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　６ 年　 ６ 月　 ２ 日現在

機関番号：３２６１２
研究種目：若手研究
研究期間：2018～2023
課題番号：１８Ｋ１８００９

研究課題名（和文）スパースモデリングによる発見的統計手法の開発

研究課題名（英文）Development of discovering statistical methods via sparse modeling

研究代表者
片山　翔太（Katayama, Shota）

慶應義塾大学・経済学部（三田）・准教授

研究者番号：５０７４２４５９

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：スパースモデリングによる発見的統計手法の開発を目指し，本研究課題では特に（1）高次元線形回帰モデルにおける差異検出および(2)超高次元パラメータを伴う2標本問題についての研究を行なった．(1)では，回帰係数ベクトルの差分を直接スパースに推定できる手法の開発を行い，その予測誤差，変数選択の一致性，バイアス除去による漸近分布の導出などの理論を整備した．（2）では，遺伝子データ解析への応用を目指して，2グループ間の差異を特徴付ける超高次元パラメータに対する統計的推測法を与えた．さらには，その手法を用いて実際のCovid-19に対する重症化患者と非重症化患者のRNA-seqデータを比較した．

研究成果の学術的意義や社会的意義
本研究課題で実施した研究(1)(2)はどちらも基礎的なものであり，それゆえに社会的意義も大きい．(1)については医療・経済・マーケティングなどへの応用が考えられ，提案手法の解釈可能性から，個体に依存した処置や介入へと繋がる．(2)については，遺伝子データからのさらなる有益な情報抽出が可能となる．また，どちらの研究も新規の方法論を開発しており，さらにはその理論保証も与えている．

研究成果の概要（英文）：Aiming at the development of discovering statistical methods via sparse modeling, particularly (1) difference detection in high dimensional linear regression models and (2) two sample problems with ultra high dimensional parameters are studied. In the theme (1), a method for directly and sparsely estimating difference in regression coefficient vectors is developed, and gave its prediction error, variable selection consistency and derivation of the asymptotic distribution based on de-biasing. In the theme (2), a statistical inference for the ultra high dimensional parameters that characterize the differences between two groups is provided for application to the analysis of gene data. Furthermore, the proposed procedure compared the RNA-seq data of high and low risk Covid-19 patients.

研究分野：高次元データ解析

キーワード：高次元データ　スパースモデリング　因果推論　多重検定

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

ビッグデータ時代の到来以降, 大規模かつ複雑なデータが世の中には溢れている. これに伴い, 新しいタイプの統計手法の開発が近年急速に進んでいる．そのひとつがスパースモデリングである．これは，一見膨大に見えるが本質的な情報は少ない（スパース）といった，ビッグデータの特徴を的確に捉えており, その有用性から爆発的に研究が進んできた. 特に, Lasso (Tibshirani, 1996)をはじめとしたスパース回帰の徹底的研究によって，1 次（平均）の構造については解析手法がほぼ確立しつつある.

しかしながら，スパース回帰では基本的に 1 次の構造しか抽出できず，得られる情報は専門家からすると当然であることも多い. 確かに, ビッグデータを検証的に解析できるようになったことは大きな進歩である. だが, 応用領域の発展にさらに貢献するためには，より複雑な構造も発見できる統計手法の開発を進めていかなければならない.

２．研究の目的

本研究では, スパースなビッグデータから情報を「発見的に」抽出する統計手法の開発を行う. 開発の軸は，パラメータを正確に 0 と推定できるスパース推定法である．この手法は，応用可能性が非常に高く，発見的手法の開発と相性が良い. だが，スパース推定への過信にも注意が必要である. 例えば，考え得る全てのパラメータをモデルに入れてスパース推定することで，一応の発見的手法を与えることができる. しかし，ビッグデータの場合はパラメータ数が爆発し，計算に膨大な時間が掛かることに加え，推定対象の大規模化に伴って理論的にも推定精度が悪くなってしまう. このような安易なモデリングも少なくない現状において，データの背景を考慮しながら，スパース推定に寄りすぎない発見的統計手法の開発を目指す.

３．研究の方法

本研究ではとりわけ，(1)高次元線形回帰モデルにおける 2 グループ間の差異検出，および，(2)超高次元パラメータを伴う 2 標本問題について研究を行なった．これらのテーマは基礎的であるが故に応用範囲も広く，研究を行う価値がある．以下では，テーマごとに研究の方法について記述する.

(1)医療・経済・マーケティングなどの分野においては，処置が結果変数に与える影響を推定することが重要な課題となる. 特に, 説明変数を所与のものとした, 条件付き処置効果(Conditional Average Treatment Effects，CATE)が重要となる．というのも，もし CATE を適切に推定できれば，どの個体に処置を行うべきかの知見が得られるためである. 本テーマでは，近年の計測機器やデータベースの発展に伴う，高次元説明変数に対する CATE の適切な推定手法について研究を行なった.

各結果変数と説明変数の間に線形回帰モデルを想定すると，CATE は各回帰係数ベクトルの差分と説明変数の内積で与えられる．そのため，その回帰係数ベクトルの差分を推定すれば良い. 従来のアプローチはグループごとにスパース推定を実行し, その差分を取るものである. しかしその方策では，CATE の推定自体は可能であるが，推定値のばらつきによって真に差がある箇所を特定することはできない．これでは，どの説明変数が CATE に効いているのか全く解釈できない．そこで本研究では，回帰係数ベクトルの直接的なスパース推定法の開発を行なった.

(2)遺伝子データ解析への応用を動機に，交絡変数が存在し得る状況において，2 標本問題すなわち 2 グループ間の比較問題について研究を行なった．もし各グループへの割り当てがランダムに行われていれば, 平均ベクトル間の比較問題に帰着するが, 遺伝子データ解析の文脈においては, ランダムな割り当てが実行不可能となるケースも多い. 例えば病気の有無などでグループ分けされている場合がそれに該当する．そのような場合, 2 グループ間の差が交絡変数を介して生じている可能性があり，交絡変数を適切にコントロールした比較が必要となる.

遺伝子データは一般的に変数の次元が 10,000 を超える．そのような超高次元データに対してはスパース推定が上手く機能しないことが経験的に知られている. そこで本研究では, 各結果変数（遺伝子）に対して比較のための検定統計量を構築し，そもそも 2 グループ間に差はあるのか，そしてその差はどの結果変数にあるのか，を検証できる方法論を開発した．これを実現するため，各結果変数と割り当て変数および交絡変数の間に一般化線形モデルを仮定し，また，割り当て変数と交絡変数の間にロジスティック回帰モデルを仮定した．一般化線形モデルを仮定した背景には遺伝子データの多様性がある．遺伝子データはその測定法によって連続・2 値・カウントなど様々な変数を有する. 上記のモデリングの下で, 割り当て変数の係数，すなわち 2 グループ間での差を特徴付けるパラメータについての推測法を与えた. なお, 割り当て変数の係数は各結果変数で異なると考える方が自然であるため，この問題は超高次元パラメータに対する統計的推測問題となる.

４．研究成果

(1)高次元線形回帰モデルにおける 2 グループ間の差異検出:回帰係数ベクトルの差分を直接的にスパース推定するために，まずは各グループの結果変数に重みを付け，その重みがどのような性質を持てば差分に直接アクセスできるかを明らかにし，その結果を元に重みの計算方法を導出した．なお，この重みは各グループへの処置確率（傾向スコア）に類似した情報を持っているが，計算の際にはその確率モデルを特定する必要は全くない．得られた重みを付与した結果変数を用いてスパース推定を行うことで，目的が達成できる．

具体的な重み計算は，各グループにおける説明変数の共分散行列をバランスさせながら，重みベクトルの一様ノルムを最小化することで達成される．その数値計算アルゴリズムを導出するために，当初は交互方向乗数法を適用する形で進めていたが，研究の途中で，説明変数の次元が数百程度でも計算負荷が非常に高いという問題が生じた．その原因のひとつは微分不可能な一様ノルムの最適化であり，それをスムーズに近似することで数値計算の効率化を行なった．

重み付けによって得られたスパース推定量の理論的性質も明らかにした．変数の次元とサンプルサイズが同時に大きくなる高次元漸近枠組みの下で，予測誤差の導出および変数選択の一致性を示した．ここで，変数選択の一致性とは，推定された推定量のサポートと真のサポートが一致する確率が 1 に近づくというものである．すなわち，構築したスパース推定量は，どの説明変数が CATE に効いているのかを正確に特定できている．また，構築した推定量は予測誤差の意味でどの程度「良い」のかを明らかにするため，最適性の検証も行なった．具体的には，傾向スコアが既知の場合において，予測誤差の minimax 最適レートを導出した．提案した推定量の予測誤差のレートは，minimax 最適なものよりも悪くなるものの，その差を規定するパラメータを明らかにすることはできた．提案推定量は傾向スコアが未知のより複雑な状況も許容しているため，その差が生じたと考えられる．

回帰係数ベクトルの差分に関する統計的推測を与えることも重要な課題である．しかしながら，スパース推定量は罰則付き最適化によって計算されており，その罰則が原因で推定量にバイアスが生じてしまう．そこで，Zhang and Zhang(2014)などで提案されている de-bias のテクニックを用いて，提案したスパース推定量からバイアスを除去し，差分に関する漸近正規性を高次元漸近枠組の下で導出した．これにより差分に対する検定・信頼区間の構築などが可能となった．

(2)超高次元パラメータを伴う 2 標本問題:まずは超高次元パラメータの同時検定問題について研究を行なった．これは「そもそも 2 グループ間に差はあるのか」の問いに答えるためである．前述したモデリングの下，Vansteelandt and Dukes(2022)によって与えられた効率的スコア関数（漸近分散が最小）に基づいて各パラメータの検定統計量を計算し，それらの最大値を取ることで同時検定問題に対する検定統計量を構築した．実際に検定を行う際には，その検定統計量の帰無仮説における分布が必要となる．そこで，高次元漸近枠組みにおける漸近帰無分布の導出を進めた．しかしながらその導出過程で，スコア関数に含まれる局外母数の推定に関する問題が生じた．局外母数の推定とスコア関数の計算において，同じサンプルを使ってしまうと，それらの依存関係の評価が困難となってしまったのである．そこで，使うサンプルを分割する Cross-fitting(Chernozhukov et al.,2018)のアイデアを適用し，その問題を回避した．そして，最大値型の検定統計量がガンベル型分布へと分布収束することを示し，さらには検出力も導出した．検出力を評価することで，少なくとも 1 箇所の結果変数において差が生じていれば，漸近的に正しくその差を検出できることが示された．

次に「どの結果変数において 2 グループ間の差が生じたのか」の問いに答えるため，各パラメータの検定統計量を用いた多重比較法を構築した．これは結果変数ごとにそのパラメータが 0 であるか否かを検定するものであり，当然ながら検定の多重性の問題が生じてしまう．そこで本研究では，Liu(2013)や Javanmard and Javadi(2019)などを参考に，偽発見率(False Discovery Rate, FDR)のコントロール法を与えた．そしてその方法が，漸近的に正しく FDR を規定の値にコントロールできることを証明した．

提案した発見的統計手法の有用性を示すため，実データへの適用も行なった．元々は Overmyer et al.(2021)で解析されている RNA-seq 遺伝子データを用いて，Covid-19 に対する重症化患者と非重症化患者の比較を行なった．
なお，この解析では重症化患者を集中治療室に入院してかつ人工呼吸器を装着した患者と定義している．年齢や性別などの分布が 2 グループ間で明らかに異なっており，それらが交絡変数になっていると考えられる．FDR をコントロールしながら差の生じた遺伝子を抽出したところ，有名な既存手法である edgeR や DESeq2 と比較しておおよそ半分の遺伝子を発見できた．この結果は，交絡変数のコントロールによって，それを介さない純粋な差を厳選

Selected gene ontology

Proposed	edgeR	DESeq2
GO:0002283	GO:0002283	GO:0002283
GO:0002446	GO:0002446	GO:0002446
GO:0042110	**GO:0034470**	GO:0042119
GO:0042119	GO:0042119	GO:0043312
GO:0043312	GO:0043312	**GO:0060491**
GO:0045785		
GO:0046635		

For reference

GO term	description
GO:0002283	neutrophil activation involved in immune response
GO:0002446	neutrophil mediated immunity
GO:0034470	ncRNA processing
GO:0042110	T cell activation
GO:0042119	neutrophil activation
GO:0043312	neutrophil degranulation
GO:0045785	positive regulation of cell adhesion
GO:0046635	positive regulation of alpha-beta T cell activation
GO:0060491	regulation of cell projection assembly

できたと考えられる．さらに，選択された遺伝子を用いて，遺伝子オントロジーエンリッチメント解析を実行した．遺伝子はその機能によっていくつかのカテゴリー（遺伝子オントロジー）に分類される．エンリッチメント解析を実行すると，選択された遺伝子群がどのような遺伝子オントロジーを多く含んでいるのか解析できる．上図は各手法によって選択された遺伝子オントロジーを示しているものであり，青字が提案手法独自のもの，赤字が既存手法独自のものである．すなわち，既存手法と比較して新たに 3 つの遺伝子オントロジーを発見できている．

今後の展望: 上述の(1)(2)の研究においては，結果変数と処置変数に関連する交絡変数がすべて観測されているという条件の下で発見的手法の開発を行なっていた．しかしながら，実際のデータ解析においては未観測の交絡変数の存在は避けられない．例えば遺伝子データを考えると，個体の先祖情報・データ収集環境などは多くの場合未観測の情報であり，これらは結果変数と処置変数の両方に関係し得る．そのため，今後の研究において，(1)(2)の研究を未観測交絡変数が存在する状況下へと拡張していく必要があるだろう．

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　2件／うちオープンアクセス　0件）

１．著者名 Cho Seonghun, Katayama Shota, Lim Johan, Choi Young-Geun	４．巻 -
２．論文標題 Positive-definite modification of a covariance matrix by minimizing the matrix ℓ_{∞} norm with applications to portfolio optimization	５．発行年 2021年
３．雑誌名 AStA Advances in Statistical Analysis	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10182-021-00396-7	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 該当する

１．著者名 Katayama Shota, Fujisawa Hironori, Drton Mathias	４．巻 7
２．論文標題 Robust and sparse Gaussian graphical modelling under cell-wise contamination	５．発行年 2018年
３．雑誌名 Stat	６．最初と最後の頁 e181
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/sta4.181	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 該当する

１．著者名 Katayama Shota	４．巻 201
２．論文標題 Computational and statistical analyses for robust non-convex sparse regularized regression problem	５．発行年 2019年
３．雑誌名 Journal of Statistical Planning and Inference	６．最初と最後の頁 20,31
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jspi.2018.11.001	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13件（うち招待講演　8件／うち国際学会　4件）

１．発表者名 片山翔太
２．発表標題 High dimensional tests under observed confounding
３．学会等名 研究集会「多変量統計学・統計的モデル選択の新展開」（招待講演）
４．発表年 2023年

１．発表者名
Katayama Shota

２．発表標題
High dimensional tests on multivariate regressions under confounding

３．学会等名
High dimensional regression in biomedical applications, EcoSta 2023（招待講演）（国際学会）

４．発表年
2023年

１．発表者名
Katayama Shota

２．発表標題
High-dimensional multiple testing under confounding

３．学会等名
Development and Integration of High-Dimensional Data Analysis, Sparse Estimation, and Model Selection Methods（招待講演）（国際学会）

４．発表年
2024年

１．発表者名
片山翔太

２．発表標題
高次元データにおける交絡調整を伴う最大値型検定

３．学会等名
統計関連学会連合大会

４．発表年
2022年

１．発表者名
岡本憲暁，片山翔太，星野崇宏

２．発表標題
未測定交絡因子が存在する場合における制御された直接効果の推定法とその性質について

３．学会等名
日本計算機統計学会第36回シンポジウム

４．発表年
2022年

１．発表者名
岡本憲暁，片山翔太，星野崇宏

２．発表標題
未測定交絡因子が存在する場合における制御された直接効果の識別

３．学会等名
日本分類学会シンポジウム

４．発表年
2022年

１．発表者名
片山翔太

２．発表標題
Direct sparse estimation of conditional average treatment effects via covariance matrix balancing

３．学会等名
統計関連学会連合大会

４．発表年
2021年

１．発表者名
片山翔太

２．発表標題
Hypothesis testing on high dimensional parameter under confounding

３．学会等名
International Symposium on New Developments of Theories and Methodologies for Large Complex Data（招待講演）（国際学会）

４．発表年
2021年

１．発表者名
片山翔太

２．発表標題
Direct estimation of individualized treatment effects via approximate balancing

３．学会等名
Doshisha statistical meeting

４．発表年
2019年

１．発表者名
Shota Katayama

２．発表標題
Direct estimation of conditional average treatment effect in high dimensions

３．学会等名
International symposium on theories and methodologies for large complex data（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
片山翔太

２．発表標題
セルワイズ外れ値に頑健なスパースグラフィカルモデリング

３．学会等名
日本行動計量学会（招待講演）

４．発表年
2018年

１．発表者名
Katayama Shota

２．発表標題
Robust and sparse Gaussian graphical modelling under cell-wise contamination

３．学会等名
Japanese Joint Statistical Meeting CSA-KSS-JSS Joint International Sessions（招待講演）

４．発表年
2018年

１．発表者名
Katayama Shota

２．発表標題
Robust and sparse Gaussian graphical modelling under cell-wise contamination

３．学会等名
International Symposium on Statistical Theory and Methodology for Large Complex Data（招待講演）

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

HP : https://sites.google.com/view/skatayama/home

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			
韓国	Seoul National University	Sookmyung Women’s University		
デンマーク	コペンハーゲン大学			