様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　24 年　5 月　30 日現在

機関番号：32641
研究種目：若手研究(B)
研究期間：2010 ～ 2011
課題番号：22700293
研究課題名（和文）
リサンプリング法による多重検定の理論的研究とゲノム解析への応用
研究課題名（英文）
Resampling-based multiple testing and its application to genome analysis
研究代表者
酒折　文武（SAKAORI FUMITAKE）
中央大学・理工学部・准教授
研究者番号：90386475

研究成果の概要（和文）：本研究では，ゲノム解析分野での応用を念頭におき，高次元における多重検定問題および変数選択問題について検討を行った．リサンプリングに基づく多重検定法については，理論的な整備，とくにパーミュテーション法の問題点を明らかにすることができた．また，L1 正則化法を用いたスパース推定法による有効な変数選択法の開発と，適切な評価指標の導入を行うことができた．

研究成果の概要（英文）：In this study, we investigated high-dimensional multiple testing problem and variable selection problem for application to the genome analysis. As a result, we have organized some theoretical problems of permutation-based method for the resampling-based multiple testing methods. Also we have developed some new variable selection methods using L1 penalized sparse estimation methods and their evaluation methods for various situations.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	2,300,000	690,000	2,990,000
2011 年度	800,000	240,000	1,040,000
総　計	3,100,000	930,000	4,030,000

研究分野：総合領域
科研費の分科・細目：情報学・統計科学
キーワード：統計的推測

１．研究開始当初の背景

ゲノム解析，数理ファイナンスなど様々な分野において，複数の仮説を同時に検証するような多重検定の問題を扱う場面が増えてきている．例えば，DNA マイクロアレイ解析における数千～数万と測定された遺伝子発現データについて変異部位を同定する問題や，一塩基多型（SNP）による疾患関連遺伝子を同定する問題，線形モデルを用いた遺伝子間の制御関係を探る遺伝子ネットワーク推定の問題などが挙げられる．

このような問題においては一般に，同時に検証すべき変数の数がとても多い，いわゆる高次元であり，変数間が独立とは限らず相関関係が未知であり，変数がスパース性を有することが多い．こうした状況下では，従来の大標本理論に基づく統計的推測では歯が立たないことも多い．

そこで近年脚光を浴びてきている研究分野に，リサンプリング法に基づく多重検定がある．また，遺伝子ネットワーク推定などで用いる線形回帰モデルや時系列モデルなどにおいては，多重検定のみならず，L1 正則化法により変数選択を行う方法の研究が盛んである．こうした近年の研究動向の整備と新たな方法論の開発を主眼として本研究を開始した．

２．研究の目的

上述したように，リサンプリングに基づく多重検定法や L1 正則化による変数選択法に関する理論整備と，ゲノム解析等の分野における種々の仮説検証・変数選択問題を解決するための新たな方法論の開発が本研究の目的である．具体的には，２つの問題について，それぞれ以下のような目的で取り組んだ．

(1) リサンプリングに基づく多重検定

ゲノム解析の分野において，多くの仮説を同時に検証する多重検定法は重要な役割を担っている．上述したような，高次元・変数間の未知の相関関係の存在・スパース性といった特徴から，ブートストラップ法やパーミュテーション（並べ替え）法などのリサンプリング法を用いた多重検定法のニーズが高まっている．これらリサンプリング法は計算機をフルに活用した手法であり，近年様々な分野で活用されてきているが，実際には理論的な検証を十分に踏まえずに適用している場面もみられる．

本研究は，こうしたリサンプリングに基づく多重検定法の理論的な整備を行い，新たな知見の獲得や新たな手法の開発を行うことを目的とした．

(2) L1 正則化法に基づく変数選択

遺伝子ネットワーク推定等の問題においては，回帰モデルや時系列モデルなど様々な線形モデルが用いられる．このようなモデルにおいて，複数の仮説を同時に検証する問題は，モデルに含まれる変数選択問題に対応する．こうした変数選択問題は一般に，多重検定法のほか，情報量規準を用いた方法，L1 正則化法を用いた方法などが用いられている．とくに，変数の数がサンプルサイズよりも大きい高次元の場合には，パラメータ推定を行ったあとに変数選択を行う多重検定法・情報量規準を用いた方法の適用には限界がある．そこで L1 正則化法を用いてパラメータ推定と変数選択を同時に行う方法が重要である．

本研究では，とくにゲノム解析で見られるデータ構造を念頭に置き，こうした L1 正則化法の理論的な整備や多重検定法との特徴の比較を行うこと，さらに具体的なモデルに対する新たな方法論を開発することを目的とした．

３．研究の方法

上記 2 つの具体的な研究目的に従い，それぞれ以下のように研究を進めた．

(1) リサンプリングに基づく多重検定

リサンプリングに基づく多重検定法の問題点としては，Pollard and van der Laan (2004) による，パーミュテーション（並べ替え）による方法に関する指摘がある．本研究ではまず，この問題点の把握とその解決法についての検討を行うべく理論研究を行った．

平均の二標本問題におけるパーミュテーション法（並べ替え法）は，両標本に含まれる個体をランダムに 2 グループに振り分ける操作（並べ替え）によって検定統計量の標本分布を求め，仮説検定を行うというものである．Pollard らの指摘は，この操作に問題があるという，手法の本質に関わる指摘である．理論的に検証した結果，帰無仮説が正しいときにはこの操作に問題はないが，対立仮説が正しいときに，検定統計量の標本分布が適切に推定されないという問題があることがわかった．したがって，P 値が適切に推定されないことになる．

この問題は，パーミュテーション法が帰無仮説の下での各観測値の交換可能性に依っていることに起因する．したがって，この問題を解決するには，対立仮説の下でも検定統計量の標本分布を適切に推定できるような新たなアルゴリズムの構築が必要である．しかしながら本研究においては，この問題を解決できる新たな手法の提案には至らなかった．

また，調査を進めるうちに，リサンプリング法と(2)で行った L1 正則化法を組み合わせた方法が近年注目されてきていることがわかった．この手法についての調査も合わせて行い，他の手法との主眼の違いや有効性を確認した．

(2) L1 正則化法に基づく変数選択

遺伝子ネットワーク推定などの問題において，線形回帰モデルや時系列モデルなどが用いられる．さらに，しばしばデータが高次元であったり，外れ値を含んだりする．とくに考慮すべきは，サンプルサイズよりも変数の数が大きい場合についてである．こうしたデータにおいてはパラメータ推定をどのように行うかが問題となり，推定を行ったあとに多重検定により検証を行うことは現実的に難しい．そこで，パラメータ推定と変数選

択を同時に行う L1 正則化法を用いることが多い.

本研究においては，既存の L1 正則化法の整理とゲノム解析における事例研究を確認した上で，以下のような 2 つのモデルに関して L1 正則化法の開発を進めた.

① 時系列モデル

遺伝子ネットワーク推定において AR モデルや VAR モデル等の時系列モデルが用いられることがある．本研究では，AR モデルと VAR モデルのみに限らず，より一般的な時系列モデルに適用可能な L1 正則化法を用いたパラメータ推定・変数選択法の開発を目指した.

時系列モデルにおける lasso など L1 正則化法については，これまでに様々な研究が行われてきている．しかしながら，時系列モデルにおいて重要な lag（時差）を考慮にいれたモデリングはこれまでに行われてきていない.

本研究では，AR モデルや VAR モデル，ADR モデルなど様々な時系列モデルに適用可能な手法として，ある種の重み付き L1 正則化法を提案した．提案手法は，時系列モデルにおける lag に着目し，lag が大きくなるにつれて変数の影響が小さくなるという時系列モデルの一般的な性質を取り入れたものである．すなわち，lag の大きい変数ほど制約の強くなるような重みをつけた L1 正則化法である．adaptive lasso の変法であると解釈できることから，推定法については adaptive lasso における方法と同様の方法を用いることができた.

② ロバスト回帰モデル

線形回帰モデルもまたゲノム解析の分野で用いられる基本的なモデルである．その中で，データに外れ値が含まれる場合には，線形回帰モデルでは適切にパラメータ推定ができず，L1 正則化法を用いた場合にもその影響は大きいと考えられる．線形回帰モデルにおける L1 正則化法についてはすでに膨大な研究が存在するが，ロバスト回帰モデルに関する L1 正則化法については，いくつかの手法が提案されているのみである．そこで本研究では，ロバスト回帰モデルにおける L1 正則化法の既存研究の整理と，新たな手法の開発を目指した.

回帰モデルに対するロバストな推定量としては，M 推定量，LTS 推定量，Regression Depth による方法など様々なものが提案されている．本研究では，こうした M 推定法や LTS 推定法についての L1 正則化法についての研究を行った．とくに，LTS 推定法は，観測された残差が大きいものに関しては外れ値として除外するため，パラメータの推定に用いることのできる観測値が少なくなる．場合によっては，パラメータの数が観測値よりも多くなってしまう可能性がある．lasso 等の方法では，観測値の数以上の変数を選択することができないため，場合によっては適切なモデルとならない可能性もある．そこで，elastic-net を用いることにより，適切なパラメータ推定と変数選択を行うことを提案した.

M 推定量や LTS 推定量などの特徴として，チューニングパラメータが必要となる点が挙げられる． L1 正則化法を用いた場合は正則化にもチューニングパラメータが必要であるため，多くのチューニングパラメータを同時に決定することが必要となる．既存研究におけるチューニングパラメータの決定法としては交差確認法を用いることが多かったが，計算コストの問題の他，交差確認法自体の不安定性の問題がある.

そこで本研究では，情報量規準を用いたチューニングパラメータ決定を考えた．用いる推定量が最尤推定量ではないことを考えると，GIC やブートストラップ法による情報量規準などを用いる必要がある．本段階では，ブートストラップ法に基づく情報量規準を検討した．さらに，計算コストの削減のため，efficient bootstrap 法を用いた EIC の適用を提案した．これによって，情報量規準におけるバイアスの推定量の分散を減少させること，すなわち，ブートストラップ法における繰り返し数を少なく取り計算コストを削減することに成功した.

４．研究成果

上記 2 つの具体的な研究目的に関する研究成果はそれぞれ以下の通りである.

(1) リサンプリングに基づく多重検定

3(1)で述べたように，パーミュテーション（並べ替え）方に基づく多重検定法についての問題点の確認を行うことはできたが，具体的な解決法の提案には至らなかった．このことから，ゲノム解析の分野でリサンプリングに基づく多重検定を行う際にパーミュテーション法が用いられることが多いが，現在の段階ではパーミュテーション法よりもブートストラップ法を用いた方が適切であることが示唆された．この知見は非常に重要であり，何らかの方法で広めていく必要があると考えられる．また，パーミュテーション法の適切な利用に関しては，今後も継続的な研究が不可欠であると考えられる.

また，2010 年に提案された stability selection 法は，(2)の研究で扱ったような線形モデルやグラフィカルモデル等で用いられる方法であるが，そのアイデアの根源と

なっているのはパーミュテーション法を用いたデータの非復元抽出である．False positive, False negative に着目している既存手法に対し，True positive に重きを置いているところに特徴がある．今後は，この方法の更なる検証や理論的考察が必要であると考えられる．

(2) L1 正則化法に基づく変数選択

① 時系列モデル

様々な時系列モデルに適用可能な手法として，lag による重み付き L1 正則化法を提案した．提案手法についての理論的性質（推定量の一致性等）については明らかにできていないが，数値実験により，既存方法よりも平均二乗誤差など複数の観点から優れていることを示した．

しかしながら，本研究で提案した方法では，時系列データに周期性がある場合には適切な推定を行うことができないという問題点がある．もちろん，分析を行う前に季節調整法等で周期性を除去した後に用いることは可能であるが，一般的な時系列データ全般において提案モデルが適用可能では無い点には注意が必要である．

今後は，変数の数や lag の数が高次元である場合についてより深い検証が必要であると考えられる．また，提案手法における推定アルゴリズムやチューニングパラメータの決定法の開発も必要であろう．

② ロバスト回帰モデル

本研究では，M 推定量や LTS 推定量などを用いたロバストな回帰モデルに関して，L1 正則化法を用いたパラメータ推定と変数選択法についての研究を行った．とくに，elastic-net を用いることにより，適切なパラメータ推定と変数選択を行うことを提案した．さらに，モデルに含まれる多くのチューニングパラメータを同時に決定する方法として，efficient bootstrap 法による情報量規準 EIC の適用を提案した．数値実験を通して，ロバストな elastic-net の有用性と，efficient bootstrap 法による分散の抑制の効果を確認することができた．以上の結果により，高次元の線形モデルにおいて，外れ値を含む場合の適切な推定・変数選択法と，モデルに含まれるチューニングパラメータの決定法を提案することができた．

今後の課題としては，より計算コストを削減するために，多数の繰り返しを必要とするブートストラップ法を用いるのではなく，GIC や Mallows の Cp のような基準によるチューニングパラメータの決定が挙げられる．

また，今回提案した efficient bootstrap 法に基づくチューニングパラメータ決定法は，ロバスト回帰モデルのみならず様々な線形・非線形モデルにおいても活用できると考えられる．こうした研究も今後の課題である．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① Park, H. and Sakaori, F., Lag weighted lasso for time series model, Computational Statistics, 査読有, 2012, 掲載決定.

〔学会発表〕（計３件）

① Park, H. and Sakaori, F., Robust lasso-type estimation and efficient bootstrap information criteria in regression modeling, Joint meeting of the 2011 Taipei international statistical symposium and 7th conference of the asia regional section of the IASC, 2011 年 12 月 18 日，台北（台湾）.

② Park, H. and Sakaori, F., Lag weighted sparse regularization for multivariate time series, The 58th world statistics congress (ISI2011), 2011 年 8 月 23 日，ダブリン（アイルランド）.

③ Park, H. and Sakaori, F., Lag weighted lasso, 日本統計学会春季集会, 2011 年 3 月 6 日，立教大学.

６．研究組織

(1)研究代表者

酒折　文武（SAKAORI FUMITAKE）

中央大学・理工学部・准教授

研究者番号：90386475