様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 ２４ 日現在

機関番号：１７３０１
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１８０１０

研究課題名（和文）探索的データ解析における統計的推論とその応用

研究課題名（英文）Statistical inference in exploratory data analysis and its application

研究代表者
梅津　佑太（UMEZU, Yuta）

長崎大学・情報データ科学部・准教授

研究者番号：６０７９３０４９

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：近年のデータ科学では，検証すべき仮説が定まらないままデータが取得されることが多い．その際，検証すべき仮説の生成と，その仮説の検証を同じデータを用いて行う場合，選択バイアスの問題が生じてしまう．本研究では，選択バイアスを解消するために，selective inferenceと呼ばれるフレームワークに着目し，既存手法の課題の解決を試みた．主要な成果は，selective inferenceのアイデアを教師なし学習へ応用したことと，データの正規性を緩和することでより広いクラスのモデルに対してselective inferenceが適用可能であることを示したことである．

研究成果の学術的意義や社会的意義
近年のデータ科学では，検証すべき仮説が定まらないままデータが取得されることが多い．その際，検証すべき仮説の生成と，その仮説の検証を同じデータを用いて行う場合，選択バイアスの問題が生じてしまう．とはいうものの，データの分割や同じ環境での再実験が困難な場合に統計的なエビデンスを提供するためには，同じデータを用いて仮説の生成と検証を行うことが求められる．本研究では，selective inferenceのアイデアに基づき，いろいろな問題に対してこのような統計解析が可能であることを示した．

研究成果の概要（英文）：In recent data science, we often observe data without determining hypothesis to be tested. Particularly, severe selection bias could be occur when the same dataset is used both for generating the hypothesis to be tested and for testing it. Here, in order to correct the selection bias, we focus on the selective inference framework, and tried to improve the existing method. Our main results are the application of the idea of selective inference to unsupervised learning and the development of the method that can be applied to more general class of statistical model by relaxing the normality of the data.

研究分野：数理統計学

キーワード：モデル選択　selective inference　高次元漸近理論　教師なし学習　教師あり学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1) 統計的データ解析は，大きく検証的データ解析と探索的データ解析に分けられる．検証的データ解析とは，統計的仮説検定に代表されるように，事前に定められた仮説をデータから検証するものである．一方，探索的データ解析では，クラスタリングやモデル選択などを通して，データを要約することで検証すべき仮説を生成するものである．データから探索的に定められた仮説に対して，同じデータを利用してその仮説を検証する場合，選択バイアスと呼ばれる問題が生じる．そのため，データの分割や同じ環境での実験が行えない限り，探索的な解析と検証的な解析を同時に行うことは困難であった．一方，よく設計された実験計画によってデータが観測されるわけではない近年のデータ駆動型科学では，同じ条件下での実験は困難であり，しかも，あらかじめ検証すべき仮説が定まっていることはほとんどないため，古典的な統計手法を適用することができなかった．

(2) こういった状況では，データから探索的に生成された仮説に対して，何らかの方法で同じデータを利用して統計的なエビデンスを保証するといった方針を取らざるを得ない．例えば，線形回帰モデルにおいて生成される仮説とは，適当なモデル選択を通して変数選択するということであり，統計的なエビデンスとは，選択された各変数に対する信頼区間の構成や仮説検定を行うということである．このような枠組みの問題は post-selection inference として知られており，選択されうる仮説の候補すべてを考慮する simultaneous inference と，選択された仮説のみに着目する selective inference と呼ばれる方法に大別される．simultaneous inference は，古典的な統計解析のアイデアに基づく手法であるものの，元のデータの次元が 20 程度であっても計算機的に実行が困難となってしまう．一方，selective inference では，FPR（false positive rate)のような，古典的な統計推論で用いられる FWER（family-wise error rate）に代わる過誤を用いた手法であり，より大規模なデータに対しても適用できることが多い．このような手法が近年広く研究されているものの，上述の計算機的な問題や，モデルや仮説探索の方法に対する制限が強いという理由で，それほど汎用的な手法であるとは言えず，さらなる拡張が重要である．

２．研究の目的

以上の背景を鑑み，近年の大規模なデータ解析に対する post-selection inference のため，本研究では selective inference の拡張を試みる．2016 年に Lee *et al.*によって提案された selective inference のアプローチ（引用文献①）では，線形回帰モデルにおける特定のモデル選択手順（selection event）を条件づけることで，選択された各変数に対する条件付き推論が可能であることが示された．一方，彼らの提案法では，分散既知の線形回帰モデルであること，および，Lasso（least absolute shrinkage and selection operator）によるモデル選択が仮定されていた．この設定では，例えば，Lasso の selection event やモデル選択後の推定量など，すべての計算を厳密に行うことができるため，自然に切断正規分布に基づく条件付き推論が可能となる．ところが，ロジスティック回帰モデルをはじめとする一般化線形モデルを考えた場合，selection event やモデル選択後の推定量の陽な表現を求めることは困難である．そこで，本研究では，これらの条件を緩和し，より一般的なモデルに対して selective inference を拡張することが最大の目的である．

３．研究の方法

(1) 近年の post-selection inference は教師あり学習に対して精力的に発展しているものの，同様の問題は教師なし学習に対しても考えられる．例えば，変化点検出では時系列の構造が変化する点をデータから推定することが仮説の選択に対応し，選択した変化点に対して実際にモデルが変化しているか否かを，同じデータを用いて検証するということである．問題ごとに選択される仮説の選択手順を具体的に考えることによって，データの正規性に基づく selective inference の開発を行う．

(2) たとえ教師あり学習でも，一般化線形モデルのような，より広いクラスのモデルに対しては，selection event やモデル選択後の推定量に対する陽な表現を得ることは困難である．また，モデルの正規性を仮定しないため，Lee *et al.*のように切断正規分布を直接用いることができない．そのため，何らかの近似的な評価による妥当な統計推論の開発が求められる．本研究では，高次元漸近理論に基づき，正規性の仮定を緩和することで，ロジスティック回帰モデルに対して，シンプルな方法でモデル選択した場合の selective inference の開発を行う．

４．研究成果

(1) 1つ目の成果はデータの正規性に基づく selective inference を多次元時系列の変化点検出へ拡張したことである．多次元時系列の変化点検出は，構造が変化する時点のみならず，どの系列の構造が変化したかを知ることも重要となる．例えば，脳波の計測を考えた場合，どの電極でいつ構造が変化したかを知りたいということが動機の一つである．正規性に基づく変化点検出では累積和が尤度比に対応するため，本研究でもこれを基礎とした理論の開発を目指した．累積和を系列ごとに並べた行列を考えた場合，変化点や変化した系列の選択はある種の最適化問題の解として与えられるため，その最大化点の選択に関するアルゴリズムが selection event になることに着目し，どの時点でどの系列が変化したかを selective inference によって検証する手法を提案した．また，いろいろな数値実験を通して，提案手法が正しく統計的過誤を制御できることを確認した．また，漸近理論により，カーネル法に基づく独立性の検定に対する selective inference の開発も行った．

(2) 2つ目の成果は，教師あり学習における selective inference の拡張に関するものである．シンプルなスクリーニングによって選択されるモデルに対して，選択された変数のみを用いたロジスティック回帰モデルのパラメータの漸近分布を高次元漸近理論に基づき導出することで，分類問題に対する selective inference の開発を行い，数値実験を通して手法の有用性を確認した．本研究では，Lasso のような最適化問題を解くことなく selection event を書き下すことができるため，ポアソン回帰のような，より一般の回帰モデルに対しても同様の手法を開発することは可能であると考えられる．また，Lasso や orthogonal matching pursuit によってモデル選択する場合，枝かりを用いた効率的なアルゴリズムを開発することで超高次元線形回帰モデルに対しても selective inference が効率的に行えることを示した．

＜引用文献＞

① Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., Exact post-selection inference with application to the lasso, *The Annals of Statistics*, 44(3), 907-927, 2016.

５．主な発表論文等

〔雑誌論文〕　計8件（うち査読付論文　8件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Matsui Hidetoshi、Umezu Yuta	４．巻 3
２．論文標題 Variable selection in multivariate linear models for functional data via sparse regularization	５．発行年 2020年
３．雑誌名 Japanese Journal of Statistics and Data Science	６．最初と最後の頁 453～467
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s42081-019-00055-x	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Shinjo Keiko、Hara Kazuo、Nagae Genta、Umeda Takayoshi、Katsushima Keisuke、Suzuki Miho、Murofushi Yoshiteru、Umezu Yuta、Takeuchi Ichiro、Takahashi Satoru、Okuno Yusuke、Matsuo Keitaro、Ito Hidemi、Tajima Shoji、Aburatani Hiroyuki、Yamao Kenji、Kondo Yutaka	４．巻 15
２．論文標題 A novel sensitive detection method for DNA methylation in circulating free DNA of pancreatic cancer	５．発行年 2020年
３．雑誌名 PLOS ONE	６．最初と最後の頁 1～18
掲載論文のDOI（デジタルオブジェクト識別子） 10.1371/journal.pone.0233782	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Suzumura Shinya、Nakagawa Kazuya、Umezu Yuta、Tsuda Koji、Takeuchi Ichiro	４．巻 14
２．論文標題 Selective Inference for High-order Interaction Features Selected in a Stepwise Manner	５．発行年 2021年
３．雑誌名 IPSJ Transactions on Bioinformatics	６．最初と最後の頁 1～11
掲載論文のDOI（デジタルオブジェクト識別子） 10.2197/ipsjtbio.14.1	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yuta Umezu, Ichiro Takeuchi	４．巻 2
２．論文標題 Selective inference via marginal screening for high dimensional classification	５．発行年 2019年
３．雑誌名 Japanese Journal of Statistics and Data Science	６．最初と最後の頁 559～589
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s42081-019-00058-8	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Makoto Yamada, Yuta Umezu, Kenji Fukumizu, Ichiro Takeuchi	4．巻 84
2．論文標題 Post Selection Inference with Kernels	5．発行年 2018年
3．雑誌名 Proceedings of Machine Learning Research	6．最初と最後の頁 152～160
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Tsubasa Hirakawa, Takayoshi Yamashita, Toru Tamaki, Hironobu Fujiyoshi, Yuta Umezu, Ichiro Takeuchi, Sakiko Matsumoto, Ken Yoda	4．巻 9
2．論文標題 Can AI predict animal movements? Filling gaps in animal trajectories using inverse reinforcement learning	5．発行年 2018年
3．雑誌名 Ecosphere	6．最初と最後の頁 1～24
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/ecs2.2447	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Takuto Sakuma, Kazuya Nishi, Kaoru Kishimoto, Kazuya Nakagawa, Masayuki Karasuyama, Yuta Umezu, Shinsuke Kajioka, Shuhei J. Yamazaki, Koutarou D. Kimura, Sakiko Matsumoto, Ken Yoda, Matasaburo Fukutomi, Hisashi Shidara, Hiroto Ogawa, Ichiro Takeuchi	4．巻 33
2．論文標題 Efficient Learning Algorithm for Sparse SubSequence Pattern-based Classication and Applications to Comparative Animal Trajectory Data Analysis	5．発行年 2019年
3．雑誌名 Advanced Robotics	6．最初と最後の頁 134～152
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/01691864.2019.1571438	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Yuta Umezu, Yusuke Shimizu, Hiroki Masuda, Yoshiyuki Ninomiya	4．巻 71
2．論文標題 AIC for the non-concave penalized likelihood method	5．発行年 2019年
3．雑誌名 Annals of the Institute of Statistical Mathematics	6．最初と最後の頁 247～274
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10463-018-0649-x	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計10件（うち招待講演　4件／うち国際学会　1件）

１．発表者名
梅津佑太

２．発表標題
超高次元スパース加法モデルにおける変数選択

３．学会等名
科研費シンポジウム「統計学と機械学習の数理と展開」

４．発表年
2019年

１．発表者名
梅津佑太

２．発表標題
超高次元加法モデルにおける変数選択

３．学会等名
第22回情報論的学習理論ワークショップ(IBIS2019)

４．発表年
2019年

１．発表者名
Yuta Umezu

２．発表標題
Selective Inference for Change Point Detection in Multi-dimensional Sequences

３．学会等名
Chile-Japan Academic Forum 2018（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
梅津佑太，竹内一郎

２．発表標題
Selective Inference に基づく変化点検出とその応用

３．学会等名
日本応用数理学会2018年度年会（招待講演）

４．発表年
2018年

１．発表者名
梅津佑太

２．発表標題
Selective Inference に基づく多変量系列の変化点検出

３．学会等名
日本行動計量学会第 46 回大会（招待講演）

４．発表年
2018年

１．発表者名
梅津佑太，竹内一郎

２．発表標題
Selective Inference に基づくスパース線形回帰モデルにおける能動学習

３．学会等名
第21回情報論的学習理論ワークショップ（IBIS 2018）

４．発表年
2018年

１．発表者名
梅津佑太，竹内一郎

２．発表標題
Selective Inference under the Local Alternative

３．学会等名
2018年度 統計関連学会連合大会

４．発表年
2018年

１．発表者名
梅津佑太

２．発表標題
カーネル法に基づく超高次元モデル選択

３．学会等名
科研費シンポジウム「多様な分野のデータに対する統計科学・機械学習的アプローチ」

４．発表年
2020年

１．発表者名 梅津佑太
２．発表標題 Sparse Regularization Method and Information Criterion
３．学会等名 2020年度 統計関連学会連合大会（招待講演）
４．発表年 2020年

１．発表者名 梅津佑太
２．発表標題 超高次元加法モデルにおけるモデル選択
３．学会等名 2020年度 統計関連学会連合大会
４．発表年 2020年

〔図書〕　計2件

１．著者名 梅津 佑太，西井 龍映，上田 勇祐	４．発行年 2020年
２．出版社 講談社サイエンティフィク	５．総ページ数 208
３．書名 スパース回帰分析とパターン認識	

１．著者名 Bradley Efron、Trevor Hastie、藤澤 洋徳、井手 剛、井尻 善久、井手 剛、牛久 祥孝、梅津 佑太、大塚 琢馬、尾林 慶一、川野 秀一、田栗 正隆、竹内 孝、橋本 敦史、藤澤 洋徳、矢野 恵佑	４．発行年 2020年
２．出版社 共立出版	５．総ページ数 600
３．書名 大規模計算時代の統計推論	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関