様式 C-19

科学研究費補助金研究成果報告書

平成　21　年　5　月　24　日現在

研究種目：若手研究(B)
研究期間：2007～2008
課題番号：19710172
研究課題名（和文）高次元オミックスデータとヘテロ医療診断データの統合的解析法の開発
研究課題名（英文）Development of integrated analysis methods for high-dimensional and heterogeneous clinical data
研究代表者
大羽　成征（OBA SHIGEYUKI）
京都大学・大学院情報学研究科・講師
研究者番号：80362838

研究成果の概要：
近年の臨床医学データ解析では、解析対象となる変量に高次元性・ヘテロ性・欠測などの問題があること、解析結果出力時に統計的検定多重性の問題があることが困難を生んでいる。これらの問題の統合的解決法として、柔軟な確率モデルと統計的多重検定法の発展応用に基づく手法を開発した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
１９年度	1,700,000	0	1,700,000
２０年度	1,400,000	420,000	1,820,000
年度			
年度			
年度			
総　計	3,100,000	420,000	3,520,000

研究分野：複合新領域
科研費の分科・細目：ゲノム科学・応用ゲノム科学
キーワード：バイオインフォマティクス，遺伝子発現量データ，多変量解析，多重検定，行列因子化

１．研究開始当初の背景

近年の臨床医学関連データ解析では様々な臨床診断情報に加えて遺伝子発現量や SNP などの高次元変量オミックスデータを複数同時に扱う必要があるが、このとき（1）各症例の特徴を示す変量の高次元性・ヘテロ性・欠測などの問題や、（2）解析結果出力時の統計的検定多重性の問題への対応が難しい。このため、言いたいことと言えることの間のギャップに足をとられるなどで、統計的処理が研究プロセスのうえでのボトルネックとなることが当該分野では多かった。機械学習分野では、構造利用学習が活発な研究領域となりつつあるが、理想化状況での理論研究が主であり、ここで想定するような医学・生物学の実問題における高次元ヘテロデータに伴う諸問題を同時にとり扱うことのできるような一つの手法はこれまでに存在しなかった。

２．研究の目的

本研究課題では最新の統計的学習・機械学

習の知見を結集することによって高次元ヘテロデータ解析における実際上の緒問題の解決法を統合しパッケージ化して供給し、現状の打開を目指すことを目的とした。具体的には、（１）の問題に対しては、適切な階層化を施したモデルを工夫することによってモデルの複雑さのコントロールを行う正則化方法の開発を行い、（２）の問題に対しては階層的モデルに基づく統計的検定手法の開発を行い、データに基づく手法の信頼性の向上と信頼性見積もりができるようにすることを目指す。またこれらの手法をエンドユーザーにとって扱いやすいインターフェイスを備えた形で公開することも目的のひとつとした。

３．研究の方法

解析の対象となるデータが一見複雑な構造をとるように見えたとしても、欠測を含むひとつの行列と見れば統一的にまとめることができる。そこで、行列データの確率モデルとしてベイズ的行列因子化モデルを基盤として、特殊な問題に発展してゆく戦略をとることとした。とくに欠測構造の特殊性の取り扱いかたに注意してモデリングを行った。

多重性については、J. Storey らが 2003 年に発表した多重性のもとでの最適検定の理論(Optimal Discovery Procedure; ODP)のの応用を考えた。

４．研究成果

本研究課題で得られた成果は大きく分けて、(1) 行列因子化法の改善、(2) 検定多重性のもとでの検出力最大化、およびこれらの組み合わせによる (3) 行列因子化を前提とした検定における検出力向上、の三つにわけられる。以下この順に記述する。

(1) 行列データの因子化法

観測 j=1,...,N における、変量 i=1,...,M の観測値を Yij とし、これをまとめたものを観測行列 **Y**={Yij} と呼ぶ。遺伝子発現量データでは Yij は i 番目の遺伝子が j 番目の標本において示した発現量（実数値）を表す。アレイ比較ゲノム（aCGH）法によって計測された DNA 異常データでは i 番目の DNA セグメントの増幅／欠損の度合いを実数値で表す。評点データでは、Yij は評価者 j が i 番目の対象に対してつけた 6 段階評点データを表す。どのデータにおいても、Yij が欠測している場合を含む。このことを欠測パターン行列 **W**={Wij} で表し、Wij=0 ならば Yij は欠測しており、Wij=1 ならば Yij は観測されているものとする。

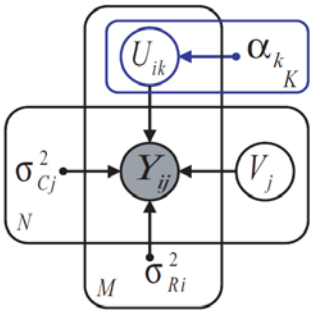
この観測行列 **Y** を、因子化表現された近似行列 **X**=**UV**' によって近似することを、行列因子化と呼ぶ。**U**, **V** はそれぞれ N×K 行列、M×K 行列であり、K を因子数と呼ぶ。

適切に推定された **X** は、観測行列 **Y** の欠測部分の予測に用いることができる。またこのとき因子行列 **U**,**V** の各行ベクトルは各変量 i 、各観測 j の特徴を K 次元で表現する特徴ベクトルの意味を持ち、データの理解に役立つ。

本研究課題では、(A) とくに遺伝子発現量解析のために、各成分と各症例の信頼性の違いが同時に存在する場合を考慮して重み付けを工夫した因子分解法「縦横二方向因子分析 (B) デジタル値評点データに対して行列因子分解を行う「重みつきマージン最大化行列分解」、「因子数推定つきマージン最大化行列分解」(C) 遺伝子発現量と aCGH データの同時解析で見られるような、欠測の構造にヘテロ性が入っているデータに対してスパースな因子負荷行列を求める「確率的ヘテロ成分分析」を開発した。

（A） 縦横二方向因子分析法の開発

マイクロアレイによる遺伝子発現量観測データの品質は、スポット毎および観測毎に異なる場合がある。このような状況を想定して以下のようなモデルを提案した。



上図は縦横二方向因子分析のモデルを図解したものである。Y, U, V の間の関係は既に述べたとおりであるが。症例毎に異なる誤差分散パラメタ σ^2_{Cj} ,j=1,...,N, 遺伝子毎に異なる誤差分散パラメタ σ^2_{Ri} ,i=1,...,M、因子毎に異なるスケールパラメタ α_k ,k=1,...,K を含み、これら全てを同時に推定対象としているところが工夫点である。このモデルは既存の因子分析モデルを特別の場合として含む。（$\sigma^2_{Cj}, \sigma^2_{Ri}$ のいずれかを定数とすれば既知の因子分析と同様のモデルである。）また α_k を推定対象とすることによって、関連次元自動決定（ARD）の効果を得る。

想定された状況を模したシミュレーションでは欠測予測性能において既存手法を越える精度が得られた。一方で公開された実データ上では大きな精度向上は得られなかった。これは公開データでは品質の低い観測があらかじめ排除されていることによるものと思われた。本手法の応用が有効と考えられ

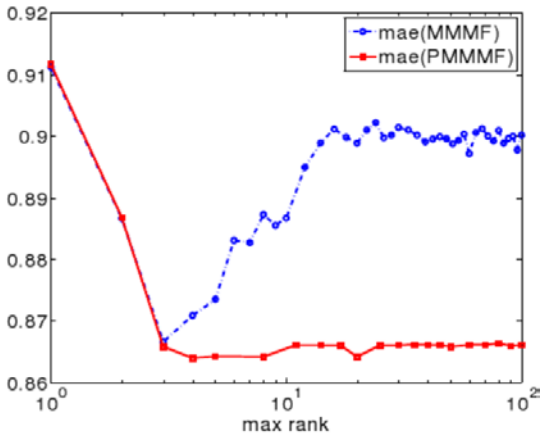
る状況は、マイクロアレイデータセットに品質の低い観測が含まれる場合である。本手法によれば、特定の遺伝子・特定の観測の品質が絶望的に低いならばそれを実質的に排除し、一定の情報が含まれるならばその情報を適切に引き出すことで、悪影響を受けずに情報を統合できる。

(B) マージン最大化行列因子化法の改善

デジタル値評点データに対して、マージン最大化行列因子化法という既存手法がある。これは、デジタル観測値 Y と実数値近似 X との間を「ヒンジ誤差」と呼ばれる特殊な誤差関数で結び、これを最小化する因子化近似行列を求めることで、デジタル値推定誤りを防ぐための安全マージンを最大化するという工夫である。

これに対して、2 点の改善を行った。第一は、観測行列の一部に着目して重みをかけることで全体的な予測精度を制御する重み付き因子化。第二は、近似行列の因子数を自動決定する ARD の仕組みの導入である。

デジタル値評点データの実例として公開されている映画評点データセットを用い、改善手法と従来手法との比較実験をおこなった。結果の一部を以下の図に示す。



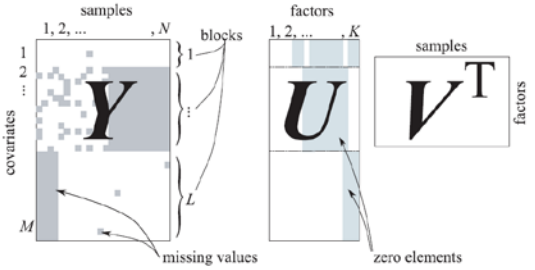
横軸は因子数 K を、縦軸は予測性能を mean absolute error で表したものである。MMMF は従来手法、PMMMF は改善手法である。PMMMF では因子数 K を大きめにとったとしても最良性能と同等の性能が常に得られているが、これは因子数自動決定のおかげで不要な因子の影響が自動的に減殺されるからである。

今後この結果を臨床医学関連のヘテロデータへも応用してゆく予定である。

(C) ヘテロ成分分析

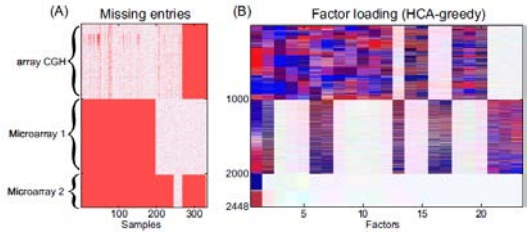
行列形の観測データ **Y** がヘテロ性を持つ場合の行列因子化法として、ヘテロ成分分析という名前の手法を開発した。ここで考えるヘテロ性とは、変量 i=1,…,M が複数のブロックに分かれており、ブロックごとに運ばれる情報の実効次元が異なり、ノイズレベルが異なり、欠測率が異なる状況である。とくに欠測については、ブロックごとまとめて欠測するような状況がときどき起こるような状況を想定する。このような状況は、遺伝子発現量マイクロアレイデータとアレイ CGH による DNA 変異データを同時に扱う場合にしばしば出会うものである。

こうしたヘテロデータに対応するために、因子行列 **U** がブロック型スパース性を持つという制限をもつ行列因子化法を提案した。



上図では、ヘテロ的な欠測を含む観測行列 **Y** を、ブロック型スパース性を持つ因子行列 **U** を用いて表現するさまをイラストで示した。

提案手法で、神経芽腫に関するヘテロデータを解析したところ、既存手法よりも少ない非ゼロ成分を持つ因子行列によって既存手法以上の欠測予測性能を得ることができた。また得られたブロック型スパース因子行列によってヘテロデータが運ぶ実効的情報の次元を表現することができた。



上図の左は、対象とした神経芽腫関連データであり、3 ブロックからなる。赤で示した要素は欠測を示しており、ブロック型の欠測がみられる。右はこれをヘテロ成分分析にかけた結果として得られたブロック型スパース性を持つ因子行列であり、白いところがゼロ値、赤と青がゼロでない実数値をあらわす。ブロックごとに実効次元が大きく異なることが分かる。

(2) 多重検定における検定力最大化

行列の形で得られたデータ **Y**={Yij} に基づいて、有意遺伝子検出を行う問題を考える。行列の各列 j=1,…,N が症例に対応しており、その一部が予後の悪い癌、残りが予後の良い癌だとしたとき、これらの症例群の間で平均発現量に違いのある遺伝子を DEG(differentially expressed genes) と呼ぶ。しかし観測データは有限であり、ノイズ

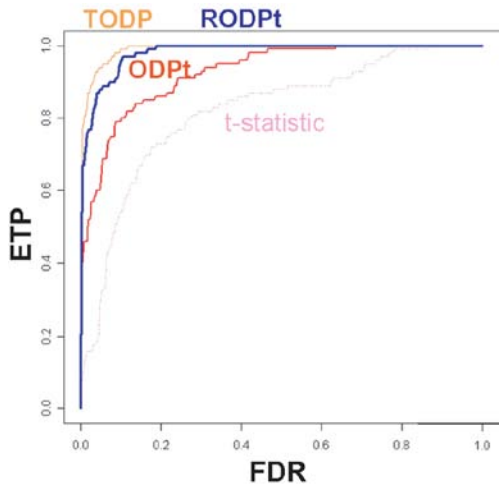
も含まれているため、観測に基づいて DEG を検出するときには、一定の偽陽性と偽陰性がどうしても含まれざるを得ない。検定の目的の第一は、検出された DEG に含まれる偽陽性率を一定未満に抑える「保守性」、そして第二はなるべく多くの真の DEG を検出する「検出力」である。この二つの目的は一般には矛盾するが、保守性を確保しつつ、検出力を最大化することを目的とするとき検定法の改善の余地が残っている。とくに、複数の検定を同時に行う多重検定では、多重性を利用することによって検出力を高める工夫の余地が大きい。

鍵になるのは 2006 年に J. Storey によって提案された統計的有意性スコア最適化原理（ODP）の理論である。DEG 検出は、第一に DEG 候補となる遺伝子に有意性スコア（しきい値関数と呼ぶ）をつけ、第二に適切なしきい値を超えた遺伝子を検出するという過程からなる。ODP 理論は最適なしきい値関数の理論的根拠を与える。しかし、最適なしきい値関数はモデルパラメタの多くを既知とした理想的な状況で定義されるものであり、実際の応用上では未知のパラメタを観測データにもとづいて近似的に推定して求める必要がある。そして既存の推定方法には大きな改善の余地があると思われた。

本課題では、ODP 理論の意味で最適なしきい値関数を観測データに基づいて定義するために、二つの方向での研究を行った。第一は、しきい値関数の逐次的改善法、第二は多次元局所 FDR 法である。

（A）しきい値関数の逐次的改善法

まず、しきい値関数の逐次的改善法とは、古典的なしきい値関数（t-統計量）に基づいて得られた検定結果を初期値として、逐次的に真の最適値に近づけてゆくことで推定の改善を図る方法である。



上図にシミュレーション結果を示す。横軸は偽陽性率（FDR）、縦軸は期待真陽性率(ETP)を示し、しきい値関数毎にさまざまなしきい値を適用することによってひとつの曲線が描かれる。曲線が左上に寄るほど良いしきい値関数である。最も左上に寄っている TODP（理想的 ODP）が最も良いが、これは理想的状況で得られるしきい値関数であり、実際には得られない。古典的な t 統計量、既存手法による推定値 ODPt と比べ、提案手法である RODPt は顕著に精度が良く TODP に近い性能を示した。

（B）多次元局所 FDR 法

多次元局所 FDR 法は、ノンパラメトリックベイズ的手法に基づく検定法の応用である。手法としては、2001 年 Efron らによって提案された既存の手法であるが、本課題のなかで研究を進めたところ、これが ODP 理論の意味で最適なしきい値関数となり得ること、および、そのための厳密な条件が明らかになった。この結果は学会等では未発表であり、現在論文を投稿中である。

(3) 行列因子化を前提とした検定

(1),(2)の結果を使用することで、行列因子化モデルを前提とした多重検定法を開発した。この手法は、１９年度に得られた２つの成果（１）行列因子化の確率的モデルをヘテロデータに拡張したヘテロ成分分析モデル（ＨＣＡ）および（２）隠れ変数モデルを前提とした多重検定手法（ＨＯＤＰ）の統合手法であり、本研究課題で当初目標としていたアイディアの骨子を全て実装した手法と位置づけられる。本手法によれば、ヘテロ性や欠測を含む行列であっても全て一段階抽象化された行列成分で代表され、遺伝子有意性検定などの知識抽出処理は抽象化された世界で行うことができる。この研究は「ニューラルコンピューティング研究会」ほかで発表され IEEE Computational Intelligence Society Japan より 2008 年度 Young Researchers Award を受賞した。

ここではこの手法の骨子のみを説明しておく。行列 **Y** の因子化によって得られる因子行列 **U** は、因子数 K のもとで各変量 i の性質を示す十分統計量となっている。そこで、因子行列の成分を統計量とした多次元局所 FDR を求めれば、これは因子数 K のもとで最適なしきい値関数となる。しかし、2 次元を越える次元での局所 FDR の計算は通常のノンパラメトリック分布推定では精度が悪くなる。そこで、因子間の条件付き独立性を仮定したもとで 2 次元ずつの局所 FDR を求めたうえでそれらを統合して、最終的なしきい値関数を求める工夫を行った。

多くのシミュレーションデータおよび、実データで既存手法との比較を行ったところ保守性を損なうことなく検出力の向上が経験的には確かめられた。しかし保守性の理論的保証はまだ十分に確かめ切れていない。この困難の解決は課題の期間内にはできなか

ったが、これを解決したうえでの論文投稿を予定している。

最後に特記事項として、本研究課題から派生した研究テーマがＪＳＴさきがけで採択され、今後とも発展的に継続の予定であることを記しておく。

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０件）

〔学会発表〕（計 10 件）
① Nomura, H., Oba, S., and Ishii, S., Re-weighted ODP for differential gene expression analysis, International Symposium on Artificial Life and Robotics (AROB 13th '08), GS5-2, 2008, 査読有り, Oita Japan
② Furuya, M., Oba, S., and Ishii, S. Collaborative filtering based on a Weighted Maximum Margin Matrix Factorization. International Symposium on Artificial Life and Robotics (AROB 13th '08), GS14-5, 2008, 査読有り, Oita Japan
③ Oba, S., Kawanabe, M., Mueller, K-R., and Ishii, S. Heterogeneous Component Analysis. In proceedings of 21st annual conference on Neural Information Processing Systems. 2007、査読有り, Vancouver Canada
④ 大羽成征、石井信、検定多重性とサンプル個性を利用した臨床ラベル関連遺伝子探索、電子情報通信学会ニューロコンピューティング研究会、2008、査読無し、琉球大学
⑤ 古谷允宏、大羽成征、石井信、最大マージン行列因子化法の確率モデル化と行列データの欠測予測、第 11 回情報論的学習理論ワークショップ、2008、査読有り、仙台国際センター
⑥ 大羽成征, 石井信、典型パターン因子を統計量として用いた経験ベイズ検定による差異発現遺伝子検出法、第 11 回情報論的学習理論ワークショップ、2008、査読有り、仙台国際センター
⑦ 袖林和宏、大羽成征、石井信、二方向因子分析による行列データの欠損予測、情報論的学習理論ワークショップ、2007、査読有り、東京工業大学
⑧ 大羽成征、石井信、隠れ変数モデルに基づく同時検定用統計量最適化、情報論的学習理論ワークショップ、2007、査読有り、東京工業大学
⑨ 大羽成征、石井信、多重検定のための検定統計量最適化：隠れ変数を仲立ちとした情報共有に基く新しいアプローチについて、統計関連学会連合大会、2007、査読無し、神戸大学
⑩ 袖林和広、大羽成征、石井信、二方向因子分析による欠測データ予測、統計関連学会連合大会、2007、査読無し、神戸大学

〔図書〕（計０　件）
〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://hawaii.sys.i.kyoto-u.ac.jp/~oba/

６．研究組織
(1)研究代表者
大羽　成征（OBA SHIGEYUKI）
京都大学・大学院情報学研究科・講師
研究者番号：80362838

(2)研究分担者
（　　）
研究者番号：

(3)連携研究者
（　　）

研究者番号：