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

平成　３０　年　　９　月　１１　日現在

機関番号：　６２６０３
研究種目：　基盤研究(B)（一般）
研究期間：　2013～2017
課題番号：　２５２８０００８

研究課題名（和文）ゲノム・オミックスデータ解析の安定化のための統計的方法論

研究課題名（英文）Statistical methodology for stabilizing of the genome-omics data analysis

研究代表者
江口　真透（Egichi, Shinto）

統計数理研究所・数理・推論研究系・教授

研究者番号：１０１６８７７６

交付決定額（研究期間全体）：（直接経費）　　12,900,000円

研究成果の概要（和文）：ゲノム・オミックスデータ解析に伴う高次元小標本の問題のためにデータ解析の安定化を目指した新たな統計的方法論の開発に挑戦した．特にゲノム・オミックスデータを用いて治療効果，予後などの予測のための従来法を超えた方法を提案し，実用的なアルゴリズムを発表した．主要な貢献として，遺伝子ランキングの安定化のために開発した‘サブサンプル繰り返し法’とデータ潜在的な異質性に対処した異なる線形予測関数を一般化平均でつなぐ‘準線形モデル’の提案と実用化が挙げられる．

研究成果の概要（英文）：In genome and omics data analysis we aimed at exploiting new statistical methods to challenge a difficult aspect caused by high-dimensional small-sample observations. In particular, we have proposed and published statistical methods and practical algorithms for prediction for treatment effect and prognosis based on genome and omics data beyond existing methods. In major contributions the subsampling replication method for a gene ranking was presented for a stable ranking, and a new modeling, called quasi-linear model, was proposed to incorporate different linear predictors into an integrated predictor connecting by a generalized mean.

研究分野：　統計科学

キーワード：　オミクス　データ　遺伝子ランキング　データ異質性　高次元・小標本　準線形モデル　表現型予測

１．研究開始当初の背景

(1) ゲノム・オミックスデータの統計的解析から高次元小標本の下で理論的・実際的な成果が得られつつある．しかし，現実のデータ次元数が標本数を超える場合は統計的結論の多重な解が生じる問題と低い汎化能力の問題が実用化の妨げになっていた。

(2) 申請者グループは統計的パターン認識，密度推定，クラスタリングのためにブースティングの理論的考察と実装化を得た。表現形の予測についてマイクロアレイ，プロテオーム，SNP を対象にし，クラスラベルとして疾病の種類や薬剤の奏功性，感受性を対象に関連する遺伝子やタンパクの予測発見について発表していた．しかし，高次元小標本のデータ解析の問題に根本的な解決策には到達できてなかった．

２．研究の目的

(1) ゲノム・オミックスデータから得られる統計的結論に多重な解が生じる問題と低い汎化能力に陥る問題の本質的な原因を解明する．このような高次元小標本での統計方法の性能の限界について理論的に明らかにすし，得られた性能の理論的限界を現実のデータ解析の場面で詳細に検討する。

(2) 統計予測問題で考えると特徴ベクトルの高次元性によってクラスラベル予測子の可能な関数形の自由度が過剰に増大する．この現象が様々な科学で起こり，例えば，医学での個人化医療を可能にしていることになる．個人が持っている医学的な詳細な情報が特徴ベクトルとして表されるとき高次元性が現れる一方で，表現型は依然，少数のクラスラベルで表さるの実データから統計予測を行うときに大きな困難に向き合わなければならない．単なる線形予測では良い性能が得られないことは自明である．このためデータ異質性を柔軟に反映させた非線形予測の提案と実用化を行いたい．

３．研究の方法

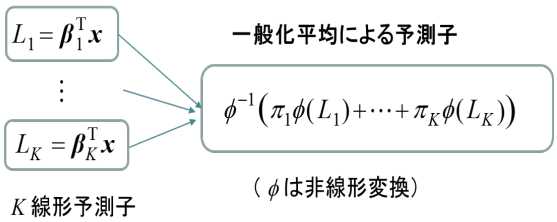
(1) ゲノム・オミックスデータ解析で広く使われている遺伝子ランキングについて安定なランキングが得られるような方法を開発する．近年，安定性解析で考えられているサブサンプリングの方法をランキングに適用する．この方法の良さをサポートするために漸近分布の導出を求める．このため，ノンパラメトリック統計の古典的な理論展開をサーベイする．

(2) ゲノム・オミックスデータの統計的解析を機械学習の枠組みから教師なし学習と教師あり学習の枠組みから整理する．この整理の下で，統計パターン認識として教師なし学習のクラスタリングと教師あり学習の統計予測を合体させることを考える．この指針から特徴ベクトルの成分をクラスタリングによってグループに分けて，グループごとの線形予測を統合した予測子を教師ありデータによって学習させるアルゴリズムを作る．数値実験を通して有効なチューニング法を決める．

４．研究成果

(1) 非線形学習のための準線形モデル　高次元小標本データの本質的に困難な問題を予測の内容で考えた．例えば，特徴ベクトルが網羅的な遺伝子の発現量のベクトル，クラスラベルが，ある疾病の患者であるか健常者であるかの場合を考えよう．このとき，被験者の患者グループの遺伝子発現パターンに潜在的な異質性があることが想定される．この潜在的な異質性は疾病の症候群に幾つかの潜在的なサブタイプから，しばしば起こることが報告されている．この研究では，次の 2 段階の手続きからなる方法を提案した．

　はじめに教師なし学習によって，潜在的な異質性によって生じる遺伝子のグループをクラスタラングによって決定する．次に教師あり学習の内容でクラスターごとの線形予測関数の学習と，一般化平均でつなぐ非線形学習を同時に行う．これによってクラスラベルの予測のためにクラスター内で有効な線形予測量を非線形な結合によって単一の予測量に統合し，全体の予測に有効な予測量を取り出すことに成功した．

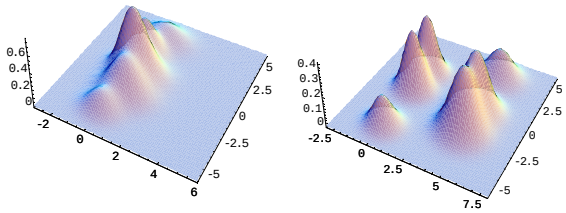


上の図のように各クラスターでは線形予測関数の学習を行いながら，全体の予測関数は生成関数ϕ によって一般化平均の形で統合している．特にϕとして指数関数を採用すると log-sum-exp の形になることが分かる．これに逆温度パラメーターを導入すると，このチューニングによって柔軟な非線形結合が得られることが分かった．

(2) 再現性の高いランキング法　2 値クラスラベルの予測問題で特徴空間の次元が膨大となるときに，特徴ベクトルの成分ごとの予測性能を 2 標本検定によって測り，その性能をランキングするアプローチがある．このランキングに基づいてこれによってトップ K の成分を選択して予測を行う方法が広く使われている．たとえば，FDA によって認可されている乳がん治療の予後予測に使われている Mamma Print は遺伝子ランキングによる

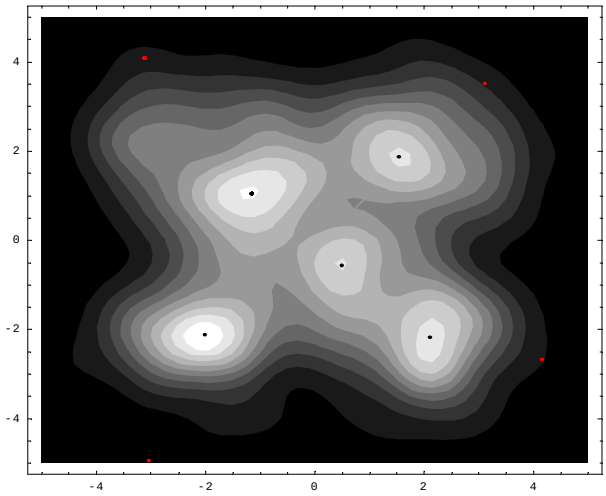
top 70 の遺伝子発現の線形予測法である．しかしながら公開されているデータの再現性テストでは，この top 70 の遺伝子の再現性は極めて悪い結果が確認された．やはり小標本に比べて膨大な数の遺伝子発現を得たことは，偶然に有意となった遺伝子があふれて安定したランキングが実行されてない．この研究ではランキング法を安定させるために全データを一度，学習させるのでなく，サブサンプルを繰り返し学習させる手続きが採用された．これによって再現性が脆弱な従来のランキング法をロバストな性能を持つ方法が提案された．理論的なサポートしてノンパラメトリック統計量の一つである U 統計量の漸近分布の導出を利用して，サブサンプルの 2 標本検定の平均によって定義される提案統計量の漸近分散の導出を行った．

(3) 一般化 t 統計量　2 標本検定で広く使われている t 統計量を利用した 2 値クラス予測の問題を考えた．線形予測子に対する t 統計量を一般化したクラスを考察して，このクラスの中で最適な統計量を与えるアルゴリズムを提案した．このクラスは生成関数 U によって記述されるので最適な U を求めるアルゴリズムとなる．さらに特徴ベクトルの高次元問題に対処するために Lasso タイプの変数選択により高次元ベクトルのモデル選択を行った．提案した統計量の理論的なサポートとして，クラスラベル 0 の分布は正規性を仮定するがクラスラベル 1 の分布は下の図で表されているように，混合正規モデルを含む多峰性を強要する広い仮定を考えた．



この仮定の下で任意の生成関数 U による統計量が統計一致性を持つことが証明された．さらに，漸近分散は U に依存する形で与えられ，最小となる形が陽に導出された．数値実験によって提案方法が従来法を明確に優越するシナリオが提示された．実データ解析においても良好な性能が検証されている．

(4) 自発的クラスタリング　教師なし学習の代表的な方法にクラスター分析がある．K 平均やモデルベイズドの方法は予めクラスター数を決める必要がある．提案した方法はクラスターセンターを決める際に自発的にクラスター数を学習する方法を考えた．これは (a) ガンマ推定アルゴリズムによって標本平均ベクトルを初期値として実行して平均の推定値を得る．(b)その平均推定値からもっとも離れたデータ点からガンマ推定アルゴリズムを実行して平均推定値を得る．(c) 2 つの平均ベクトル推定値が異なれば，さらに，最も離れたデータ点からガンマ推定アルゴリズムを実行する．このようにガンマ推定アルゴリズムの収束値が新たな収束値を持つまで繰り返し，得られた平均ベクトル推定値をクラスターセンターとして割り当てる．下の図では 5 つのクラスターセンターが得られているが，これはガンマ推定の超ロバスト性から一つのクラスターセンターは他のクラスターデータを外れ値として見なし，そのクラスターに属すデータだけを学習して平均ベクトルを推定している．このようにガンマ・ダイバージェンスの持つ自発学習性がうまく利用されている．



(5) マルチタスク学習の偽指数モデル　2 値予測問題を考えるときに偽指数モデルに対して一致性を満たすブースティングは Itakura-Saito ダイバージェンスに基づくものに限ることが示された．このことより，マルチタスク学習において，この Itakura-Saito ダイバージェンスを積極的に援用したブースティングの学習アルゴリズムを提案して，その性能について詳細に検討した．

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１５件）

① K. Omae, O. Komori and S.Eguchi, Quasi-linear score for capturing heterogeneous structure in biomarkers, BMC Bioinformatics, 18(308),2017, DOI:10.1186/s12859-017-1721-x, ページ無し，査読有

②Reproducible detection of disease-associated markers from gene expression data
BMC Medical Genomics 9(53) 2016 DOI: 10.1186/s12920-016-0214-5, **ページ無し**, **査読有**

③Robust Clustering Method in the Presence of Scattered Observations, A. Notsu and S. Eguchi
Neural Computation, 28(6), 1141-1162, 2016, DOI: 10.1162/NECO_a_00833, **査読有**

④Risk assessment of radioisotope contamination for aquatic living resources in and around Japan, Okamura H, Ikeda S, Morita T, Eguchi S.
Proceeding of National Academy of Science, 113(14), 3838-3843, 2016, **査読有**

⑤An asymmetric logistic regression model for ecological data, O. Komori, S. Eguchi, S. Ikeda, H. Okamura, M. Ichinokawa, S. Nakayama, Methods in Ecology and Evolution, 7(2) , 249-260, 2016, DOI: 10.1111/2041-210X.12473, **査読有**

⑥S. Kato and S. Eguchi., Robust estimation of location and concentration parameters for the von Mises-Fisher distribution., Statistical Papers, 2015, DOI10.1007/s00362-014-0648-9, **査読有**

⑦ O. Komori, S. Eguchi and J. Copas., Generalized t-statistics for two-group classification, Biometrics, 71(2), 404-416, 2015, DOI: 10.1111/biom.12265, **査読有**

⑧Kanao K., Komori O., Nakashima J., Ohigashi T., Kikuchi E., Miyajima A., Nakagawa K., Eguchi S. and Oya M., Individualized prostate-specific antigen threshold values to avoid overdiagnosis of prostate cancer and reduce unnecessary biopsy in elderly men, 2014, Japanese Journal of Clinical Oncology, 44 (9) , 852 – 859, DOI:10.1093/jjco/hyu133, **査読有**

⑨S. Eguchi, O. Komori and A. Ohara., Duality of maximum entropy and minimum divergence., Entropy, 16(7), 2014, 3552-3572, DOI:10.3390/e16073552, **査読有**,

⑩Notsu, A., Komori, O. and Eguchi, S., Spontaneous clustering via minimum gamma-divergence", Neural Computation, 26(2),421-448,2014, DOI:10.1162/NECO_a_00547, **査読有**

⑪ Chen, P-W., Hung, H., Komori, O., Huang, S-Y. and Eguchi, S., Robust independent component analysis via minimum gamma-divergence estimation, IEEE Journal of Selected Topics in Signal Processing, 7, 4, 614-624 (2013), DOI: 10.1109/JSTSP.2013.2247024, **査読有**

⑫Ohara, A. and Eguchi, S., "Geometry on positive definite matrices induced from V-potential function", Geometric Science of Information. Lecture Notes in Computer Science, 8085, 621-629 (2013) DOI: 10.1007/978-3-642-40020-9_69.

⑬ Ohara, A. and Eguchi, S., "Group invariance of information geometry on q-Gaussian distributions induced by beta-divergence", Entropy, 15, 11, 4732-4747 (2013), DOI:10.3390/e15114732.

⑭Notsu, A., Kawasaki, Y. and Eguchi, S., "Detection of heterogeneous structures on the Gaussian copula model using projective power entropy", ISRN Probability and Statistics, Volume 2013 (2013), Article ID 787141, DOI:10.1155/2013/787141.

⑮Komori, O., Pritchard, M. and Eguchi, S., "Multiple suboptimal solutions for prediction rules in gene expression data", Computational and Mathematical Methods in Medicine, Vol. 2013 (2013) Article ID 798189, DOI:10.1155/2013/798189. [**査読有**]

〔**学会発表**〕（**計２３件**）

①Shinto Eguchi, Katsuhiro Omae., Information Geometry of Predictor Functions in a Regression Model., 3rd conference on Geometric Science of Information, 09 Novembre 2017, Mines ParisTech, Paris (France)

②**江口真透．一般化平均によるモデルと推定．科研費シンポジュウム「統計学, 機械学習の数理とその応用」,公立はこだて未来大学, 2017 年 9 月 21 日 9 月 22 日**

③**林賢一，江口真透．擬似線形関数に基づくクラスタ毎の異質性を考慮した回帰モデル．統計連合大会，2017 年 9 月 6 日，南山大学名古屋キャンパス**

④ **江口真透，大前勝弘．回帰モデルの予測関数の情報幾何．統計連合大会, 2017 年 9 月 5 日，南山大学名古屋キャンパス**

⑤**小森理，三枝祐輔，江口真透．生態データのためポアソン点過程－準線形モデリング－, 統計連合大会，2017 年 9 月 5 日，南山大学名古屋キャンパス**

⑥**野津昭文，大前勝弘，江口真透, 一般化ガンマクラスタリングについて, 統計連合大会，2017 年 9 月 4 日, 南山大学名古屋キャンパス**

⑦**大前勝弘，江口真透．一般化平均を用いたコックス比例ハザードモデルの拡張．統計連合大会，2017 年 9 月 4 日，南山大学名古屋キャンパス**

⑧ S.Eguchi, Two paradigms in statistical prediction, The International Conference on Bioinformatics and Biostatistics for Agriculture Health and Environment, 2017**年** 1 **月** 21 **日**, take place January 20-23, 2017 at the

University of Rajshahi in Bangladesh.

⑨S.Eguchi, Information Geometry associated with two generalized means, Information Geometry and its applications IV, 2016 年 6 月 13 日, Liblice, Czech Republic

⑩Eguchi, S., Path connectivity on a space of probability density function space., 招待講演, Information geometry for machine learning, 和光, 2015.12.3.

⑪江口 真透、小森 理.，２群判別のための多重マーカーの一般化 t-統計量について:セミパラメトリックスとラッソ，招待講演，バイオ統計学の挑戦と貢献，福岡 2015.2.3

⑫Eguchi S., Komori, O., Ohara,, A. Duality in a maximum generalized entropy model., 34th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Amboirs France, 2014.09.25

⑬Komori, O., Eguchi, S., Maximum power entropy method for ecological data analysis., 34th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering., Amboise France. 2014.09.23, (ポスター)

⑭Takenouchi T. Komori O, Eguchi S., A novel boosting algorithm for multi-task learning based on the Itakuda-Saito divergence., 34th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Amboirs France, 2014.09.23

⑮江口 真透，統計学から学んだ三つのレッスン， 招待講演，統計関連学会連合大会，東京，2014.09.16

⑯大前 勝弘、小森 理、江口 真透，マイクロアレイデータによるロバストな遺伝子ランキングを用いた表現型予測，統計関連学会連合大会，東京，2014.09.14

⑰Komori, O, Eguchi, S., Assessment of fishery status based on mis-label model, The 3rd Institute of Mathematical Statistics Asia Pacific Rim Meeting , Taipei Taiwan, 2014.06.30

⑱Omae, K., Komori, O., Eguchi, S., Robust ranking method via repeated random partition for gene expressions data., The 3rd Institute of Mathematical Statistics Asia Pacific Rim Meeting , Taipei Taiwan, 2014.06.28

⑲ Eguchi, S., Komori, O., Generalized entropy and divergence in statistical learning., The 3rd Institute of Mathematical Statistics Area Pasific Rim Meeting, Taipei Taiwan, 2014.06.27, 招待講演

⑳Eguchi, S., “Possible generalization of MAXENT”, International symposium on innovation and challenges for fisheries assessment and management，2014 年 3 月 5 日，慶應義塾大学理工学部 矢上キャンパス

㉑江口真透，「情報幾何の展開–アダブーストからポアンカレ予想まで」，特別講演会，2013 年 11 月 15 日，千葉大学理学部

㉒江口真透，「2 値判別分析におけるモデルと推定の関係について」，科学研究費シンポジウム「一般化線形モデルの最新の展開とその周辺」，2013 年 11 月 9 日，千葉大学理学部

㉓Komori, O., Hung, H., Chen, P., Huang, Su-Yun and Eguchi, S., “A class of u-statistics combining multiple markers for two-group classification”, Joint Meeting of the IASC Satellite Conference for the 59th ISI WSC and the 8th Conference of the Asian Regional Section of the IASC, Aug.23, 2013, Seoul, Korea

〔図書〕（計７件）

①Eguchi, S., A. Notsu, O. Komori
Computational Information Geometry
(担当:分担執筆, 範囲:One chapter (79-99))
Springer International Publishing 2017 年

②Shinto Eguchi, Katsuhiro Omae., International Conference on Geometric Science of Information, Springer,Cham, 561-568., 2017,

③Eguchi, S., A. Notsu, O. Komori, Spontaneous Learning for Data Distributions via Minimum Divergence, Springer International Publishing 2017, 総ページ数 299

④ Eguchi, S., A. Notsu, O. Komori, Computational Information Geometry, Springer International Publishing, 2017, pp.79-99

⑤Osamu Komori and Shinto Eguchi., Statistical and Machine-Learning Methods for Class Prediction in High Dimension. Chapter 14 in Design and Analysis of Clinical Trials for Predictive Medicine. Eds S. Matsui, M. Buyse, R. Simon. Chapman & Hall/CRC Biostatistics Series 2015, pp. 253-270

⑥T Takenouchi, O Komori, S Eguchi. A novel boosting algorithm for multi-task learning based on the Itakuda-Saito divergence, 2014, Vol. 1641, pp. 230-237, AIP Publishing.

⑦S Eguchi, O Komori, A Ohara. Duality in a maximum generalized entropy model, 2014, Vol. 1641, pp. 297-304

〔産業財産権〕

○出願状況，取得状況 なし.

〔その他〕
ホームページ等

６．研究組織

(1) **研究代表者**

江口　真透（EGUCHI, Shinto）
統計数理研究所・数理・推論研究系・教授
研究者番号：１０１６８７７６

(2)**研究分担者**

松浦　正明（MATSUURA, Masaaki）
帝京大学・大学院・公衆衛生学研究科・教授
研究者番号：４０１７３７９４

松井　茂之（MATSUI, Shigeyuki）
名古屋大学・大学院・医学系研究科・教授
研究者番号：８０３０５８５４

小森　　理（KOMORI, Osamu）
福井大学・工学研究科・特命講師
研究者番号：６０５８６３７９

間野　修平（MANO, Shuhei）
統計数理研究所・数理・推論研究系・准教授
研究者番号：２０３７２９４８

野間　久史（NOMA, Hisashi）
統計数理研究所・データ科学研究系・准教授
研究者番号：７０６３３４８６

(3)**連携研究者**

竹之内　高志（TAKENOUCHI, Takashi）
公立はこだて未来大学・システム情報科学部・准教授
研究者番号：５０４０３３４０

逸見　昌之（HENMI, Masayuki）
統計数理研究所・データ科学研究系・准教授
研究者番号：８０４６５９２１

(4)**研究協力者**

John B. COPAS
University of Warwick, Department of Statistics, Emeritus Professor

Su-Yun Huang (**陳素雲**)
Institute of Statistical Science, Academia Sinica

HUNG HUNG (**洪弘**)
National Taiwan University, Graduate Institute of Epidemiology and Preventive Medicine, Associate Professor