様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　　３　日現在

機関番号：　１７１０２
研究種目：　基盤研究(C)（一般）
研究期間：　2011～2015
課題番号：　２３５００３５３

研究課題名（和文）離散観測される確率場における分布理論の開発とその空間統計学への応用

研究課題名（英文）Developing distribution theory for discretely observed random field and its application to spatial statistics

研究代表者

二宮　嘉行（Ninomiya, Yoshiyuki）

九州大学・マス・フォア・インダストリ研究所・准教授

研究者番号：５０３４３３３０

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：主研究成果は「疑似相関を用いた新しい多重性調整法の開発」と「L1 正則化法である LASSO における正則化パラメータ選択のための AIC の導出」である．前者では，多重検定において検出力を向上させる新しい方法を導き，実際に実データ解析において有効であることを示した．後者では，一般化線形モデルの枠組みでの LASSO に対し，古典的な情報量規準と同じルーツをもつ情報量規準として唯一となるものを導いた．

研究成果の概要（英文）：The main results are ``Development of a new multiplicity adjustment using spurious correlations'' and ``Derivation of AIC for regularization parameter selection in the LASSO.'' In the former, a new method is developed to improve the power of multiple tests, and the method is shown to be efficient in real data analysis. In the latter, for the LASSO in the framework of generalized linear models, an information criterion is derived as a unique criterion which has the same root as the classical criteria.

研究分野：　数理統計学

キーワード：　確率場理論　ゲノム科学　スパース推定　多重検定　統計的漸近理論　モデル選択

１．研究開始当初の背景

(1) 多重検定について

大量のデータが得られるようになった現在，多重検定問題において大量の数の検定を扱うための手法が必要となっている．特に，相関の高い検定が存在している場合の手法は，ゲノム科学の分野と空間統計学を必要とする諸分野（疫学・画像解析・環境科学など）で需要が高く，本研究では後者を主のターゲットとする．

(2) ターゲットとする応用の一例

ある疾病について発生率の高い地域を検出したいという疫学の問題は「集積性検出」と呼ばれ，候補となる地域ごとに検定が構成されるために多重検定問題となる．地域の候補としては，例えば半径 r，位置 (s,t) の円領域が用いられ，$\theta=(r,s,t)$ の候補を集めた集合 Θ の要素数が多ければ検定は多数となる．θ に対する検定統計量を T_θ とすれば，$P(\max_{\theta\in\Theta}T_\theta>c)$ という裾確率の評価が多重検定の p 値を評価するために通常必要となる．

(3) 離散観測される確率場の超過確率

上述のような例においては，Θ として離散集合を考えることが多い．したがって，$\{T_\theta \mid \theta\in\Theta\}$ は添字 θ が近ければ相関が高くなるといった通常の性質を有する確率場（上述の例ならば三次元確率場）において離散観測したものとみなせ，$P(\max_{\theta\in\Theta}T_\theta>c)$ はその（閾値 c の）超過確率ということになる．

(4) 近年以前の研究動向

上述の超過確率に対する容易な評価の中で最も基本的なものは Bonferroni の不等式を用いた $\Sigma_{\theta\in\Theta}P(T_\theta>c)$ であり，その改良も昔から考えられてきた．しかしそれは $\{T_\theta\}$ が単純かつ特殊な相関構造をもつ場合の改良であり，その統計問題への応用は限られていた．

(5) 連続観測される確率場の超過確率

Θ が連続集合であるときの $P(\max_{\theta\in\Theta}T_\theta>c)$ の評価については，微分幾何を用いたアプローチが 1939 年に提案された． 1980 年代に入って再注目を浴び，研究を積み重ねられたその手法は，近年積分幾何を用いたアプローチによりさらなる発展を遂げた．

(6) 近年の研究動向

上述の微分幾何・積分幾何アプローチを利用して，あるいはそれに触発され，離散観測ケースに対する方法が提案された．Naiman & Wynn (1997) は，$\{T_\theta \mid \theta\in\Theta\}$ が低次元の確率変数で表現されるときに位相幾何を用いて Bonferroni の不等式を改良する理論を構築した．Efron (1997) は，Θ が一次元であるときに Θを曲線でつなげて微分幾何アプローチを用いる方法を提案した．また Taylor et al. (2007) は，Θ が格子点状である（二次元ならば $\{(a_i,b_j)\mid 1\leqq i\leqq I,1\leqq j\leqq J\}$ のように書ける）ときに積分幾何アプローチに基づく評価を与えた．

(7) 着想までの経緯

Ninomiya & Fujisawa (2008) において，私は Efron (1997) のアイディアを拡張し，Θ が二次元格子点状であるときに Θ を曲面でつなげて微分幾何アプローチを利用することを基にした評価を与えた．この方法の目的･応用が Taylor et al. (2007) のそれと同じであることに気づき，両方法の長所を活かすように融合させるべく知見を積み重ねた．結果，ただの融合にとどまらず，一般化の可能性に気づいたのが現況といえる．

(8) 本研究の位置づけ

集積性検出の例から想像できるように，本研究は基本的な統計問題を対象としている．それにも関わらずこれまでに扱われてこなかったのは，数年前の方法では取り組めない位置に問題が存在したからである．着想につながっている二つの論文は 2007 年と 2008 年のものであり，本研究は道具が揃って今取り組めるようになったもの，今取り組むべきものといえる．

２．研究の目的

(1) 新方法の提案

Taylor et al. (2007) と Ninomiya & Fujisawa (2008) を互いの長所を活かすように融合させ，超過確率 $P(\max_{\theta\in\Theta}T_\theta>c)$ を評価する新しい方法を提案する．両論文が扱っているのは Θ が低次元かつ格子点状であるケースだが，応用の一例で挙げた「集積性検出」では低次元であるものの格子点状にはならない Θ を考えるのが通常であるため，それにも対応できるように微分幾何アプローチの結果を加える．

(2) 方法の拡張・一般化

例えば集積性検出において候補となる地域の形状が円領域や長方形領域のような決まったものではない場合，Θ の次元は低くおさえられなくなり $\{T_\theta \mid \theta \in \Theta\}$ の相関構造は一層複雑になる．その相関構造をうまく掴むアルゴリズムを構築し，掴んだ後も困難が予想される確率評価に対して微分幾何・積分幾何・位相幾何アプローチを駆使し，どんな $P(\max_{\theta\in\Theta} T_\theta > c)$ に対しても性能の高い評価を与えきることが本研究の最大目標となる．

(3) 方法のカスタマイズ

集積性検出と同系統の応用問題として，画像解析における信号検出や環境科学におけるホットスポット検出がある．しかし同系統でも $\{T_\theta \mid \theta \in \Theta\}$ の相関構造には個々の問題特有の性質があるため，より性能の高い評価を目指し，その性質に即した方法を構築していく．またそれら個々の評価方法に対し，数値的な性能評価だけにとどまらず，裾確率（超過確率）の閾値に関する漸近理論に基づいた性能評価を行う．

3．研究の方法

(1) Taylor et al. の方法と Ninomiya & Fujisawa の方法の融合

評価する $P(\max_{\theta\in\Theta} T_\theta > c)$ の $\{T_\theta \mid \theta \in \Theta\}$ は，添字が近ければ相関が高く，遠ければ相関が低いとい確率場であり，マルコフ性に近い性質を有するといえる．したがって，例えば添字 $\boldsymbol{\theta}_0$ の近傍を $N_{\boldsymbol{\theta}_0}$ として $\{T_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in N_{\boldsymbol{\theta}_0}\}$ を条件付けると，$T_{\boldsymbol{\theta}_0}$ と $\{T_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \notin \{\boldsymbol{\theta}_0\} \cup N_{\boldsymbol{\theta}_0}\}$ はほぼ独立になる．この性質を用い，$\{T_\theta \mid \theta \in \Theta\}$ に関する大域的な確率の問題を各 $\{T_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \{\boldsymbol{\theta}_0\} \cup N_{\boldsymbol{\theta}_0}\}$ に関する局所的な確率の問題に落とすことが，Taylor et al. (2007) の方法の第一段階で行われている．この考え方はそのまま Ninomiya & Fujisawa (2008) の方法にも適用できるはずであり，その実現を計画の第一歩とする．

(2) 画像解析のための開発

Taylor et al. (2007) の方法と Ninomiya & Fujisawa (2008) の方法とでは近傍の定義の仕方が異なり，後者の方が前者を含む形になる．上で局所的な確率の問題に落とすと述べたが，そこには「ほぼ」独立となることから導いた近似が用いられており，したがって大きな近傍を用いた Ninomiya & Fujisawa (2008) の方法の方が性能の高い評価を与えることが予想される．このことを，両方法の共通応用問題「画像解析において形状既知の信号を検出する問題」で確かめていく．

(3) 空間疫学のための開発

上述の信号検出では Θ は格子点状に並んでいるが，「空間疫学における疾病などの集積性を検出する問題」では，たとえ候補地域の形状が既知であったとしても通常 Θ は格子状にならない．解析的評価がまだ与えられていないこのケースでも，局所的な確率の問題に落とすための近傍 $N_{\boldsymbol{\theta}_0}$ をうまく定義し，さらに微分幾何アプローチに基づく結果を加えることで，$P(\max_{\theta\in\Theta} T_\theta > c)$ の性能の高い評価を与えていく．

4．研究成果

(1) 23 年度

離散観測される正規確率場の超過確率（互いに相関をもつ正規確率変数の最大値の裾確率）を評価する Taylor et al. (2007) の方法と Ninomiya and Fujisawa (2008) の方法を融合して超過確率の新しい評価方法を与えることについて，理論上どのような相関構造をもつ正規確率場にも対応できるような完成形を与えた．これにより，実際に精度のよい評価を与えるかどうかは別にして，空間統計学における集積性検出・信号検出・ホットスポット検出やゲノム科学における QTL 検出などを目的とする検定問題の確率値を与えられるようにした．そして，実際にそれを遺伝的不適合が観測されるマウスにおける関連遺伝子探索のための QTL 解析に応用した．具体的には，位置情報のあるマーカーにおける遺伝型がわかっているマウスに対し，遺伝的不適合に関連する遺伝子の存在を検証するための検定問題を構築した．関連遺伝子の位置を未知とするのは当然のこと，優性効果のタイプも未知としたため，検定問題は相関が高くかつ個数の大きい多重検定問題として定式化される．ただし，マーカーの位置はわかっているため，また優性効果のタイプの候補は定めているため，多重検定間の相関構造は導出可能である．これより，この多重検定問題の確率値は離散観測される正規確率場の超過確率で評価できることになり，実際にその評価式を新しい方法で与えた．そして，二つの関連遺伝子を検出することに成功した．

(2) 24 年度

多重検定間の相関がわかっているとき，その相関の値を利用し，検定のサイズ（FWER; family-wise error rate）をコントロールしつつなるべく検出力が大きくなるように確率値を評価する一般式を与えたのが前年度である．これに対し，多重検定間の相関がわからないとき，前年度の評価式を利用しつつ新しい考え方を導入してより検出力を向上させたのが当該年度の成果である．例えば生物種間において複数の特徴量（例えば身長と体重）を比較する多重検定問題では，特徴量間の相関が一般に未知であるため，多重検定間の相関もわからない．このような場合，特徴量間の相関を一致推定量で置き換え，FWER を漸近的にコントロールする方法を用いるのが最も簡単な対処法である．しかし，特徴量間の相関がそれほど高くないとき，相関を利用してもそれほど確率値の改良（検出力のゲイン）は得られない．そこで，必ずしも一致推定量にはならない統計量を相関の代わりに用いることにより，FWER を漸近的にコントロールしつつ検出力を上げる方法を導いた．検出力を上げる方法は現在も開発され続けているがそのような発想を用いたものは存在せず，開発した手法による検出力のゲインは従来手法のそれよりはるかに大きいことが見込まれている．相関の代わりとなる統計量は容易に求められるものであり，この問題における標準手法となる価値をもつと確信している．統計量を用いて確率値を求める際には離散観測される正規確率場の超過確率の評価式を用いており，したがってこの開発は本研究課題において開発した理論が活かされる応用の一つと位置づけられる．

(3) 25 年度

グループ間の多次元データの平均を比較するという基本的な多重検定問題において，多次元データの相関が未知である設定を考える．このとき，必ずしも一致推定量にはならない統計量，具体的にはある種の標本疑似相関を相関に代入して用いることにより，多重検定のサイズを漸近的にコントロールしつつ検出力を上げる，というアイディアを得たのが前年度である．これに対し，アイディアを精錬させて手法の完成形を与えた後，数値実験を通して手法の有用性を示し，実データ解析に手法を適用したのが当該年度の成果といえる．具体的には，標本疑似相関のあるクラスの中で，ある種の最適性を満たすようなものを与え，それを実際に代入して用いるときに生じる数値計算上の問題を克服した．また，検出力を上げる既存手法であるステップダウン法と組み合わせ，さらなる検出力向上を達成した．数値実験では既存手法と比較し，常に検出力を上げることとともに，設定によってはステップダウン法によるゲインとは比較にならないほどのゲインがあることを示した．実データ解析では，コントロールマウスとその染色体の一本を別のマウスのものと入れ替えたコンソミックマウスに対し，身長や体重などの特徴量を比較する多重検定問題を扱った．この問題に対する過去の研究では相関の利用すら行われていなかったため，既存手法の適用のみで 47 個の検定を新たに棄却することになったが，本手法の適用でさらに 13 個の検定を棄却することに成功した．

(4) 26 年度

本研究課題に関して進んだ研究は，主に「関数データの変化点解析における情報量規準の導出」と「領域ごとに構造の複雑度が異なるデータに対する正則化法の開発」の二つである．通常のパラメトリック解析と異なり，関数データ解析では真の分布が仮定したモデルに属することを仮定せず，ある特殊な構造を仮定する．その仮定ゆえに変化点モデルに対して Ninomiya (2014) で導かれた AIC を用いることができないため，代わりに Cp 基準を漸近理論に基づいて評価したのが前者の研究である．これにより，変化点数の推定をおこなうことができる．評価のテクニックは Ninomiya (2014) におけるそれと同様のものだが，得られた情報量規準の形は Ninomiya (2014) におけるそれからは想像できないものとなっている．現在は数値実験により性能を評価している段階である．一方，構造の複雑度に応じて正則化パラメータの値を変動させる，という新しいタイプの正則化法を提案したのが後者の研究である．そして，そこに現れる（ハイパー）チューニングパラメータの選択のため，ある適当な漸近論に基づいて情報量規準を導いている．回帰分析あるいは判別分析において，場所によって回帰曲面あるいは判別境界の滑らかさが異なると，通常の正則化法では過適合または適合不足となる場所が生じてしまう傾向があるが，提案手法はその回避を目的としている．そしてその目的を果たすことを数値実験で確認し，さらに実データ解析をおこなっている．

(5) 27 年度

本研究課題に関して進んだ研究は，主に

「L1 正則化法である LASSO における正則化パラメータ選択のための AIC の導出」である．変数選択と推定を同時に遂行する LASSO は，空間統計学に限らないあらゆる統計・機械学習分野のホットトピックであり，この問題に対してもこれまでいくつかの情報量規準が提案されてきた．それらのほとんどはモデル選択の一致性を満たすように作られているが，結局その中から最適なものを決定する手段が存在せず，使うべき情報量規準が結局定まらないという大きな弱点をもつ．一方，一致性ではなく予測誤差最小化という観点で，LASSO のための AIC の有限補正は正規線形回帰モデルの枠組みで与えられている．これは上述の問題点を有さないが，一般化線形モデルの枠組みで与えられない．そこで，漸近論を用いた AIC 元来の定義に基づき，一般化線形モデルの枠組みで Kullback-Leibler ダイバージェンスの漸近不偏推定量を与えた．ちなみに，LASSO の漸近論は通常の推定量に対する漸近論の枠組みには含まれないため，確率場の最大値に関する理論が必要となり，本研究課題において積み上げてきたものを活用している．具体的には，凸解析に基づき，AIC の導出の際に必要となる種の確率場の確率収束を，ある正則条件のもとで導いた．与えた情報量規準は正規線形回帰モデルの枠組みで AIC の有限補正となるため，その一般化とみなせる．簡易な形で与えられるため，交差確認法に比べて計算コストは少ないが，パフォーマンスはほとんど同等か上回ることを数値実験で確認した．

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 9 件）

① Umezu, Y., Shimizu, Y., Masuda, H. and Ninomiya, Y., AIC for non-concave penalized likelihood method, *arXiv preprint*, arXiv:1509.01688, 2015.12., Non-refereed.

② Umezu, Y., Matsuoka, H., Ikeda, H. and Ninomiya, Y., Defect rate evaluation via simple active learning, *Pacific Journal of Mathematics for Industry*, 7 (8), 1-8, 2015.10., Refereed.

③ Ninomiya, Y., Change-point model selection via AIC, *Annals of the Institute of Statistical Mathematics*, 67, 943-961, 2015.10., Refereed.

④ 吉田 久男・二宮 嘉行, 変動化罰則付き最尤法による二群判別, *ISM Research Memorandum*, 1191, 2015.02., 査読無.

⑤ Kim, D., Kawano, S. and Ninomiya, Y., Adaptive basis expansion via l1 trend filtering, *Computational Statistics*, 29, 1005-1023, 2014.10., Refereed.

⑥ Ninomiya, Y. and Kawano, S., AIC for the LASSO in generalized linear models, *ISM Research Memorandum*, 1187, 2014.05., Non-refereed.

⑦ Ninomiya, Y., Signal detection and model selection, *Mathematical Approach to Research Problems of Science and Technology (eds. Nishii, R. et al., Springer Japan)*, 239-248, 2014.03. Non-refereed.

⑧ 二宮 嘉行, 信号検出と統計的モデル選択, *Kyushu University COE Lecture Note*, 64, 167-174, 2013.02., 査読無.

⑨ 二宮 嘉行・吉本 敦, ベイズ法を用いた単木成長予測モデル, ***森林資源管理と数理モデル***, 10, 333-349, 2011.04., 査読有.

〔学会発表〕（計 20 件）

① 因果推論モデルにおける周辺構造の選択のための情報量規準, 第 10 回日本統計学会春季集会, 2016.03.05.

② 疑似相関を用いた多重性調整, 研究集会「遺伝学と統計学における数理とモデリング」, 2016.01.25.

③ Lq 正則化法のための AIC, 研究集会「スパース推定と情報量基準」, 2016.01.07.

④ Regularization Parameter Selection for the Lasso in Generalized Linear Models, 9th Conference of the Asian Regional Section of the International Association for Statistical Computing, 2015.12.17.

⑤ LASSO による変数選択のための AIC, 統計関連合同大会, 2014.09.14.

⑥ AIC-type Information Criterion for LASSO, The 3rd Institute of Mathematical Statistics Asia Pacific Rim Meeting, 2014.07.01.

⑦ 疑似相関を用いた多重性調整およびその応用, 応用統計学会, 2014.05.22.

⑧ AIC による変数選択の問題点および LASSO について, ワークショップ「気候モデルの農業への応用 2」, 2014.01.16.

⑨ 変数選択問題と AIC, 第 3 回 数学・数理科学とシステム制御との連携研究集会, 2013.11.14.

⑩ P-value evaluation for multiple testing of means under the existence of positive correlations, 8th International Conference on Multiple Comparison Procedures, 2013.07.11.

⑪ 変量間に相関があるときの平均に関する多重検定について, 日本計量生物学会年会, 2013.05.23.

⑫ 統計的モデル選択に関する近年の動向, 計測自動制御学会, 2013.03.08.

⑬ 罰則付尤度比検定とチューブ法, 研究集会「数理統計学の沃野」, 2012.11.24.

⑭ A p-value evaluation for multiple testing problem based on highly correlated test statistics, International Biometric Conference, 2012.08.

⑮ Improving a geometrical approach for multiple testing problems, IMS Asia Pacific Rim Meeting, 2012.07.

⑯ AIC for estimating the number of structural breaks, Workshop "Time Series: Models, Breaks and Applications", 2012.02.

⑰ Likelihood ratio test for exploratory factor analysis model, 7th Conference of the Asian Regional Section of the IASC, 2011.12.01.

⑱ 信号モデルに対するモデル選択理論について, USN シンポジウム, 2011.10.

⑲ 因子分析における因子数選択のための分布理論, 統計関連学会連合大会, 2011.09.

⑳ スキャン統計量のための裾確率評価, 計量生物学会, 2011.06.

〔図書〕（計 1 件）

① 二宮 嘉行, *確率と確率変数*, 東京図書, 統計学 （日本統計学会編）, pp. 3--29, 2013.04.

〔その他〕
ホームページ等：
http://www.imi.kyushu-u.ac.jp/academic_staffs/view/51

６．研究組織
(1)研究代表者
二宮 嘉行（NINOMIYA YOSHIYUKI）
九州大学・大学院数理学研究院・准教授
研究者番号：50343330

(2)研究分担者
なし

(3)連携研究者
なし