様式 C-19

科学研究費補助金研究成果報告書

平成２２年　５月　７日現在

研究種目： 若手研究(B)
研究期間： 2007 ～ 2009
課題番号： 19700270
研究課題名（和文）　ロバスト推測法の展開とゲノムデータ解析への応用

研究課題名（英文）　Robust inference and its application to genome data

研究代表者
藤澤　洋徳（FUJISAWA HIRONORI）
統計数理研究所・数理・推論研究系・准教授
研究者番号： 00301177

研究成果の概要（和文）： 統計的手法の多くは，外れ値が存在すると，妥当性を失うことがある．外れ値が存在しても妥当性を失わない統計的手法はロバスト推測と言われている．特に，外れ値が存在したとしても潜在バイアスを小さく抑えて，手法の妥当性を保つことが，大きな目的の一つとなっている．本申請課題では，ロバスト推測の研究を発展させ，その手法をゲノムデータの解析に応用した．

研究成果の概要（英文）： When outliers are present, many statistical methods lose the validity of the result for analyzing the data. The statistical method which gives the valid result even if outliers are present is called the robust inference. In particular, it is important to make a latent bias small even when outliers are present. I have discussed robust inference and applied the method to genome data.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	700,000	0	700,000
2008 年度	600,000	180,000	780,000
2009 年度	600,000	180,000	780,000
年度			
年度			
総　計	1,900,000	360,000	2,260,000

研究分野： 統計科学
科研費の分科・細目： 統計科学
キーワード： ロバスト推測．遺伝子発現データ．一塩基多型．

１．研究開始当初の背景

データを解析するとき，多くのデータ解析手法は，データに外れ値が含まれていないことが前提になっている．まずは理想的な状態で最善の手法が提案されるわけである．ところが，データには，しばしば外れ値が含まれている．この瞬間に，最善の手法は，場合によっては最悪の手法になる．使おうとしているデータ解析手法が外れ値に大きく引きずられやすいデータ解析手法であれば，目の前に提示される解析結果は全く信用が置けないものになる．このような外れ値による悪影響を抑えようという手法がロバスト推測法である．

ロバスト推測法の研究の歴史は長い．特に，

外れ値が少ない場合には，様々な方法が提案されていた．しかしながら，外れ値が多い場合には，外れ値が少ない場合ほどには，十分に妥当な方法が提案されていなかった．モデルを特定しての場合に応じた提案はある程度はなされていたが，一般的なモデルに対しての統一的な方法は十分には提案されていなかった．

研究代表者らは，近年になって，一般的なモデルに対してのパラメータの推定において，外れ値が多い場合にも，ロバストの考えをきちんと整理することで，外れ値が少ない場合と同様の議論が可能であることを示し，同時に，パラメータ推定値を得る際に，かなり汎用的で使いやすい数値アルゴリズムも提案した．この研究は，学会やシンポジウムで何度も発表し，高い評価を得ている．現在は二つの論文にまとめて投稿中である．

遺伝子発現データや SNP データなどのゲノムデータに基づく解析は，最近になっても盛んに行われている．研究代表者は 5 年ほど前からその研究に携わっている．そこで，常に問題となることの一つが，外れ値の問題であった．新しく出現したタイプのデータには良くある話である．どうしても，データの採取の段階から手探りが続くため，その先にあるデータ解析の方法論との整合性を含めて，全てに流れ作業を整えた実験計画に基づいて，データを採取するには至らない．必然的に，何らかの意味でのたくさんの外れ値を扱うことになる．

上述のパラメータのロバスト推定は，もともとは純粋に理論的な動機から出発したのであるが，その最中に，ゲノムデータの解析に関わって，外れ値を扱うことの重要性を感じたことはさらに強い動機となった．現在は遺伝学研究所と癌研究所のゲノムデータの解析に関する研究に参加している．当初は，既存のロバスト推測法や，研究代表者らが開発したロバスト推定法で対処していたが，現在では，外れ値が多い場合の検定やモデル選択にまで，そのニーズが求められている．本研究では，そのようなニーズに応えられる統計的手法の開発と，その手法を実際に現場に適用し，そこから湧き起こる問題に対してもさらに対処していくことを目指している．

２．研究の目的

以下に，現在のところの具体的な研究テーマを，簡単に三つ挙げて説明したい．具体的な話は「３．研究の方法」に記したい．

一つ目は，遺伝学研究所との議論において発生した問題で，外れ値グループが多いときに，それを検定によって有意であると導出する手法の開発である．この問題は，簡単なようで決定版はなく，古くから統計学でマスク効果と呼ばれている問題である．さらに，マスク効果を克服するための過去の手法は，外れ値グループが多い場合をきちんと克服していない．その問題に対して，マスク効果を克服する道筋はある程度は見えているが，まずは基本的な問題をきちんと克服して，さらに一般的にマスク効果を克服するための道筋も作りたいと考えている．

二つ目は，癌研究所との議論において発生した問題で，外れ値が多い場合のモデル選択問題である．外れ値にロバストなモデル選択規準は，外れ値にロバストなパラメータ推定の一つ上の段階にある．過去に色々と提案されてはいるが，その妥当性は必ずしも明確でなく，パラメータ推定のときの考えを，なんとなく適用しているという様相である．本研究では，モデル選択規準自体をある意味でのパラメータと捉えることで，可能な限り普通のパラメータ推定と同じ土俵に載せて，モデル選択規準自体をロバスト推定するという考えで立ち向かいたいと考えている．

三つ目は，外れ値が多い場合のロバスト推測に関する，様々な幾何学的な解釈を整理することである．研究代表者らが提案したパラメータのロバスト推定法は，自然な幾何学的解釈をもつことが示され，結果的に，そういう幾何学的な解釈をもつ方法は唯一つしかないことが証明され，さらに，幾何学的な解釈を考えることで，パラメータ推定値を得るための自然な数値アルゴリズムを提案することも可能になった．幾何学的な解釈を整備することで，提案している方法が最適であるのか，より妥当な方法がないのか，などを突き詰めて考えやすくなる土壌が揃う．そのような幾何学的解釈をできるだけ掴んで，さらに有効なロバスト推測法を考える原動力としたい．

３．研究の方法

平成 19 年度

遺伝学研究所では，日本晴ゲノムをベースにした遺伝子発現データを利用して，様々なイネの様々な遺伝子の機能を調べようとしている．そのときに，日本晴以外のイネを結合しようとすると，うまく結合しない場合が頻繁に発生して，そのような大量の外れ値データをどのように扱うかということが問題とされていた．

単にモデルを仮定して何らかのパラメータを推定するだけであれば，研究代表者らが提案したロバスト推定法で十分である．しかしながら，遺伝学研究所では，さらに，外れ値グループの幾つかを，検定を使って有意差を提示したいという要求であった．外れ値が多いため，普通の方法では，統計学で有名なマスク効果の問題が起きる．ゆえに，うまい

方法で，外れ値を扱う必要がある．

歴史的には，マスク効果を克服するのに，外れ値グループの個数が幾つであるかを事前に想定して問題点を克服するなどの方法が提案されていた．しかしながら，研究代表者らが提案したロバスト推定法に基づいて，検定統計量を構築すれば，外れ値グループの個数を，事前に想定することなく，合理的にマスク効果を克服できると想像している．

その基本となるアイデアは次である．

まずは，外れ値グループであると検証したいグループ以外の残りのデータに対して，ロバスト推定法を利用して，大枠の基本構造を同定する．このとき，検証したいグループ以外の残りのデータにも，外れ値グループが存在することは多く，残りのデータにおいては，外れ値は多いと想像すべきことがポイントである．研究代表者らが提案したロバスト推定に基づいて考えれば，その大枠の基本構造を同定することは可能となる．

次に，外れ値グループであると検証したいグループにも，基本構造はあると想定する．大枠の基本構造から見て外れ値グループであっても，そのグループ内では基本構造が存在することが想定できたためである．その基本構造もロバスト法で同定する．

しかるのちに，大枠の基本構造から見て，外れ値グループであると検証したいグループの基本構造が外れの位置にあるかを検証することにするのである．つまり，何度もロバスト法を用いて，小さな構造ごとに外れ値を除いていって，もともと外れ値がなかったかのように最終的に検証を行えるようにするわけである．

まずは，上述のアイデアの数理的部分をきちんと整理したい．そして，実際の大規模データに応用して，どこまできちんと働くのかを試す予定である．さらに，遺伝学研究所からは，他にも，様々な外れ値グループ同定の状況を想定されている．そのような話にも対処できる方法を考えて行きたい．

平成 20 年度以降

前年度に引き続き，データが得られる状況に応じて，マスク効果の問題を考えることになるであろう．そして，それに対応した解析手法を構築しつつ，そこから類推される中心的部分を抜き出し，マスク効果への一般的な処方箋を構築したい．

それと同時に，次の二つの問題も考えて行きたい．

一つは，癌研究所の情報解析グループとの議論で発生している話題である．SNP データに基づいてハプロタイプブロックをどう同定するかが問題となっている．この同定問題は，ハプロタイプブロック構造によって決まる最適な統計モデルを選択する問題として，定式化することができる．

そのときに問題となるのは，非常に低頻度で起こる突然変異や組み換えなどの結果として生じる，低頻度のハプロタイプの扱いであった．つまり，ある種の外れ値と捉えることができる．現状では，外れ値の影響を自動的に取り除く方法ではなくて，そのようなハプロタイプを，アドホックに取り除いた後に，モデル選択を行っている．

本来であれば，そのようなアドホックさなしに，データから自動的に最適なモデルを選択したい．つまり，外れ値に影響されないロバストなモデル選択規準が欲しい．そのような研究を行いたい．

モデル選択規準自体をある意味でのパラメータと捉えることで，可能な限り普通のパラメータ推定と同じ土俵に載せて，モデル選択規準自体をロバスト推定するという考えで立ち向かいたいと考えている．研究代表者らが議論した相互エントロピーを利用することで，そのようなロバスト推定は最低限は可能である．

ただし，普通のパラメータの推定のときと違って，ロバストさは保てるが，ある意味でのバイアスが内在している．これの影響がはっきりしない．潜在的に含まれている外れ値の割合に依存する量なので，そのバイアス自体を適当に推定して補正するなどの対処を考えたりすることで，バイアスのない良いモデル選択規準が提案できるのではと考えている．

上述の内容で，統計学の基本的な話題である，推定・検定・モデル選択を扱うことになる．その後には，外れ値が多い場合のロバスト推測に関する，様々な幾何学的な解釈を整理したい．その意義については「２．研究の目的」で書いたとおりである．ただし，具体的な方法は，上述の検定とモデル選択に関する議論が進む方向に依存するので，具体的に書くことは難しい．パラメータのロバスト推定のときに考えたように，何らかの意味での射影であったり，ピタゴリアン関係であったりの観点から考えたいと思っている．

なお，本研究については，遺伝学研究所と癌研究所との議論を前提として書いている．その関係は良好であり，これからも継続されることは間違いないと考えている．万が一，その関係が途切れたとしても，本研究の継続には大きな支障は生じない．なぜならば，基本となる問題点については，既に研究代表者が十分に理解しているところであり，理論的な展開については支障は生じない．ゲノムデータの解析についても，二つの研究所から既にゲノムデータの基本部分は使用許可を頂いており，また，開発されるであろう手法は，二つの研究所からのデータを動機としているが，一般的な話題でもあるため，そういう

意味では支障は生じない．二つの研究所との議論があれば，さらに動機が深まるであろう，と考えている．

４．研究成果

外れ値の割合が小さい場合に対処するための方法論や理論は，これまでにかなり整理されていた．ところが，外れ値の割合が大きい場合には，幾つかの方法論や理論はあるものの，本来の目的である潜在的なバイアスを小さくする，ということを調べるための理論が整備されていなかった．研究代表者は，外れ値の割合が小さいか大きいかに関わらず，外れ値に対処するための理論を完成させた．そして，過去に提案されていた，ある方法が，その目的に適うことを示した．さらに，その目的に適う方法が，本質的に一つしかないことを証明した．これは画期的な結論だった．その論文が Journal of Multivariate Analysis に受理された．

Affymetrix の GeneChip 上のプローブには SNP が混在しているときがある．そのとき，得られたデータをそのまま解析すると，SNP の影響で，思ったような結果が得られないことがある．そのような SNP の影響を外してデータを解析するためにはどうすればよいのか，という問題がある．これまでは，そのような SNP に影響されているプローブを同定して，そのプローブの結果を使わないことにしてデータを解析すると言う二段階の解析が主流であった．本研究では，それらを同時に扱える方法を開発し，対応するソフトウェア SNEP を開発した．SNP に影響されているプローブから得られるデータを外れ値として考えることで，そのような目的を可能にした．特に，そのようなプローブの割合は小さいとはいえず，外れ値の割合が大きい場合として捉えられ，上述の方法論が見事にマッチした．他の類似方法と比較して，圧倒的にパフォーマンスが優れていることも確認した．補足的だが重要な知見も得られた．遺伝子発現データは，しばしば，当たり前のように，正規化という作業が行われる．ところが，それが逆効果をもたらす可能性を示唆することになり，同時に，SNEP の考え方を使えば，正規化なしにうまく行く可能性が示唆された．

上述の内容を塩基多型と表現多型を同時に同定する方法としてまとめた．投稿した論文は BMC Bioinformatics に受理された．対応するソフトウェア SNEP に関連した HP も完成した．その後，雑誌から，「あなたたちの論文は非常に高いアクセス数である」との報告をも頂いた．

この結果に基づいて，共著者を含んだ遺伝学研究所の倉田研究室を中心として，さらなる研究が進んでいる．現実のデータに対して SNEP のパフォーマンスがより詳しく調べられている．さらに，塩基多型の検出だけに特化した，より精度の高い方法を，別の方向から構築中である．SNEP には，プローブセット内に塩基多型が多すぎる場合には，パフォーマンスが落ちる問題点があったが，それの克服を目指している．

そのほかに，外れ値の割合が大きい場合におけるロバスト推測法において，単なる推定を超えた検定やモデル選択規準の開発があった．特に，外れ値の割合を，別に推定する必要があると考えていた．検定の方は，実は，外れ値の割合を明示的に推定する必要がなく，意外と簡単に構築できることとなった．モデル選択規準の方は，やはり，外れ値の割合を推定する必要がありそうである．その割合を推定する方法を考えて，モデル選択法の構築まで辿り着いた．今後は，そのパフォーマンスを調べていく予定である．

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 4 件）

① Fujisawa, H., Horiuchi, Y., Harushima, Y., Takada, T., Eguchi, S., Mochizuki, T., Sakaguchi, T., Shiroishi, T. and Kurata, N. (2009). SNEP: Simultaneous detection of nucleotide and expression polymorphisms using Affymetrix GeneChip. BMC Bioinformatics, Vol.10, No.131. 査読有.

② Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. Journal of Multivariate Analysis, Vol.99, 2053-2081. 査読有.

③ Fujisawa, H., Isomura, M., Eguchi, S., Ushijima, M., Miyata, S., Miki, Y., and Matsuura, M. (2007). Identifying haplotype block structure by using ancestor-derived model. Journal of Human Genetics, Vol.52, 738-746. 査読有.

④ Ninomiya, Y. and Fujisawa, H. (2007). A conservative test for multiple comparison based on highly correlated test statistics. Biometrics, Vol.63, 1135-1142. 査読有.

〔学会発表〕（計 16 件）

① 藤澤洋徳: Affymetrix GeneChip を利用した塩基多型と発現多型の同時同定，統計的推測法の発展と応用，2010.3.17，広

島
② 藤澤洋徳：SNEP: Simultaneous detection of nucleotide and expression polymorphisms using Affymetrix GeneChip, 統計科学セミナー, 2010.1.8, 福岡
③ Fujisawa, Hironori: Statistical Analysis of Affymetrix GeneChip Data, 2nd joint research meeting of ISI, ISM and ISSAS, 2009.10.12, Mishima, Japan
④ Fujisawa, Hironori: A simple selection of smoothing parameter in penalized spline regression, Joint Statistical Meting 2009, 2009.8.4, Washington, D.C., USA
⑤ Kuriki, Satoshi : Multiplicity adjustments in detecting reproductive barriers caused by Loci interactions, BIRS Workshop 09w5040 Random Fields and Stochastic Geometry, 2009.2.26, Banff, Canada
⑥ Fujisawa, Hironori: Robust parameter estimation with a small bias against heavy contamination, Statistical Science's colloquium series at Cornell University, 2009.2.25, New York, USA
⑦ 藤澤洋徳: 罰則付スプライン回帰におけるスムージングパラメータの簡単な選択法, 統計関連学会連合大会, 2008.9.8, 神戸
⑧ Fujisawa, Hironori: A sequential selection of smoothing parameter in penalized spline regression, 2008 Joint Meeting of ISI, ISM and ISSAS, 2008.6.19, Taipei, Taiwan
⑨ 藤澤洋徳: SNP を考慮に入れた遺伝子発現データ解析, シンポジウム「バイオインフォマティックスおよび経時観察データの解析」, 2008.2.8, 廿日市
⑩ Fujisawa, Hironori : A unified method for detecting single feature polymorphisms and gene expression level differences, Pasific Symposium on Biocomputing, 2008.1.6, Hawaii, USA
⑪ Fujisawa, Hironori : A unified method for detecting single feature polymorphisms and gene expression level differences, 1st joint research meeting of ISM and ISSAS (on Bioinfomatics), 2007.11.30, Tokyo, Japan
⑫ 藤澤洋徳： 異分野コミュニケーション～円滑にするコツは？～, 情報・システム研究機構 若手クロストーク研究会, 2007.11.26, 三島
⑬ 栗木哲: 遺伝子座間の相互作用による生殖的隔離障壁の検出と多重性調整, 統計関連学会連合大会, 2007.9.7, 神戸
⑭ 藤澤洋徳: SNP を考慮した遺伝子発現データ解析, 統計関連学会連合大会, 2007.9.7, 神戸
⑮ 藤澤洋徳: 数理統計学とデータ解析, 統計数理研究所 外部評価シンポジウム, 2007.8.13, 東京
⑯ 藤澤洋徳: SNP を考慮した遺伝子発現データ解析, 統計若手サマーセミナー, 2007.8.7, 指宿

〔その他〕
Fujisawa et al. (2009, BMC Bioinformatics) で開発したソフトウェア「SNEP」のホームページ
http://www.ism.ac.jp/~fujisawa/SNEP/

６．研究組織
(1)研究代表者
藤澤　洋徳 (FUJISAWA HIRONORI)
統計数理研究所・数理・推論研究系・准教授
研究者番号： 00301177