\in[0,1]$ is a parameter controlling the smoothness of b, $\mathbf{E}_x^{a,b}$ is the expectation over $\{X_t^{a,b}\}_{t\geq 0}$ with $X_0^{a,b}=x$, $f^{a,b}\left(\theta\right)=\int M\left(\xi\right)\left[\left(b^{\mathrm{m}}\left(\xi,\theta\right)-b\left(\xi\right)\right)^{\otimes 2}\right]\Pi^{a,b}\left(\mathrm{d}\xi\right)+J\left(\theta\right)$ is the loss function, and $\Pi^{a,b}$ is the invariant probability measure of $X_t^{a,b}$. The contribution (ii) yields the existence of c such that the inequality holds uniformly in $S:=\{(a,b)\}$, a class of coefficients of SDEs satisfying the same regularity conditions; hence, the risk bound is uniform in S. Note that $\bar{\theta}_n$ estimates the best $\theta\in\Theta$ (or the quasi-optimal parameter; see [4]) with $b^{\mathrm{m}}\left(,\theta\right)$ closest to the true b in the $L^2\left(\Pi^{a,b}\right)$-distance. Moreover, if the model b^{m} correctly specifies b, that is, for all $(a,b)\in S$ there exists θ such that $b=b^{\mathrm{m}}\left(\cdot,\theta\right)$, then we can obtain the following bound:

$$\sup_{(a,b)\in S}\mathbf{E}_x^{a,b}\left[f^{a,b}\left(\bar{\theta}_n\right)\right]\leq c\left(\frac{\log nh_n^2}{\sqrt{nh_n^2}}+h_n^{\beta/2}\right).$$

One simple but significant outcome of the above discussion is a non-asymptotic risk guarantee of the following online gradient descent for linear models such that an arbitrary initial value $\theta_1\in\Theta$,

$$\theta_{i+1}:=\mathrm{Proj}_\Theta\left(\theta_i+\frac{1}{\sqrt{i}}\left(\partial_\theta b^{\mathrm{m}}\left(X_{(i-1)h_n}^{a,b},\theta_i\right)\right)\left(\Delta_i X^{a,b}-h_n b^{\mathrm{m}}\left(X_{(i-1)h_n}^{a,b},\theta_i\right)\right)\right),$$

where $b^{\mathrm{m}}\left(x,\theta\right)$ is the possibly misspecified parametric model whose components are linear in $\theta\in\Theta$. Note that it corresponds to the case $M\left(x\right)=I_d$, $J\left(\theta\right)=0$. As evident, the uniform risk bound for the estimator $\bar{\theta}_n:=\frac{1}{n}\sum_{i=1}^n\theta_i$ over a certain family S of the coefficients a,b holds: for some $c>0$,

$$\sup_{(a,b)\in S}\sup_{\theta\in\Theta}\left(\mathbf{E}_x^{a,b}\left[\int\left\|b^{\mathrm{m}}\left(\xi,\bar{\theta}_n\right)-b\left(\xi\right)\right\|_2^2\Pi^{a,b}\left(\mathrm{d}\xi\right)\right]-\int\left\|b^{\mathrm{m}}\left(\xi,\theta\right)-b\left(\xi\right)\right\|_2^2\Pi^{a,b}\left(\mathrm{d}\xi\right)\right)$$
$$\leq c\left(\frac{\log nh_n^2}{\sqrt{nh_n^2}}+h_n^{\beta/2}\right).$$

If we assume that b^{m} correctly specifies b, then

$$\sup_{(a,b)\in S}\mathbf{E}_x^{a,b}\left[\int\left\|b^{\mathrm{m}}\left(\xi,\bar{\theta}_n\right)-b\left(\xi\right)\right\|_2^2\Pi^{a,b}\left(\mathrm{d}\xi\right)\right]\leq c\left(\frac{\log nh_n^2}{\sqrt{nh_n^2}}+h_n^{\beta/2}\right).$$

References

[1] Bhudisaksang, T. and Cartea, A. (2021). Online drift estimation for jump-diffusion processes. *Bernoulli*, 27(4):2494–2518.

[2] Sharrock, L. and Kantas, N. (2022). Joint online parameter estimation and optimal sensor placement for the partially observed stochastic advection-diffusion equation. *SIAM/ASA Journal on Uncertainty Quantification*, 10(110):55–95.

[3] Surace, S. C. and Pfister, J.-P. (2019). Online maximum-likelihood estimation of the parameters of partially observed diffusion processes. *IEEE Transactions on Automatic Control*, 64(7):2814–2829.

[4] Uchida, M. and Yoshida, N. (2011). Estimation for misspecified ergodic diffusion processes from discrete observations. *ESAIM: Probability and Statistics*, 15:270–290.

NON-SPARSE HIGH-DIMENSIONAL ASYMPTOTICS: THEORY AND PRACTICE

MASAAKI IMAIZUMI[1]

[1]*The University of Tokyo*

1. Introduction

This talk will deal with the recently developed high-dimensional asymptotics. Traditionally, high-dimensional statistics based on sparsity have been mainly investigated. In contrast to the literature, in recent years, several high-dimensional analysis without sparsity has been attracting attention, motivated by the recent success of non-sparse large-scale statistical analysis. For example, there is a high-dimensional analysis using the spectrum of data matrices and an analysis of high-dimensional models using the Gaussian comparison theorem. In this presentation, I will overview those theories and introduce some research results applying them.

This presentation consists of several studies. First, we present an extension of the concentration inequality for empirical averages of high-dimensional random matrices (Koltchinskii and Lounici, 2017), and related over-parameterized linear regression studies (Bartlett *et al.*, 2020). We extend these results to distributions with thick hems. We then present methods for dealing with high-dimensional models under Gaussianity. These are done using the Gaussian comparison theorem (Thrampoulidis *et al.*, 2015, 2018) and the proud message propagation method (Bayati and Montanari, 2011; Sur and Candès, 2019). We develop M-estimators under general regularization as well as generalizations of these estimators in the case of unknown link functions.

Date: December 8, 2022.

References

Bartlett, P. L., Long, P. M., Lugosi, G. and Tsigler, A. (2020) Benign overfitting in linear regression, *Proceedings of the National Academy of Sciences*, **117**, 30063–30070.

Bayati, M. and Montanari, A. (2011) The lasso risk for gaussian matrices, *IEEE Transactions on Information Theory*, **58**, 1997–2017.

Koltchinskii, V. and Lounici, K. (2017) Concentration inequalities and moment bounds for sample covariance operators, *Bernoulli*, **23**, 110–133.

Sur, P. and Candès, E. J. (2019) A modern maximum-likelihood theory for high-dimensional logistic regression, *Proceedings of the National Academy of Sciences*, **116**, 14516–14525.

Thrampoulidis, C., Abbasi, E. and Hassibi, B. (2018) Precise error analysis of regularized m-estimators in high dimensions, *IEEE Transactions on Information Theory*, **64**, 5592–5628.

Thrampoulidis, C., Oymak, S. and Hassibi, B. (2015) Regularized linear regression: A precise analysis of the estimation error, in *Conference on Learning Theory*, PMLR, pp. 1683–1709.

Sparse Bayesian inference on gamma-distributed observations

Yasuyuki Hamura[1], Takahiro Onizuka[2], Shintaro Hashimoto[2], Shonosuke Sugasawa[3]
[1]Graduate School of Economics, Kyoto University
[2]Department of Mathematics, Hiroshima University
[3]Center for Spatial Information Science, The University of Tokyo

In various statistical applications, we often face a sequence of positive-valued observations such as machine failure time, store waiting time, survival time under a certain disease, an income of a certain group, and so on. A common feature of the data is "sparsity" in the sense that most of the underlying means of observations are concentrated around a certain value (grand mean) while a small part of the means is significantly away from the grant mean. To reflect the sparsity structure, a useful Bayesian technique is an idea of "global-local shrinkage" (e.g. Carvalho et al., 2010; Polson and Scott, 2010; Datta and Dunson, 2016) that provides adaptive and flexible shrinkage estimation of underlying means; when the observations are around the grand mean, the posterior mean strongly shrinks the observation toward the grand mean, but the observations that are away from the grand mean remain unshrunk.

In this presentation, we consider the following gamma inverse-gamma hierarchical model:

$$y_i \mid \lambda_i \sim \mathrm{Ga}\left(\delta_i, \frac{\delta_i}{\lambda_i}\right), \quad \lambda_i \mid u_i \sim \mathrm{IG}(1+\tau u_i, \beta\tau u_i), \quad u_i \sim \pi(\cdot), \quad i = 1, \ldots . n,$$

where $\beta > 0$ and $\tau > 0$ are unknown hyper-parameters for which we assign conjugate gamma priors, and λ_i is a parameter of interest. The prior mean of λ_i is $\mathbb{E}[\lambda_i] = \beta$, so that β is interpreted as a grand mean of underlying heterogeneous means. A fixed constant δ_i is selected in given context. As considered in Liu and Stephens (2016), if y_i and λ_i are sampling and true variances, respectively, the choice is $\delta_i = n_i/2$, where n_i is a sample size used to compute y_i. Moreover, if y_i is a sample mean based on n_i samples generated from an exponential distribution $\mathrm{Exp}(1/\lambda_i)$, it holds that $\delta_i = n_i$, and it reduces the framework of a sequence of exponential data when $n_i = 1$, considered in Donoho and Jin (2006). Unlike existing shrinkage priors, our new prior is a *shape-scale mixture of inverse-gamma* distributions, which has a desirable interpretation of the form of posterior mean and admits flexible shrinkage. In fact, the posterior mean of λ_i is given by

$$\mathbb{E}[\lambda_i \mid y_i] = \mathbb{E}\left[\frac{\delta_i y_i + \beta\tau u_i}{\delta_i + \tau u_i} \;\middle|\; y_i\right] = y_i - \mathbb{E}[\kappa_i \mid y_i](y_i - \beta),$$

where $\kappa_i = \tau u_i/(\delta_i + \tau u_i) \in (0,1)$ is known as *shrinkage factor* that determines the amount of shrinkage of y_i toward the grand mean β. As desirable properties of κ_i, $\mathbb{E}[\kappa_i \mid y_i]$ should be close to 1 when y_i is close to the grand mean, leading to strong shrinkage toward β, while $\mathbb{E}[\kappa_i \mid y_i]$ should be sufficiently small for y_i having large $y_i - \beta$ to prevent bias caused by over-shrinkage. We also note that the global parameter τ determines the overall shrinkage effect, whereas the local parameter u_i allows κ_i to vary over different observations. As priors of u_i that satisfy such desirable properties, we propose the following scaled beta (SB) and inverse rescaled beta (IRB) priors.

$$\pi_{\mathrm{SB}}(u_i) = \frac{1}{B(a,b)} \frac{{u_i}^{a-1}}{(1+u_i)^{a+b}},$$
$$\pi_{\mathrm{IRB}}(u_i) = \frac{1}{B(b,a)} \frac{1}{u_i(1+u_i)} \frac{\{\log(1+1/u_i)\}^{b-1}}{\{1+\log(1+1/u_i)\}^{b+a}},$$

where $a, b > 0$ are hyperparameters and $B(a,b)$ is the beta function. We discuss the roles of the hyperparameters, a and b, of the proposed priors, and then we propose particular choices of the hyperparameters. Under the proposed priors, we can construct an efficient sampling algorithm for posterior inference via Markov chain Monte Carlo (MCMC) method. Furthermore, we show that the proposed priors have two desirable theoretical properties;

(i) robust shrinkage rules for large observations, and

(ii) Kullback-Leibler super-efficiency under sparsity.

The performance of the proposed method is demonstrated through numerical studies, and we apply the method to two real datasets related to the average length of hospital stay for COVID-19 in South Korea and variance estimation of gene expression data.

References

[1] Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97, 465–480.

[2] Datta, J. and Dunson, D. (2016). Bayesian inference on quasi-sparse count data. *Biometrika*, 103(4), 971–983.

[3] Donoho, D. and Jin, J. (2006). Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *The Annals of Statistics*, 34, 2980–3018.

[4] Lu, M. and Stephens, M. (2016). Variance adaptive shrinkage (vash): flexible empirical Bayes estimation of variances. *Bioinformatics*, 32, 3428–3434.

[5] Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian Statistics* 9, 501–538.

Minimum information dependence modeling for mixed domain data

Keisuke Yano (the Institute of Statistical Mathematics),
Tomonari Sei (the University of Tokyo)

Abstract

In this talk, we introduce a joint statistical model for mixed-domain data that is proposed by [1]. The proposed model contains multivariate Gaussian and log-linear models. We show the existence and uniqueness of the proposed model under fairly weak conditions. To estimate the dependence parameter in our model, we present a conditional inference together with a sampling procedure and show it provides a consistent estimator of the dependence parameter.

1 Minimum information dependence model

Let $(\mathcal{X}_i, \mathcal{F}(\mathcal{X}_i), \mathrm{d}x_i)$ for $i = 1, \dots, d$ be measure spaces and denote their product space by $\mathcal{X} = \prod_{i=1}^d \mathcal{X}_i$ and $\mathrm{d}x = \prod_{i=1}^d \mathrm{d}x_i$. For index i, use the notation $-i$ to indicate the removal of the i-th coordinate, e.g., $x_{-i} = (x_j)_{j \neq i}$, $\mathcal{X}_{-i} = \prod_{j \neq i} \mathcal{X}_j$, and $\mathrm{d}x_{-i} = \prod_{j \neq i} \mathrm{d}x_j$.

Let $r_1(x_1; \nu), \dots, r_d(x_d; \nu)$ be statistical models of marginal densities on $\mathcal{X}_1, \dots, \mathcal{X}_d$, respectively, where ν denotes parameters characterizing the marginal densities. We can assign, if necessary, independent parameters to each r_i as $r_i(x_i; \nu_i)$ by setting $\nu = (\nu_1, \dots, \nu_d)$.

We consider a class of probability density functions

$$p(x; \theta, \nu) = \exp\left(\theta^\top h(x) - \sum_{i=1}^d a_i(x_i; \theta, \nu) - \psi(\theta, \nu)\right) \prod_{i=1}^d r_i(x_i; \nu), \tag{1}$$

where $\theta \in \mathbb{R}^K$ is a K-dimensional parameter representing the dependence, and $h : \mathcal{X} \to \mathbb{R}^K$ is a given function. The functions $a_i(x_i; \theta, \nu)$ and $\psi(\theta, \nu)$ are simultaneously determined by constraints

$$\int p(x; \theta, \nu) \mathrm{d}x_{-i} = r_i(x_i; \nu), \quad i = 1, \dots, d, \text{ and} \tag{2}$$

$$\int \sum_{i=1}^d a_i(x_i; \theta, \nu) p(x; \theta, \nu) \mathrm{d}x = 0. \tag{3}$$

Note that the density (1) is reduced to the independent model $\prod_{i=1}^d r_i(x_i; \nu)$ if $\theta = 0$.

Definition 1. A statistical model (1) together with the constraints (2) and (3) is called a *minimum information dependence model*. The parameter θ is called the *canonical parameter*, ν is the *marginal parameter*, $h(x)$ comprises the *canonical statistics*, $a_i(x_i; \theta, \nu)$s are the *normalizing functions* and $\psi(\theta, \nu)$ is the *potential function*.

Figure 1 displays a two-dimensional histogram of samples from the minimum information dependence model for mixed variables (discrete and $[0, 1]$) with negative correlation, which shows that the minimum information dependence model easily expresses the dependence between mixed-domain variables.

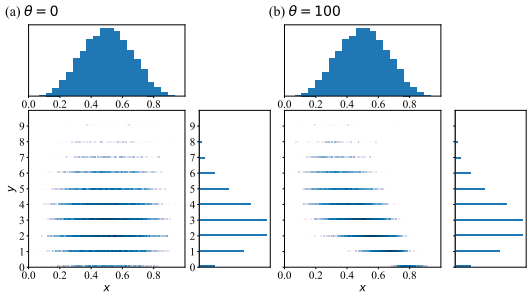

Figure 1: Two-dimensional histograms of 10000 samples from the minimum information dependence model with the Beta Beta$(10, 10)$ and Poisson Po(3) marginals. The canonical statistic $h(x, y)$ is given by $h(x, y) = x/(y + 1)$. The joint histogram and marginal histograms are plotted. (a) Joint histogram with $\theta = 0$. (b) Joint histogram with $\theta = 100$.

Let $p_0(x) := \prod_{i=1}^d r_i(x_i; \nu)$. We say that a function $H \in L_1(p_0(x)\mathrm{d}x)$ is *feasible* if there exist measurable functions $\{a_i(x_i) : i = 1, \dots, d\}$ and a real number $\psi \in \mathbb{R}$ such that the function $p(x) = e^{H(x) - \sum_{i=1}^d a_i(x_i) - \psi} p_0(x)$ satisfies

$$\int p(x)\mathrm{d}x_{-i} = r_i(x_i; \nu) \qquad \text{for each } i = 1, \dots, d \quad \text{and}$$
$$\int \sum_{i=1}^d a_i(x_i) p(x)\mathrm{d}x = 0.$$

Theorem 1 (Theorem 1 of [1])**.** A function $H \in L_1(p_0(x)\mathrm{d}x)$ is feasible if there exist $\{b_i \in L_1(r_i(x_i)\mathrm{d}x_i) : i = 1, \dots, d\}$ such that

$$\int e^{H(x) - \sum_{i=1}^d b_i(x_i)} p_0(x)\mathrm{d}x < \infty. \tag{4}$$

Furthermore, if H is feasible, then $\sum_{i=1}^d a_i(x_i)$ and ψ are unique.

References

[1] T. Sei and K. Yano. Minimum information dependence modeling, 2022. arXiv:2206.06792.

高次元データ学習における特徴学習の優位性

Taiji Suzuki[1,2]
Joint work with Sho Okumoto[1], Jimmy Ba[3], Murat A. Erdogdu[3], Zhichao Wang[4], Denny Wu[3], Greg Yang[5]
[1]Graduate School of Information Science and Technology, the University of Tokyo,
[2]RIKEN Center for Advanced Intelligence Project
[3]University of Toronto and Vector Institute, [4]University of California, San Diego,
[5]Microsoft Research AI

In my presentation, I have presented two topics: (1) Learnability of convolutional neural networks for infinite dimensional input, and (2) High-dimensional asymptotics of feature learning. Each topic is based on the papers Okumoto & Suzuki (2022) and Ba et al. (2022) respectively.

1 Learnability of convolutional neural networks for infinite dimensional input

In the first part, I have presented a recent result about the learning ability of convolutional neural networks for infinite dimensional input Okumoto & Suzuki (2022). Let λ be the uniform probability measure on $([0,1], \mathcal{B}([0,1]))$ where $\mathcal{B}([0,1])$ is the Borel σ-field on $[0,1]$, and let λ^∞ be the product measure of λ on $([0,1]^\infty, \mathcal{B}([0,1]^\infty))$ where $\mathcal{B}([0,1]^\infty)$ is the product σ-algebra generated by the cylindric sets $\cap_{j\le d}\{x \in [0,1]^\infty : x_j \in B_j\}$ for $d = 1, 2, \dots$ and $B_j \in \mathcal{B}([0,1])$. Let P_X be a probability measure defined on the measurable space $([0,1]^\infty, \mathcal{B}([0,1]^\infty))$ that is absolutely continuous to λ^∞ and its Radon-Nikodym derivative satisfies $\|\frac{\mathrm{d}P_X}{\mathrm{d}\lambda^\infty}\|_{L^\infty([0,1]^\infty)} < \infty$. Then, suppose that there exists a true function $f^\circ : [0,1]^\infty \to \mathbb{R}$, and consider the following nonparametric regression problem with an infinite dimensional input:

$$Y = f^\circ(X) + \xi,$$

where X is a random variable taking its value on $[0,1]^\infty$ and obeys the distribution P_X introduced above, and ξ is a observation noise generated from $N(0, \sigma^2)$ (a normal distribution with mean 0 and variance $\sigma^2 > 0$). Let P be the joint distribution of X and Y obeying the regression model. Then, we analyzed the estimation accuracy of deep convolutional network to estimate f° from n input-output observations $(X_i, y_i)_{i=1}^n$ and investigated how the smoothness of the target function f° affects the convergence rate.

Among a wide range of success of deep learning, convolutional neural networks have been extensively utilized in several tasks such as speech recognition, image processing, and natural language processing, which require inputs with large dimensions. Several studies have investigated function estimation capability of deep learning, but most of them have assumed that the dimensionality of the input is much smaller than the sample size. However, for typical data in applications such as those handled by the convolutional neural networks described above, the dimensionality of inputs is relatively high or even infinite. In our analysis, we investigated the approximation and estimation errors of the (dilated) convolutional neural networks when the input is infinite dimensional. Although the approximation and estimation errors of neural networks are affected by the curse of dimensionality in the existing analyses for typical function spaces such as the Hölder and Besov spaces, we show that, by considering anisotropic smoothness, they can alleviate exponential dependency on the dimensionality but they only depend on the smoothness of the target functions. Our theoretical analysis supports the great practical success of convolutional networks. Furthermore, we show that the dilated convolution is advantageous when the smoothness of the target function has a sparse structure.

2 High-dimensional asymptotics of feature learningt

In the second part, I have presented the asymptotic analysis of predictive accuracy of two layer neural networks with feature learning developed in Ba et al. (2022).

We consider the training of a fully-connected two-layer neural network (NN) with N neurons,

$$f_{\mathrm{NN}}(\boldsymbol{x}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_i \rangle) = \frac{1}{\sqrt{N}} \boldsymbol{a}^\top \sigma(\boldsymbol{W}^\top \boldsymbol{x}),$$

where $\boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{W} \in \mathbb{R}^{d \times N}, \boldsymbol{a} \in \mathbb{R}^N$, σ is the nonlinear activation function applied entry-wise, and the training objective is to minimize the empirical risk. Our analysis will be made in the *proportional asymptotic limit*, i.e., the number of training data n, the input dimensionality d, and the number of neurons N jointly tend to infinity. Intuitively, this regime reflects the setting where the network width and data size are comparable, which is consistent with practical choices of model scaling.

We showed that the first gradient update contains a rank-1 "spike," which results in an alignment between the first-layer weights and the linear component of the teacher model f^*. To characterize the impact of this alignment, we compute the prediction risk of ridge regression on the conjugate kernel after one gradient step on W with learning rate η, when f^* is a single-index model. We consider two scalings of the first step learning rate η. For small η, we establish a Gaussian equivalence property for the trained feature map, and prove that the learned kernel improves upon the initial random features model, but cannot defeat the best linear model on the input. Whereas for sufficiently large η, we prove that for certain f^*, the same ridge estimator on trained features can go beyond this "linear regime" and outperform a wide range of random features and rotationally invariant kernels. Our results demonstrate that even one gradient step can lead to a considerable advantage over random features, and highlight the role of learning rate scaling in the initial phase of training.

References

J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, D. Wu, and G. Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. to appear.

S. Okumoto and T. Suzuki. Learnability of convolutional neural networks for infinite dimensional input via mixed and anisotropic smoothness. In *International Conference on Learning Representations*, 2022.

量子乱数プロトコルを用いた量子計算機の安定性評価

鹿野　豊
群馬大学大学院理工学府
Institute for Quantum Studies, Chapman University
email: yshikano@gunma-u.ac.jp

概 要

All computing devices, including quantum computers, must exhibit that for a given input, an output is produced in accordance with the program. The outputs generated by quantum computers that fulfill these requirements are not temporally correlated, however. In a quantum computing device comprising solid-state qubits such as superconducting qubits, any operation to rest the qubits to their initial state faces a practical problem. On the implementation of the scalable quantum computers, the health check (or stability check) algorithms are needed. We propose that the quantum random number generation is one of the candidates of the health check algorithms in any quantum computing devices.

1　はじめに

量子計算機に関して、量子アルゴリズムと呼ばれる計算アルゴリズムの根本的な変更により、これまで知られていた従来型の計算量とは別の量子計算量という概念を創出した。特筆すべきは、現代暗号の計算量的安全性の基盤として用いられている「素因数分解のアルゴリズム」は **NP** (Non-deterministic Polynomial time) のクラスであることは知られているが、量子コンピュータによって誤り確率が高々1/3 の多項式時間で解ける決定問題の複雑性クラスである **BQP** (Bounded-error Quantum Polynomial time) で解けるという Simon-Shor のアルゴリズムというものが知られている。そのため、量子計算機を物理系として実現するための開発のレシピを必死に模索してきており、現在は小規模ではあるが、クラウド上でユーザーが自由にアルゴリズムを実行できる量子計算機も実現している。一方で、量子状態をそのまま情報のソースとして用いるためにノイズに対して従来型のデジタル計算機より耐性がないことが知られている。このため、一つ一つの計算素子の単位である量子ビットの正確な評価を行い、物理的な原因を追跡した上で量子ビットの改善をこれまで行ってきた。しかし、このようなボトムアップでのやり方は量子ビット数が増えれば増えるほど、評価だけに時間がかかってしまうため、ユーザーが計算機として使える時間が減ってしまうということが知られている。そのため、簡易に量子計算機の評価・診断ができるプロトコルの開発が必要であるということが必要であることは必然であろう。

そこで、本論稿では、量子計算機を計算機システムの一つだと捉えると、インプットおよびアウトプットにはビット列を用いたブラックボックスモデルとシステム的な観点で考えることが出来る。この際、デジタル計算機と量子計算機の本質的な差は、インプットに対してアウトプットが確定的に出てくるデジタル計算機に対して [1]、インプットに対してアウトプットが確率的に出てくるため、何度も同じアルゴリズムを実行し、その結果から統計的な判断を加えなければならないのが量子計算機の特徴である。すると、量子計算機の特徴を調べるためには、本質的に確率的な出力を行うシステムであるということを統計的に判定する問題に落とし込むことが出来る。そのため、本論稿では、量子計算機が確率的出力を本質的にするシステムであることをチェックするためにシードのない乱数生成器 [1][2]であるということを検証する方法を検討する。

2　量子乱数プロトコル

我々の提案する量子乱数プロトコルは非常に単純である。量子状態を用意する空間として量子力学の公理より 2 次元の複素ヒルベルト空間 $\mathcal{H} := \mathbb{C}^2$ を考える。そして、初期の量子状態として

$$|0\rangle := \begin{pmatrix} 1 \\ 0 \end{pmatrix} \in \mathbb{C}^2 \tag{1}$$

[1]当然、デジタル計算機にも必須なノイズがあるため、誤り訂正符号という技術を実装させ、確定的にアウトプットのビット列を出力させなければならない。

[2]シードのある乱数生成器のことの代表例はデジタル計算機の中で実装される疑似乱数生成器である。物理乱数生成器の多くは初期条件の違いに対して出力結果が鋭敏であることを利用しているため、真性乱数生成器とは呼ばれていない。

を用意する。そして、ここにアダマールゲート $H \in U(2)$ という状態を作用させると

$$|\psi\rangle := H|0\rangle = \frac{1}{\sqrt{2}}\left(|0\rangle + |1\rangle\right) = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix} \tag{2}$$

となる。量子力学の公理より測定結果に関しては、量子状態に対して自己共役作用素 A をスペクトル分解する。今回の場合、$A = \sigma_z \in U(2)$ というパウリ行列を採用すると、

$$\sigma_z = +1|0\rangle\langle 0| + (-1)|1\rangle\langle 1| = +1\begin{pmatrix}1&0\\0&0\end{pmatrix} + (-1)\begin{pmatrix}0&0\\0&1\end{pmatrix} =: +1\hat{P}_0 + (-1)\hat{P}_1 \tag{3}$$

となる。ここで、$\langle 0| := (|0\rangle)^\dagger, \langle 1| := (|1\rangle)^\dagger$ という Dirac 記法を採用している。これは、射影作用素 $\hat{P}_0, \hat{P}_1$ に対して、それぞれに対応する固有値が測定結果として得られる。その固有値が得られる確率はボルンの確率規則より

$$\Pr\,(*) := \|\hat{P}_*|\psi\rangle\|^2,\ (* \in \{0,1\}) \tag{4}$$

となることから、今回の例においては

$$\Pr\,(0) = \Pr\,(1) = \frac{1}{2} \tag{5}$$

となる。そして、測定結果を得た後に初期状態にリセットさせる。こうすることで何度でもアルゴリズムが実行できるようになる。これは 1 量子ビットに対して、全て同じ操作を行っていることからシードのない乱数生成器と捉えることが出来る。

本論稿では、これを用意されたクラウド量子計算機の全ての量子ビットに対して、同じ操作を「同時に」[3]。次に、このアルゴリズムを 5 日間実行させ、その結果として量子ビットが安定的な確率的な出力を得ているかということを単純に実行回数に対して「1」の出た積算回数をグラフ化することにより直線になっていなければ不安定に動作したということが分かった [3]。そして、初期状態へのリセットが不十分な場合、出力されたビット列に対して時間相関が検出されてしまう可能性があるため、チェックを行ったところ、時間相関がランダムに行っている可能性が示唆された [4]。このような検定を行うことで量子計算機の安定動作に関する指標を構築できる可能性があるということを示した。

3　おわりに

量子計算機をシステムとして捉えることで、量子乱数プロトコルが量子計算機の簡易診断プロトコルになるということを指摘した。本論稿で提案したプロトコルをクラウド量子計算機で実行したところ、全て安定的に動作されていないということが統計的な処理により明らかになった。

今後、NIST Test Suites を拡張させ、バイアスのある確率出力がある場合での乱数検定手法を開発させ、更には、安定性動作指標として本論稿では実行回数に対して「1」の出た積算回数をグラフ化することにより直線になっていなければ不安定に動作したということを判定基準としたが、その他にもランダム化試験のような統計的手法を適用することが重要である。統計学的基準を利用した量子計算機の指標づくりは、これまでも行われてきたが、今後、更に重要であると予想するため理論的な側面のみならず、実際のデータ処理をした際のソフトウェアの整備まで含めて開発していかねばならない宿命である。

参考文献

[1] Y. Shikano, AIP Conference Proceedings **2286**, 040004 (2020).

[2] K. Tamura and Y. Shikano, TUCS Lecture Notes **30**, 13 - 25 (2019); Cryptology ePrint Archive, Paper 2020/078.

[3] K. Tamura and Y. Shikano, International Symposium on Mathematics, Quantum Theory, and Cryptography, Mathematics for Industry, vol 33 (Springer, Singapore, 2021) 17– 37.

[4] Y. Shikano, K. Tamura, and R. Raymond, EPTCS **315**, 18– 25 (2020).

[3] ハードウェアの制約上、同時に本当に実行されているわけではなさそうであるが、残念ながら量子ビット数を増やして動作させるという方向性に昨今は焦点が当てられており、どれくらい正確に量子操作が実行できるかという点においてはあまり改善が見られていない。特に長期的に安定的に動作できるシステムではないということは強調しておきたい。更には、2022 年 10 月現在であっても全ての乱数検定をパスするクラウド量子計算機は存在していない。

Bootstrap for Selecting a subset which contains all populations better than a standard

Jun-ichiro Fukuchi, Gakushuin University, Tokyo, Japan

1 Framework

When samples are obtained from k populations with the different means, statistician is often interested in finding populations whose means are larger than that of the standard population. For the case of normal populations, Gupta and Sobel (1958) formulated subset selection approach. In this approach, the goal is to select a subset which contains all populations better than a standard. At this presentation, we only describe the method for unknown standard case. The method for known standard is similar and it is omitted. Papers related to this methodology are Paulson (1952), Dunnet (1955), Tong (1969), Huang, Panchapakesan and Tseng (1984) among others.

A related problem is selecting a subset which contains the population with the largest mean. The basic framework is built by Gupta (1965). Swanepoel (1983) investigated the bootstrap method for subset selection of the largest mean. Fukuchi (2020) proved consistency and the second-order correctness of the bootstrap under the assumption that population distributions belong to the same location family. This type of homogeneity assumption is needed for the selection of largest mean problem since the statistic used for selection rule depends on the distribution function associated with the largest mean asymmetrically. Subset selection problem of better than standard is simpler in some sense and as a result the bootstrap is shown in this paper to be successful without homogeneity assumption on the distributions even when sample sizes are unequal.

2 Selection rule and the infimum of the probability of the correct selection

Let $\Pi_1, \ldots, \Pi_k$ be k populations with means $\mu_1, \ldots, \mu_k$, and the common variances σ^2. Let Π_0 be the standard population. Denote the distribution function of Π_i by F_i, $i = 0, 1, \ldots, k$. It is assumed that the functional forms of F_i are unknown. Let $\mu_{[1]} \leq \mu_{[2]} \leq \cdots \leq \mu_{[k]}$ be ordered means. Let $\{X_{ij} : j = 1, \ldots, n_i\}$ is a sample from the population $\Pi_i (i = 0, 1, \ldots, k)$. Let $\overline{X}_i$ be the sample mean of $\{X_{ij}, j = 1, 2, \ldots, n_i\}$, $i = 0, 1, \ldots, k$ and S^2 be the pooled variance.

2.1 Equal sample sizes case

Consider the following selection rule proposed by Gupta and Sobel (1957).
Selection rule R_1 : Retain the population Π_i in the selected subset if and only if

$$\overline{X}_i \geq \overline{X}_0 - \frac{c_1}{\sqrt{n}} S \tag{1}$$

where c_1 is the smallest constant such that $\inf_\Omega P(\mathrm{CS}|\mathrm{R}_1) = p$ for prespecified $p \in \left(k^{-1}, 1\right)$. Let $\Omega := \{(\mu_0, \mu_1, \ldots, \mu_k) : \mu_i \in \mathbb{R}, i = 0, \ldots, k\}$. We make following assumption.

Assumption 1. For each $i = 0, \ldots, k$, F_i belongs to a location family, that is $F_i(x) = G_i(x - \mu_i)$ for some distribution function G_i.

Theorem 1. *Under Assumption 1,*

$$\inf_\Omega P(\mathrm{CS}|\mathrm{R}_1) = P_\Omega \left(\max_{1 \leq i \leq k} \left(\frac{\overline{X}_0 - \mu_0}{S/\sqrt{n}} - \frac{\overline{X}_i - \mu_i}{S/\sqrt{n}} \right) \leq c_1 \right)$$

2.2 Unequal sample sizes case

In this section, we consider the unequal sample size case. Consider the following selection rule.
Selection rule R_2 : Retain the population Π_i in the selected subset if and only if

$$\overline{X}_i \geq \overline{X}_0 - c_2 S \sqrt{\frac{1}{n_0} + \frac{1}{n_i}}, \tag{2}$$

where c_2 is the smallest constant such that $\inf_\Omega P(CS|R_2) = p$ for prespecified $p \in (k^{-1}, 1)$. See Bechhofer, Santner and Goldsman (1995), p130 for this selection rule for normal populations.

Theorem 2. *Under Assumption 1,*

$$\inf_\Omega P(\mathrm{CS}|\mathrm{R}_2) = P_\Omega \left(\max_{1 \leq i \leq k} \frac{(\overline{X}_0 - \mu_0) - (\overline{X}_i - \mu_i)}{S\sqrt{\frac{1}{n_0} + \frac{1}{n_i}}} \leq c_2 \right)$$

Let $T_{2,n} := \max_{1\le i\le k}\left((\overline{X}_0-\mu_0)-(\overline{X}_i-\mu_i)\right)\left(S\sqrt{\frac{1}{n_0}+\frac{1}{n_i}}\right)^{-1}$. Then c_2 is the p-th quantile of $T_{2,n}$ where p is prespecified number.

3 Estimation of c_2 by bootstrap

Estimating $c_2 = c_2(p)$ amounts to estimating the distribution function of $T_{2,n}$. Since F_i's are unknown, the constant $c = c_n(p)$ needs to be estimated from the sample. Let $n = (n_0, n_1, \ldots, n_k)$ and $\mathcal{X}_n = \{X_{ij} : j = 1, \ldots, n_i,\ i = 0, 1, \ldots, k\}$. Let $\mathcal{X}_n^* = \{X_{ij}^* :\ j = 1, \ldots, n_i, i = 0, 1, \ldots, k,\}$ be the bootstrap sample. Let P^* denote the conditional distribution of X_{ij}^*, $j = 1, \ldots, n_i, i = 1, \ldots, k$ given $\mathcal{X}_n$. Based on the bootstrap sample $\mathcal{X}_n^*$, the bootstrap version $T_{2,n}^* := \max_{1\le i\le k}\left((\overline{X}_0^*-\overline{X}_0)-(\overline{X}_i^*-\overline{X}_i)\right)\left(S^*\sqrt{\frac{1}{n_0}+\frac{1}{n_i}}\right)^{-1}$ is computed. The bootstrap estimator $\hat{c}_2(p)$ of $c_2(p)$ is defined by the p-th quantile of the bootstrap distribution $P^*\left(T_{2,n}^* \le x\right)$. Theorem 3 states that bootstrap method described above is consistent. Unlike the bootstrap for subset selection of the largest mean, the bootstrap for subset selection of better than standard is consistent under rather weak assumptions on the population distributions and the rates how sample sizes increase.

Theorem 3. *Let $X_{i1}, \ldots, X_{in_i}$ be i.i.d. random variables with $\mathrm{E}(X_{ij}) = \mu_i$, $\mathrm{V}(X_{ij}) = \sigma^2$, $i = 0, \ldots, k$. Assume that*
(i) $\{X_{0j}\}_{j=1}^{n_0}, \ldots, \{X_{kj}\}_{j=1}^{n_k}$ are independent.
(ii) There exist $d_i \in [0, \infty]$, $i = 1, \ldots, k$ such that $n_i/n_0 \to d_i$.
Then, as $\min\limits_{0\le i\le k} n_i \to \infty$

$$\sup_{x\in\mathbb{R}} |P^*(T_{2,n}^* \le x) - P(T_{2,n} \le x)| \to 0,\ a.s. \tag{3}$$

and $|\hat{c}_2(p) - c_2(p)| \to 0,\ a.s.$ for any $p \in (1/k, 1)$.

When sample sizes are equal and population distributions are absolutely continuous, by Theorem 3 of Fukuchi (2020), bootstrap estimation of $T_{2,n}$ is second-order correct meaning that bootstrap approximation is better than normal approximation.

4 Concluding remarks

In this presentation, properties of bootstrap when it is used to estimate the constant in subset selection rule of selecting better than standard are investigated. We assumed the common unknown variance. Consistency and second-order correctness for the proposed bootstrap method are proved. Results of simulation study is given at the presentation.

References

[1] Bechhofer, R. E., Santner, T. J. and Goldsman, D. M. (1995). *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons.* Wiley.

[2] Fukuchi, J. (2020). A note on bootstrap for Gupta's subset selection procedure. *Sankhya A*, volume 82, p. 96–114 (2020).

[3] Gupta, S. S. (1965). On Some Multiple Decision (Selection and Ranking) Rules. *Technometrics*, 7, No. 2, 225-245

[4] Gupta, S. S. and Hsiao, P. (1983). Empirical bayes rules for selecting good populations. *Journal of Statistical Planning and Inference*. Volume 8, Issue 1, September 1983, Pages 87-101

[5] Gupta, S. S. and Sobel, M. (1958). On selecting a subset which contains all populations better than a standard. *Annals of Mathematical Statistics* Vol. 29, No. 1 (Mar., 1958), pp. 235-244.

[6] Huang, D.Y., Panchapakesan, S. and Tseng, S. T. (1984). Some locally optimal subset selection rules for comparison with a control. *Journal of Statistical Planning and Inference*. Volume 9, Issue 1, January 1984, Pages 63-72

[7] Paulson, E. (1952) On the Comparison of Several Experimental Categories with a Control. *The Annals of Mathematical Statistics*, Vol. 23, No. 2 (Jun., 1952), pp. 239-246.

[8] Tong, Y. L. (1969). On partitioning a set of normal populations by their locations with respect to a control. Ann. Math. Statist. 40, 1300-1324.

[9] Swanepoel, J. W. H. (1983). Bootstrap selection procedures based on robust estimators. *Comm. Statist. Theory Methods*, 12(18), 2059-2083.