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　　元 年　　５ 月 ２１ 日現在

機関番号：１３９０１
研究種目：挑戦的萌芽研究
研究期間：2016～2018
課題番号：１６Ｋ１２４０１

研究課題名（和文）多次元データ解析の新しい枠組み：構造推定と判別解析

研究課題名（英文）A new framework of discrimination analysis for high-dimensional data based on estimation of data structure

研究代表者
松井　茂之（Matsui, Shigeyuki）

名古屋大学・医学系研究科・教授

研究者番号：８０３０５８５４

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：生物学・医学領域におけるゲノムデータなどの高次元データを用いた興味ある形質の判別解析において、形質とゲノムデータの関連構造を捉えた階層混合モデルに基づいて判別式を構成し、モデルベーストに判別精度を推定するという新しい枠組みを検討した。併せて、がんなどの疾患にみられる分子レベルでの異質性の構造を捉えたネスト型混合モデリングとこれに基づく疾患判別法について検討した。

研究成果の学術的意義や社会的意義
ゲノムデータなどの多次元データを用いた提案する判別・予測解析は、ゲノムデータがもっている自然な関連構造、疾患の分子レベルでの異質性を明示的に考慮しており、統計・機械学習の新しい枠組みを提案するものである。一方で、本研究で開発した方法を適用することで、疾患の診断法の開発はもとより、疾患の分子機構の理解、新規治療法の分子標的の発見に役立つと期待できる。

研究成果の概要（英文）：We developed a novel framework of discrimination analysis of phenotype classes using high-dimensional genomic data in biomedical researches. This framework is based on hierarchical mixture models of the underlying structure on the association between the phonotype and genomic data and is expected to allow for stable discrimination and also for estimation of discrimination accuracy based on the model. We also considered incorporation of disease heterogeneity at the molecular level. One approach is the use of nested mixture models that can identify clusters of genes that are associated with the phonotype in particular subsets of disease patients. We applied the developed methods to real datasets from clinical genomic researches in cancer and other diseases.

研究分野：統計科学

キーワード：判別・予測解析　機械学習　疾患の異質性　統計モデリング

１．研究開始当初の背景

ゲノムデータなどの高次元データは，解析で用いる変量（例えば，遺伝子）の数がサンプルサイズよりもはるかに大きいことで特徴付けられる．高次元データの解析としてよく見られるものは，ある興味のある現象や形質，例えば，疾患発生の有無，疾患分類などに対して，関連のある遺伝子を検出したり，関連遺伝子を用いて判別・予測を行うことである．従来，前者に対しては多重検定，後者に対しては，統計的判別解析/機械学習の枠組みで多くの研究が行われてきた．しかし，多次元データのもとでは，前者は偽陽性，後者は過適合（overfitting）という根本的で深刻な問題を抱える．

一方，高次元データを全体としてみると往々にしてある構造をもっており，例えば，ゲノムの全遺伝子の内,形質と関連をもつ遺伝子とそうでない遺伝子に分けられ(混合構造),さらに，関連遺伝子に着目すると，一部の遺伝子（例えば，同じ pathway に含まれる遺伝子）は形質に対してほぼ同じ関連のパターン・大きさをもっており，関連遺伝子全体でみると，関連パターン・大きさはある分布を構成すると考えられる（階層構造）．しかしながら，上記の構造を明示的に捉えた判別分析の枠組みはこれまで十分に研究されていない．

２．研究の目的

ゲノムデータなどの高次元データを用いた形質変数の判別解析において，形質変数とゲノムデータの関連構造を捉えた階層混合モデルに基づいて判別式を構成し，モデルベーストに判別精度を推定するという新しい枠組みを検討する．併せて，がんなどの疾患にみられる分子レベルでの異質性の構造を捉えた統計モデリングとこれに基づく疾患判別法について検討する．

３．研究の方法

高次元データの構造推定のための階層モデルと経験ベイズ推定の妥当性・有効性を遺伝子発現データ/多型データ,脳画像データなどの様々な高次元データを想定して確認する.その上で，推定した関連構造に基づいた判別解析の方法を検討する．個々の遺伝子に関する効果サイズや分散に関する縮小推定量を用いた判別式の安定化，縮小推定量に基づいて遺伝子選択によるバイアスを除去した人工データの作製，さらに，これに判別アルゴリズムを適用することでの判別精度の改善，判別精度を人工データから直接推定したときの推定の性能についての検討を数値実験や実データへの適用を通して行う．

疾患の異質性を考慮した解析については，がんサンプルと正常サンプルの比較において，異なるがん関連プロファイルを想定した判別アルゴリズムの構築を行う．これは，ある関連プロファイルでは，がんサンプルの一部のみで遺伝子が高（低）発現し，別の関連プロファイルではまた別のがんサンプルのみが高（低）発現，という特殊な構造である．この構造は，遺伝子×がんサンプルの二方向（two-way）クラスタリングの形となり，混合モデルに基づく構造推定を行う．このとき，各関連プロファイルに属する事後確率を全関連プロファイルで統合することでがんの判別アルゴリズムを構築する.数値実験等により従来の判別法との性能比較を行う．併せて，疾患異質性を想定した様々な統計モデリングのアプローチと判別法の検討を行う．

なお，研究全体を通して，医学研究の実データを用いた適用研究も併せて行い，開発した判別解析法の有効性，有用性について検討する．

４．研究成果

（1）関連構造の推定に基づく判別解析

まず，本研究で提案する判別解析の枠組みの基礎となる階層混合モデルと経験ベイズ推定の評価を中心に行った．マイクロアレー遺伝子発現データ（正規的変量）と二つのクラスをもつ表現型変数の関連解析において，クラス間の平均，クラス内の分散を階層モデリングの対象として両者の縮小推定を行う方法を開発した．併せて，統計的有意性の高い一部のマーカーセットを用いた対角線形判別解析（diagonal linear discriminant analysis）を縮小推定量に基づいて構成することで，判別式の安定化を試みた．その上で，階層モデルに基づいて判別精度を推定する方法を検討した．以上の方法の実践として，がんの第二相臨床試験における薬剤奏功例の判別解析を想定したシミュレーション評価を行い，従来のクロスバリデーションに基づく判別解析と比べ（図 1），判別精度の推定精度が大きく改善する

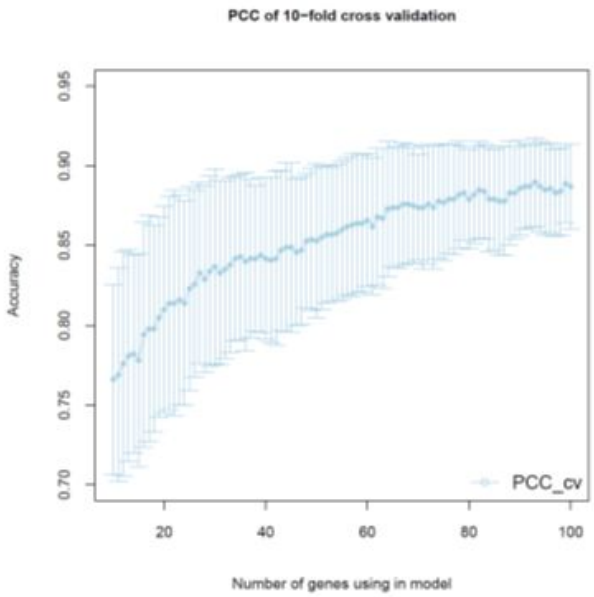

図 1. 従来のクロスバリデーションによる判別精度の推定．シミュレーション実験での 200 回の試行における判別に用いた遺伝子数（横軸）に対する正判別率の分布（縦軸）．

ことを確かめた（図 2）.

一方，別のアイデアとして，学習データに対して縮小推定量に基づく線形変換を施すことで，マーカー選択のバイアスを補正した人工データを生成し，これを用いて（テストサンプルを用いることなしに）判別精度を評価するという，これまでにない新しいアプローチについても検討した．シミュレーション実験の結果，判別精度の偏りのない推定のためには，クラス内分散の縮小推定量の使用が有効であることがわかった（特に小標本）．これより，分散の縮小推定量の改善を図った．逆ガンマ分布を事前分布に用いた方法を中心に検討し，小標本のもとで一定の改善を確認したものの推定が安定しないケースも多くみとめられた．一方で，ノンパラメトリックな事前分布も検討したが，計算負荷が大きくなり，ごく限られた条件下でのみしか性能を確認できなかった．

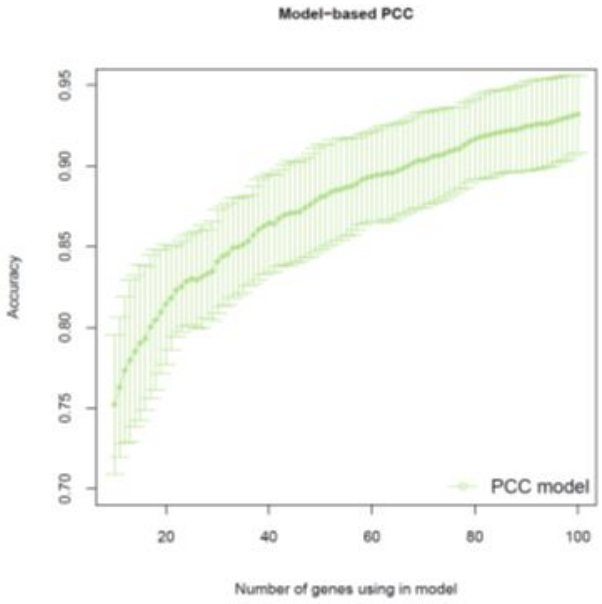

図 2. 提案するモデルベースの判別解析による判別精度の推定．シミュレーション実験での 200 回の試行における判別に用いた遺伝子数（横軸）に対する正判別率の分布（縦軸）．図 1 よりも正判別率の分布の幅が小さい．

なお，階層モデル解析については，遺伝子発現データ解析以外に，一塩基多型データ，脳画像データを用いた場合の拡張についても検討し，シミュレーション実験，及び，様々な疾患の実データの解析により，モデル推定法の妥当性を確認した．また，形質変数が生存時間である場合，複数の疾患・治療サブループがある場合の拡張についても検討した．

（2）疾患の異質性を考慮した解析

多くのがんなどに見られる分子レベルでの疾患異質性を考慮した関連解析として，ネスト化正規混合モデリングに基づくパラメトリック法を開発した．基本性能の評価を比較的な単純なデータ構造を想定したシミュレーション実験により行い，骨髄異形成症候群（MDS：myelodysplastic syndromes）のマイクロアレー遺伝子発現データへの適用を行った（図 3）.

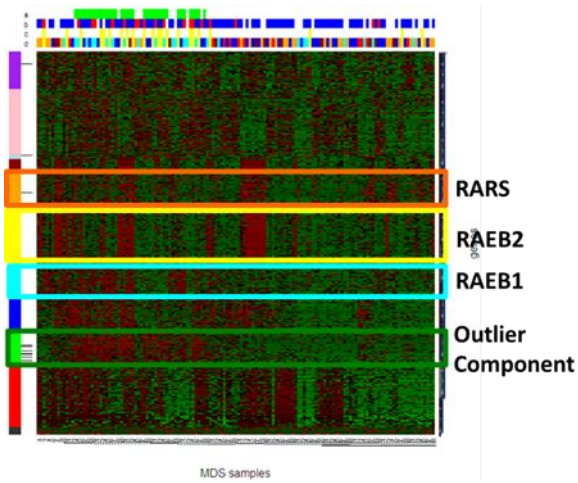

図 3. MDS のマイクロアレー遺伝子発現データへの適用. 行は遺伝子, 列は MDS サンプル. 長方形の枠は, ネスト型混合モデルにより検出されたがん関連遺伝子のクラスター. 各クラスターに対して全サンプルでなく, 一部のサンプルが高発現または低発現している.

併せて，推定モデルに基づいて，がんと健常人の判別解析法の開発を行った．これは，新たなサンプルの発現量データが与えられたとき，がん関連遺伝子の各コンポーネントに対してそのサンプルががんである事後確率を計算することで，判別を行う方法である．従来の判別法（Fisher の線形判別分析，サポートベクターマシンなど）との性能比較を行った結果，疾患の異質性の度合いが高い場合には提案法が優れるが，異質性の度合いが低い場合には従来法が優れるという結果となった．この結果を受けて，提案法の性能をさらに高めるために従来の判別法とのハイブリットな方法について検討中である．

別のアプローチとして，疑似スコアを用いた判別法，遺伝子別にがん固有の外れ値を健常人などのレファレンスサンプルの分布上での分位点として捉え，全遺伝子を通して分位点分布を構成することでがんの判別を行うノンパラメトリックな方法を開発した．シミュレーション実験等により，これらの方法についても上記のパラメトリック法と同様の性能を有していることがわかった．すなわち，疾患異質性の度合いが特に高い状況下で従来法よりも優れた性能を発揮するということである．

以上の方法の適用に関して，引き続き現在も米国の研究者との共同研究として多発性骨髄腫の患者コホートデータを用いた疾患関連解析を実施中である．

５．主な発表論文等

〔雑誌論文〕（計 8 件）

① Emura T, Matsui S, Chen HY. compound.Cox: univariate feature selection and compound covariate for predicting survival. Computer Methods and Programs in Biomedicine. 2019; 168: 21-37. 査読あり. DOI: 10.1016/j.cmpb.2018.10.020.

② Nishino J, Kochi Y, Shigemizu D, Kato M, Ikari K, Ochi H, Noma H, Matsui K, Morizono T, Boroevich K, Tsunoda T, Matsui S. Empirical Bayes estimation of semi-parametric hierarchical mixture models for unbiased characterization of polygenic disease architectures. Frontiers in Genetics 2018; 9: 115. 査読あり. DOI: 10.3389/fgene.2018.00115.

③ Baek S, Komori O, Ma Y. An optimal semiparametric method for two-group classification. Scandinavian Journal of Statistics. 2018; 45: 806-846. 査読あり. DOI: 10.1111/sjos.12323.

④ Matsui S, Noma H, Qu P, Sakai Y, Matsui K, Heuck C, Crowley J. Multi-subgroup gene screening using semi-parametric hierarchical mixture models and the optimal discovery procedure: Application to a randomized clinical trial in multiple myeloma. Biometrics. 2018; 74: 313-320. 査読あり. DOI: 10.1111/biom.12716. 6.

⑤ Omae K, Komori O, Eguchi S. Quasi-linear score for capturing heterogeneous structure in biomarkers. BMC Bioinformatics. 2017: 18: 1-15. 査読あり. DOI: 10.1186/s12859-017-1721-x.

⑥ Komori O, Eguchi S, Saigusa Y, Okamura S, Ichinokawa M. Robust bias correction model for estimation of global trend in marine populations. Ecosphere. 2017; 8: 1-9. 査読あり. DOI: 10.1002/ecs2.2038.

⑦ Nishikimi M, Matsuda N, Matsui K, Takahashi K, Ejima T, Liu K, Ogura T, Higashi M, Umino H, Makishi G, Numaguchi A, Matsushima S, Tokuyama H, Nakamura M, Matsui S. CAST: a new score for early prediction of neurological outcomes after cardiac arrest before therapeutic hypothermia with high accuracy. Intensive Care Med. 2016; 42: 2106-2107. 査読あり. DOI: 10.1007/s00134-016-4492-3.

⑧ Omae K, Komori O, Eguchi S. Reproducible detection of disease-associated markers from gene expression data. BMC Medical Genomics. 2016: 9: 53. 査読あり. DOI: 10.1186/s12920-016-0214-5.

〔学会発表〕（計 8 件）

① 江村剛志, 松井 茂之, Hsuan-Yu Chen. 単変量 Cox 回帰にもとづく遺伝子選択と複合共変量による生存期間の予測. 2018 年度統計関連学会連合大会, 2018/9/10-13, 東京都.

② 小森 理, 江口真透. 一般化エネルギー関数に基づくクラスター分析. 2018 年度統計関連学会連合大会, 2018/9/10-13, 東京都.

③ Ryo Emoto, Atsushi Kawaguchi, Hisako Yoshida, Shigeyuki Matsui. Hierarchical Mixture Modeling for Multiple Testing and Effect Size Estimation in Voxel-Level Inference of Neuroimaging Data. Joint Statistical Meeting 2018, July 28- August 2, 2018, Vancouver, Canada.

④ Ryo Emoto, Atsushi Kawaguchi, Hisako Yoshida, Shigeyuki Matsui. Multiple Testing Based on Semi-Parametric Hierarchical Mixture Models under Dependency in Disease-Association Studies with Neuroimaging Data. Eastern North American Region International Biometric Society 2018 Meeting, March 25-28, 2018, Atlanta, USA.

⑤ 松井茂之. Omics 研究における検証的解析と探索的解析. 2017 年度統計関連学会連合大会日本計量生物学会シンポジウム「医学・農学研究における p 値 - $p < 0.05$ を超えて-」. 2017 年 9 月 4 日, 名古屋（招待）.

⑥ 松井孝太, 大浦智則, 松井茂之. 入れ子型混合モデルに基づく cancer outlier profile の推定とがん診断への応用について. 科研費シンポジウム「統計的モデリングと計算アルゴリズムの数理と展開」, 2017 年 2 月 1 日, 名古屋.

⑦ Osamu Komori. Asymmetric logistic regression model. The International Conference on Bioinformatics and
Biostatistics for Agriculture Health and Environment, Jan. 22, 2017, Rajshahi, Bangladesh.

⑧ Tomonori Oura, Kota Matsui, Shigeyuki Matsui. Cancer Outlier Analysis Based on a Nested Two-Way Clustering. XXVIIIth International Biometric Conference, July 10-15, 2016, Victoria, Canada.

〔図書〕（計 0 件）

〔産業財産権〕

出願状況（計０件）
取得状況（計０件）

〔その他〕
ホームページ等：
http://www.nagoya-biostat.jp/

６．研究組織
(1)研究分担者
小森　理（KOMORI, Osamu）
成蹊大学・理工学部・准教授
研究者番号： 60586379

(2)研究協力者
John Crowley
Board Chair, Chief of Strategic Alliances
Cancer Research And Biostatistics