

# 判別分析の新理論と遺伝子解析のための新手法2

— Matroska Feature Selection Method for Microarray Data (新手法2)の解説 —

新 村 秀 一

## 1. はじめに

本稿では、筆者の一生の研究テーマ（新村，2012）であり整数計画法（IP）を用いて1997年から行ってきた「新しい判別分析の理論」を3章で、2015年末に幸運の女神がほほ笑んでめぐり合い、僅か41日で15編のフリーペーパーで確立した「Matroska Feature Selection Method for Microarray Data（新手法2）」を4章で概略を紹介する。しかし、この理論は線形分離可能なデータ（Linearly Separable Data, LSD）の判別分析が分からないと理解しにくい。そこでスイス銀行紙幣データ[6]と日本車44車種のデータ[52]の判別結果を例に、遺伝子解析の新手法2[34-48]を5章と6章で説明し、2章で研究の経緯を紹介する。2015年末にいつ終わるか分からないと思っていた一生の研究テーマが「小標本のための100重交差検証法（新手法1）」で判別係数の95%信頼区間（CI）の解釈がうまくいってほぼ終了した[30,31,33,51]。しかし、世界にそれを認めさすにはインパクトに弱かった。2015年10月末に6種類のMicroarray Data[11]を分析して、自分の研究が遺伝子解析のために最適なことが分かった[49]。筆者の開発した改定IP-OLDFだけが[19-26]、数千から数万の遺伝子情報を変数とする判別分析で、数十個の判別係数だけが0でなく残り全てが0になる。変数選択（遺伝子解析ではFeature Selectionと言っている）をしないで判別分析するだけで変数選択が自然に行える唯一の判別手法を提案したことになる。そして、LSDの判別では筆者の開発した最小誤分類数（Minimum Number of Misclassification, Minimum NM, MNM）が0になるが、その中に幾つものMNM=0になる部分空間を入れ子状に含むMatroska構造をもっていることが分かった。さらにMicroarray Dataは、数十個のMatroskaとMNM $\geq$ 1以上の多次元部分空間の排他的な和集合という思っても見なかった構造であることが分かった。十年以上に渡り世界中の多くの統計家が多次元データ解析と銘打ち研究してきたが、大きな成果を得られなかった。しかし、筆者が見つけた個々のMatroskaは、ほぼ100件\*100個以内の変数という小標本であり、簡単に統計分析できる。またMatroskaに含まれる一番小さなMatroska（Basic Gene Set, BGS）が分かればMicroarray DataのMatroska構造が完全に記述できる。すなわち、将来癌の遺伝子を修復する技術が確立されれば、これらのBGSを直接修復すればよいであろう。すぐにできる応用研究は、例えば胃がんという特定のがん患者と正常者の判別を行いがん遺伝子を特定する。そして、手術を除く抗がん剤や放射線療法などで5年以上延命した患者と正常者の判別から癌遺

伝子を特定する。治療が効果的であれば、必ず前者のがん遺伝子が後者において修正されているはずである。すなわち、癌の治療効果の客観的な評価ができると考えている。そこで、論文や国際会議で発表するには時間や研究費がかかるので、英文で本を出版し、世界にそれを問うことにした[50]。2015年まで一般的なデータで、改定IP-OLDFの検証標本の平均誤分類確率(M2)が他の判別関数より優れていることを示してきたがインパクトに弱かった。しかし、遺伝子解析は今のところ改定IP-OLDFでしか分析できない。数年後に判別分析の新理論と2個の新手法が世界的に認められればと願っている。

## 2. 判別分析研究の経緯

1971年に大学を卒業し、大阪府立成人病センターとNECの共同プロジェクトの「心電図自動診断解析システム」の一員として社会人のスタートを切った。プロジェクトリーダーの故野村裕医師から、東大医学部の疫学研究で著名な高橋氏編著の『計量診断学(東大出版会)』を読んでおくようにと渡された。暫くして「理解できましたか?ここに約3000人の32個の異常心電図所見群と正常所見群の入ったデータがあります」といって大きなMTを渡された。「このデータを用いて統計手法を使って、異常所見と正常所見を判別する診断論理を作成してください。既存の診断論理は、これらの医学書を読んでください」といって数冊の医学書を渡された。水泳をやりすぎて大学院に落ちて研究者の道を諦めていたが、数学ではないが医療工学(Medical Engineering)で研究者人生のスタートを切った。4年間研究を行ったが、野村医師の開発した「枝分かれ論理」に歯が立たなかった。当初は自分の能力がないためと考えたが、暫くしてFisher[4-5]が統計的判別分析の前提とした「Fisherの仮説」が、医学診断に適していないと考えるようになった。すなわちFisherの仮説は正常群と異常群は平均だけが異なる同じ正規分布と仮定しているが、次のように問題があると考えた。

- 1) 正常から異常はある計測値が連続的に大きく(あるいは小さく)なることで異常になる。
- 2) 異常所見の典型例は異常群の平均でなく正常から一番離れた症例である。そこで、正常群は地球、異常群は水平線から上に突き出た山脈と考える「地球モデル」を考えたら恐れ多くて発表の機会がなく、数年後にOR誌の編集委員を長らくやっていて医療特集号を組む際に、編集長の慶応大学理工学部の柳井氏の許可をもらいやっと発表の機会を得た(新村, 1984)。

一方判別手法の問題点を解消すべく、筆者の考えを「ベイズの定理」を用いて、ある計測値が連続的に大きく(あるいは小さく)なることで異常群に属する確率が0から1になる「スペクトル診断(新村他, 1973; 新村他1974)」を日本ME学会で発表した。余談になるが定年を控えた2013年末にインターネットで自分の研究が検索できるか否かを調べたら、この一番古い予稿が検索できた。研究発表を大切にしたい学会で発表することが重要であることに

初めて気づいた。しかし、30歳を過ぎて東京に移り、東大医学部で開催されていた東大医学部の伝説の秀才の一人である故開原氏の主宰する研究会に参加させてもらった。ここで、米国のフラミングハム研究でロジスティック回帰が用いられ、医学データの判別で効果を得ていることを知った。暫くして、「スペクトル診断」よりもロジスティック回帰の方がより洗練されていて、地球モデルに適していると判断した。今日では、日本の医学界では医学診断ではロジスティック回帰が用いられている。また、品質管理の故田口氏は、兼ねてより「正規分布を前提とした統計手法に批判的である」ことを知った。そして田口理論[57]では正常状態の正規分布から異常状態までのマハラノビスの汎距離を用いて、この距離が大きいほど異常と提案していることを知った。しかし、FisherのLDFと同じく分散共分散行列を用いているという点では、Fisherの手の内にあるとも考えられる。Fisherはコンピューターなどの恵まれた環境にない時代に明晰な頭脳で、「もし現象が正規分布で表されているとすれば」という前提で、我々に大きな研究分野を開拓してくれたと考えるべきである。計算機環境が整い、便利なソフトが利用できる時代になっても、現実に目をつぶり正規分布を拠る所に研究するのが可笑しいのである。筆者の研究では、Fisherが評価に用いたIrisデータ[3, 53]だけが、かろうじて数理計画法(MP)による線形判別関数(LDF)やロジスティック回帰に比べて見劣りがしないだけである。多くのデータの検証で、改定IP-OLDF(や改定IPLP-OLDF[29])、ロジスティック回帰とSVM4(ソフトマージン最大化SVMでpenalty  $c=10000$ )、SVM1 (penalty  $c=1$ )、FisherのLDFの順に判別成績が悪くなる。またFisherの仮説を検証する良い統計量はなく「NMがMNMに収束するときにデータはFisherの仮説を満たしていると考えられるしか方策がない」のが現状である。また開原氏の研究会で、SPSSの普及で有名な三宅一郎先生と間違い、日本医科大学の三宅章彦氏に声をかけ、判別関数の誤分類確率の研究や[13]、ヒューリスティックなOLDFの研究[14, 18](三宅, 新村, 1980)や、日本医科大学産婦人科のCPDデータの解析や、丸山ワクチンの解析などの機会を得た。その後、書籍で統計手法を習得する限界に目覚め、28歳の頃SASに巡り合い、日本に紹介を兼ね自分の統計の先生とした。その後、シカゴ大学ビジネススクールにLinus Schrage教授を訪ね、会話型数理計画法ソフトLINDOの代理店になり、数理計画法の勉強の助けとした(新村, 2011b)。企業人の時代、これらのソフトウェアを駆使し、まず解説書を書いて、ソフトウェアを体系的に理解した。その上で、多くの実証研究を行い論文を発表した。また企業にこれらを販売し、問題解決に当たってきたことが研究の助けになったようだ。

1996年に幸いにも成蹊大学に職を得た。数年前から厚生省の「介護保険システム」の開発に携わっていた富山中部高校の1年先輩の土肥医師の相談に載っていた関係から、着任早々厚生省の課長から厚生省の開原先生が委員長委員会の委員の依頼を大学のロビーで受けた。厚生省の委員になれば、そのうち潤沢な研究費ももらえることは知っていたが断ったの

は、じっくりと何かまとまった研究をしたかったからと土肥医師が外されていたからである。その後、国内の学会に年4回以上、国際会議に2回以上発表するようになり、大学の研究費でカバーできない部分を個人負担で行うことになり多少後悔した。企業人の時代、東大医療工学の古川教授、京大の産業オートメーションの大家で三宅教授との共同研究のヒューリスティックな最適線形判別関数 (Optimal Linear Discriminant Function, OLFDF) の発表で懇意になった桑原道義教授、パデュュー大学のK.S.Fu教授やAHPのSaaty教授などと知り合い、彼らの処で博士号を取る誘いや考えもあったが、なぜかしっくりしなかった。岡山大学の垂水氏から、計算機統計学会で会うたびに博士号を取る意思がないか誘いがあったが、東北大学の計算機統計学会開催時に夜の国分町で会い再度勧められ「考えてみます」と答えた。2-3日研究テーマを考えていて、これまで統計と数理計画法は筆者にとって別々の存在であり、いつか融合できないかと思っていた。そして、ヒューリスティック OLFDF でアプローチして途中で中断していた研究がIPで簡単に定式化できることを思いつき、論文博士号のテーマとした。IP-OLFDFでFisherのアイリスデータとCPDデータ(新村, 1996)<sup>1</sup>を用いたIP-OLFDFの研究(新村, 1998)と検証標本がないので乱数で115組の教師(内部)標本と検証(外部)標本を作製して検証した論文(新村 & 垂水, 2000)の2編で学位を得た[20]。その後、統計のテキストに用いていたTinyな「学生データ」を用いると、IP-OLFDFが「データが一般位置にない場合、正しい最適凸体 (Optimal Convex Polyhedron, OCP) の頂点を求めない」ことが分かった。また、スイス銀行1000フラン紙幣の真札と偽札の6個の計測値で、以下のことが分かった。

- 1) 6変数のスイス銀行データでは、63個の判別モデルが考えられる[9]。2変数 (X4, X6) のモデルで線形分離可能すなわち  $MNM=0$  であり、(X4, X6) を含む16個のモデルが線形分離可能である。そして、残り47個のモデルは線形分離可能でない。この時スイス銀行データはLSDであり、Matroska構造を持つことまで考えなかった。また線形分離可能な最小モデル (X4, X6) を、新理論2では基本遺伝子集合 (Basic Gene Set, BGS) と呼んでいるが、BGSが分かればLSDのMatroska構造が完全に把握できる。これが新理論2の骨子である。
- 2) MNMは単調減少性 ( $MNM_k \geq MNM_{(k+1)}$ ) があり、 $MNM_k=0$  であればk個の変数を含む全てのモデルは  $MNM=0$  であることを発見した。すなわち、一番小さな  $MNM_k=0$  になるモデル (BGS) を見つければ、全ての  $MNM=0$  になるモデル (Matroska) が分かる。

これらに加えて、日本の学術論文で「判別分析は重回帰分析と異なり、推測統計手法でない(問題4)」といえ一蹴に付される危険があるので、それを隠して「小標本のための100重交差検証法[27]」という新手法1で、学習標本と検証標本の平均誤分類確率M1とM2で

<sup>1</sup> 本研究は、多重共線性のあるデータから3つの共線性を見つける方法と、その影響を取り除く実証研究であり意義がある。

モデル選択を行い、判別係数の95%の信頼区間を加えて『最適線形判別関数(新村, 2010a)』を出版し、3章で取り上げる問題1, 問題2と問題4を不完全に解決し基礎研究を終えた。しかし、ほとんど手ごたえがなかった。そこで、LSDの判別分析が行われていないこと(問題2)に注目し、試験の合否判定をテーマに取り上げた(新村, 2011a)。ここで、医学診断の再開とくに遺伝子情報の解析も頭にあったが、分析しやすく教育にも貢献できる合否判定を選んだために、2点の問題で2年から3年の無駄な研究を行った。そのうちの 하나가、2次判別関数(QDF)と正則化判別分析(Regularized Discriminant Analysis, RDA [7])で、数学の合格群が全て不合格群に誤判別されるという問題3である。これは「何か今まで知られていない試験の設問方法の違いで、考えられない特殊なデータ構造が問題をひき起こしているのでは?」という間違った予見をし、多変量的な検討を2年間行い、2013年の計算機統計学会で敗北宣言を行った。数日後の深夜に100個の設問の1変数の分布の検討をなぜか行っていなかったので、一元配置の分散分析を行い結果はすぐに分かった。合格群のある設問が全員解答し正解の1という定数を取り、不合格群が0と1をとってバラツキていることが原因であるということが分かった。筆者は、これまでJMPを信頼しきっていたが、まさかこのような場合を想定した製品検査をしていなかったとは思わなかった。それと正則化技術に詳しくなく、それを用いたRDAも関係したため解明に時間がかかった。

- 1) この問題は一般化逆行列技術の問題で、単に一定値に乱数を加えて変動させることで解決できる。
- 2) これをよい機会と考えて、改定IP-OLDFが統計的な判別分析より優れているので、JMPにインストールすることを米国のJMPに申し出た。

これまでのSAS社との関係で楽観視していたが提案を拒否された。さらに、定数に乱数を加える改善方法は、他の統計ソフトが採用したと考えられるのに今の時点で行われていない。RDAはユーザーが2個のパラメータを $[0, 1]$ の範囲で自分で決めることになったが、QDFは未だに解決されていない。筆者が日本にSASを導入した際、JMPの開発者のSall博士の“Regression Application”とGoodnight社長の“Sweep Operator (吐出し演算子)”と「一般化逆行列」の3冊のテクニカルレポートに感激した。前の2論文は訳著として出版(新村, 1986)したが、最後の論文はすっかり忘れていた。公式見解は聞けないが、どうも一般化逆行列技術はSAS社の大切な技術遺産であるようだ。2013年の定年前に、このままでは研究が日の目を見ないまま忘れられると危機感を持った。そこで、2014年からは日本の費用のかかる学会発表をやめ、英語論文を年3本以上執筆することにした。それが実現できると分かったが、単に出していてもインパクトがない。たまたま国際的な研究者DBのResearch Gateを知った。「判別関数に関する英語論文を中心に絞ってPDFをUpload」した。インターネットの世界、あっという間に万を越える手ごたえが得られるかと当初期待したが、週に最大でも100件以下の

Read数で右往左往しているが、取りあえず実績が積み上がっていく。Read数以外、引用数、履歴閲覧数、Impactファクターやどの国の誰が論文を読んだかが分かって便利である（新村，2015）。そして日本の学術誌ではまず採択されない「判別関数は推測統計学でない（問題4）」ことを「100重交差検証法（新手法1）」を用いた95%信頼区間の2編の論文が米国のSOICに採択され[31,33]，RGにUPすることができた。一応，これで自分の一生の研究テーマが退官前に片がついたと感じたが，今1つインパクトに欠けていた。2015年10月26日に富山市で開催された科研費シンポジウムで判別係数の95%信頼区間の発表を終え，筑波大学の院生のMicroarrayデータの高次元データの主成分分析の発表を聞いて，論文発表に用いたデータで公開されているものがあることを知り，彼女からデータの所在を得た。早速データをダウンロードし，改定IP-OLDFと他のLDFで判別した。改定IP-OLDFだけが自然に数千から数万の遺伝子から数十個の遺伝子のMatroskaを特定できることが分かった[34]。その後，Microarrayデータは複数のMatroskaの排他的な和集合であることが分かり，41日間で15本のフリーペーパーをRGに発表し，一応一生の研究テーマが退官前に辛うじて解決できた。多くの優秀な先生方が，退官後数年で研究現場から去って行かれる。しかし，RGに自分の世界を築けば，

- 1) 少なくとも1年半で2500人以上の閲覧者と500人以上のFollowerを得た。国際会議で不特定多数を対象に研究発表するよりも効果的である。退官後も，研究費を抑えて研究が継続できる。
- 2) また15本の論文は書いた時点ですぐにUPすればタイムラグがなく発表ができる。しかも癌と正常を線形分離できる遺伝子のリストを載せた60頁以上の論文も発表できる利点がある。

もっと若い時代から，日本の学会発表や日本語の論文や書籍の出版をほどほどにして，RGを通して世界に情報発信することに取り組んでおけばと後悔した。僅か1年半で手ごたえが得られたので，多くの教員の利用を勧めたい。

### 3. 判別分析の新理論

筆者が開発した判別分析の新理論とは，整数計画法（IP）を用いたMNM基準に基づくIP-OLDFと改定IP-OLDF，線形計画法（LP）による改定LP-OLDFと改定IPLP-OLDF[29]のMPによる4つのLDFである。IP-OLDFで定義された判別係数の空間上で判別関数とNMの関係が分かった。改定IP-OLDFは問題1と問題2，新手法1は問題4，そして分散共分散行列の一般化逆行列の瑕疵の問題3を解決した。Fisherは計算機環境のない時代に対象とする現象を正規分布と仮定し判別分析の理論を定式化してくれた。そして多くの分野に適用された。このことを筆者は感謝している。しかし，計算機環境とソフトが充実した現在，実際の研究対象に目をつむり楽であるからといって正規分布を大前提として研究を進めることには問題があると考

えている。

### 3.1 4個の問題と2個の知見

Fisherは、分散共分散行列に基づいてFisherのLDFを提案し、判別分析の理論を確立した。しかし、判別分析には4個の重要な問題がある[28, 32]。そこで、改定IP-OLDFを開発した。これは、MNM基準に基づいて誤分類数を最小にしている。直接、判別係数空間上で定義したOCPの内点に対応する判別係数を求めている。これまで見つけた主要な問題点は次の通りである。

**問題1**： $p$ 変数の $f(\mathbf{x}_i)$ を任意のLDFとする。判別規則は非常に簡単で、拡張された判別スコアが正 ( $y_i * f(\mathbf{x}_i) > 0$ ) であればケース $\mathbf{x}_i$ が正しく分類され、負 ( $y_i * f(\mathbf{x}_i) < 0$ ) であれば $\mathbf{x}_i$ が誤分類される。この判別規則に等号が入る余地はない。すなわち判別超平面上のケースをいずれの群に判別するかは未解決の問題であることを心電図の医学診断を研究していて分かった。2010年に基礎研究を終えて日科技連出版から『最適線形判別関数』を出版した後、周りの統計研究家にこの点をどう考えるか調査したが、群1に含めることに何ら問題を感じていない多くの研究者から、実に様々な間違った他の考え方があることに驚愕した。この問題は、ケースの値を線形超平面の係数として、判別係数の空間でIP-OLDFを定義すると、有限個の凸体(CP)に分割できる。凸体の内点に対応したLDFは $NM = k$ 個の同じケースを誤分類し、凸体の頂点や辺に対応したLDFは必ず $p$ 個以上のケースが判別超平面上にきてその帰属を決定できないので、統計ソフトの出力する誤分類数は増える可能性がある。また凸体は有限個なので、必ず最小のNMすなわちMNMを持つOCPがある。この凸体の頂点を求めるIP-OLDFはデータが一般位置にない場合は、正しいOCPを求めないことが分かった。そこで内点を直接求める改定IP-OLDFを提案し、Fisher以降の新しい判別理論の中核とした。

- 1) またMNMは単調減少性 ( $MNM_k \geq MNM_{(k+1)}$ ) があり、フルモデルが必ず最小になるので学習標本ではMNMをモデル選択に使えないが、
- 2)  $MNM_k = 0$ になれば、この $k$ 個の変数を含む全てのモデルが線形分離可能になるという重要な事実が分かった。この事実に満足し、線形分離可能なデータ(LSD)はMatroska構造をもつことを見逃していた。これが、応用研究としての「遺伝子解析」のポイントである。

**問題2**：Vapnik[59]は、ハードマージン最大化SVM(H-SVM)、ソフトマージン最大化SVM(S-SVM)およびカーネルSVMを提案した。H-SVMはLSDを明確に示したが、“ $MNM = 0$ ”でもってLSDを明確に定義できる。多くの研究者は、判別分析の目的はLSDの判別でなく、重複データを判別することであると主張している。しかし、LDFの世界で“ $MNM > 1$ ”で初めて重複データを定義できるので、この主張は完全に間違っている。すなわちLSDの判別の研究は、2010年以降の応用研究で筆者が初めて行い、ほぼ2015年末に漸く完成した。なぜ研

究が行われなかったかの理由は、

- 1) H-SVMはLSD以外のデータに適用できないので、誰も実際の判別に利用しなかった。また Kernel SVMのアイデアに多くの研究者が興味を持って、H-SVMの研究はスルーされた。
- 2) IP-OLDFでスイス銀行がLSDであることが分かったのは、63個全ての判別モデルを改定IP-OLDFで検討するハードワークで初めて分かった。
- 3) LSDデータを研究に用いているデータから探すためには、上記のように63個全ての判別モデルを検討する必要がある。このようなアプローチをとっているのは、分析作業が大変なので筆者しかいない。

しかし試験の合否判定を大門4問の得点合計が50点以上を合格とした場合、 $f=T1+T2+T3+T4-50$ という自明なLDFがあり、 $f \geq 0$ であれば合格、 $f < 0$ であれば不合格と判定できる。等号を含めることができるのは、説明変数で判別規則が記述できるからである。この他、2群の平均値を拡大することで、LSDのデータを作り出せることをその後の研究で示した。

**問題3**：問題3は、JMPのQDFとRDAが用いている一般化逆行列の欠陥である。2013年に一つの群に属する変数値が一定であり、他群の値が変化した場合、QDFはクラス1に属する全てのケースをクラス2に誤分類することが分かった。しかし一定値に小さな乱数を加えることで問題3を解決できる。

**問題4**：Fisherは、LDFの誤判別率と判別係数の標準誤差(SE)を定式化していないので、判別分析は回帰分析のような推論手法ではない。そこでリサンプリングと相互検証を用いた新手法1を開発した。これで、筆者は検証標本における平均誤分類確率(M2)が最小のモデルを「最適モデル[51]」とするモデル選択を、一つとっておき法(LOO)[12]に代わって提案した。8個のLDFの最適モデルを比較すると、改定IP-OLDFは8個のLDFの中で多くの検証で最小になる。Vapnikは「サポートベクトルは汎化能力を持っている」と主張しているが、最適モデルは非常に簡単に優れたモデル選択手法であり、選ばれたモデルは汎化能力があると考えられる。また、次のように8個のLDFのおおよその順位を得た。：改定IP-OLDF(または改定IPLP-OLDF)、ロジスティック回帰、SVM4(C=10000)、そしてFisherのLDF是最悪である。改定LP-OLDFは問題1に弱く、SVM1(C=1)は多くの場合でSVM4より劣っている。

### 3.2 検討する8個のLDF

本稿では、2個の統計的なLDFと6個のMPによるLDFを評価する。FisherのLDFと式(1)のロジスティック回帰はJMP[16]で分析した。日本SASインスティテュート社のJMP部門は、JMPスクリプトで新手法1を実行するプログラムの作成をしてくれた。これらに加えて、QDFとRDAで教師データの判別を行なう。



$$\text{Log}(p/(1-p)) = f(x) \quad (1)$$

Where

p: the probability belongs to class1; x: the independent variables.

LINGO [17] は6個のMPによるLDFを定義する。式(2)の改定IP-OLDFは、IPでMNMを見つけることができる。改定LP-OLDFは誤分類されるケースに限定して判別超平面からの距離の総和を最小化することをLPで定式化したL1ノルムLDFの一種である。問題1の影響が一番受けることが分かったが、1) LSDの判別ではH-SVMより好成績であり、2) H-SVMと異なり、LSDでない判別も可能であり、3) QPでなくLPで解けるので高速である。H-SVMはLSDの判別を明確に示してくれたが、応用上は改定LP-OLDFより劣っているのではないかと考えている。式(2)で整数変数を非負の実数変数にかえると、改定IP-OLDFが改定LP-OLDFになる。改定IPLP-OLDFはMNMの近似値を高速で探す(新村, 2007a)。しかし2012年以降は、IPソルバーの高速化で、改定IP-OLDFより遅くなった[29]。改定LP-OLDFで正しく判別されたケースを $e_i = 0$ に固定し、誤判別されたケースだけに改定IP-OLDFを適用する混合モデルである。

$$\text{MIN} = \sum e_i; y_i * (x_i b + b_0) \geq 1 - M * e_i; \quad (2)$$

Where

$e_i$ : 0/1 integer decision variable; M: big M constant (M=10000);

$b_0$ : free decision variables.

式(3)でVapnikはLSDの概念を明確に示すH-SVMを提案した。今まで多くの研究者は、判別の目的は重複データを判別することであると主張している。H-SVMと改定IP-OLDF以外のLDFは、LSDの判別を理論的に保証しないので、「オーバーラップ」または「オーバーラップしない」の状態を定義できない。すなわち「MNM = 0」はLSDを意味し、「MNM > 1」が2個のクラスが重なることを意味する。実際のデータはほとんどLSDでないため、式(4)のS-SVMが提案された。「ペナルティ c」は、2個の目的式を結合するが、正しい「c」を決定する規則がない。本研究では、SVM4 (C = 10000) と SVM1 (C = 1) の2個のS-SVMを検討する。学習標本と検証標本の両方で、SVM4の平均誤分類確率 (M1とM2) はSVM1より優れている。SVMは二次計画法 (QP) で定式化される。非線形計画 (NLP) で定式化されるカーネルSVMはLDFでないので検討しない。また、S-SVMのペナルティ cやRDAの2つのオプションでチューニングするという方法は、広く統計で行われているが、最適化手法に恣意的な判断を取り入れる欠点があるという認識に欠けているように考える。

$$\text{MIN} = \|b\|^2/2; y_i * (x_i b + b_0) \geq 1; \quad (3)$$

Where

$y_i = 1 / -1$  for  $x_i \in \text{class1/class2}$ ;  $x_i$ : p-independent variables (p-variables);

$b$ : p-discriminant coefficients;  $b_0$ : the constant and free variable.

$$\text{MIN} = \|b\|^2/2 + c * \sum e_i; \quad y_i * (x_i b + b_0) >= 1 - e_i; \quad (4)$$

Where

$c$ : penalty  $c$ ;  $e_i$ : non-negative decision variable.

MPによるLDFの表記は似ているが、LP、QP、およびIPソルバーで結果が異なってくる。また統計家に上のモデルを示すと、解法のアロゴリズムの記述が不明確と指摘されることが多い。解法はLP、QP、IPおよびNLPである。また制約のあるなしに関係なく、MPソルバーで目的関数の最大値/最小値を求めることができるので数学ソフトの解けない関数の最大/最小が分かる。このことを前提にScrage教授は回帰モデルの幾つかの定義を導入した。QPは最小二乗法を定義し、LPは「絶対値最小化(LAV)回帰」を定義し、NLPでLpノルム回帰が定式化できる(新村, 2011b)。しかし回帰分析の研究論文は少ないのに、MPによる判別関数の多くの研究論文がある。これらの研究は、実データで評価しなかったため統計ユーザーは利用しなかった。1997年以前のMPによる判別モデルを総括するStamの論文[56]の後、MPによる判別関数の研究の第一段階が終了したと考えられる。RGに筆者のこの点に触れた論文をUPして、暫くするとStamをはじめとするこの分野の先達が筆者の論文を詳しく検討しているという通知がRGより届いた。しかし、私の論文に対する批判はない。またStam教授は、その後彼の論文をUPしたというメールが来たが、勝手に「この論文も読んだ方がいいよ」というメッセージと解釈している。一方、Vapnikは実データによる検証を行い3つの異なるSVMモデルを統計とORという気難しい分野でなくパターン認識の分野を中心に提案したのは賢明である。しかし筆者のように、8個の異なるLDFの比較を17種以上のデータを用いて体系的に行っていないと考えている。

### 3.3 新手法1と最適モデルの選択

問題4の解決のため、新手法1を提案した。これによって判別理論は伝統的な推測統計学でないが、その欠点を補うことができる。ロジスティック回帰は、Fisherの開発した最尤推定法で求めたヘシアン行列からロジスティック回帰係数のSEを出している。小西ら(1992)は、Bootstrap法で判別関数の誤分類確率のSEの式を求めている。しかし新手法1は、直接研究に用いているデータの個々の95%信頼区間(confidence interval, CI)を求めていて、実際のデータを分析する研究者に便利である。これらのコンピューターを利用した方法は、確率分布から導かれた伝統的な推測統計学と一線を画すべきである。

- 1) 最初「 $K = 10$ 」にしていたが、95% CIを求めるために「 $k = 100$ 」にした。そして元の標本を100回コピーし、検証標本として擬似母集団を生成する。
- 2) この標本に乱数を追加し、標本を昇順にソートする。この標本を100個に分割し1から100

までの部分標本番号を追加する。

- 3) 100個の部分標本を学習標本とし、擬似母集団を検証標本とする。この方法は、100個の部分標本を擬似集団からサンプリングしたのと同じ効果がある。一つの部分標本を学習標本とし、残り99個を検証標本とするようなLOO法的な扱いは効果的でない。検証標本は擬似母集団であり、ユニークでなければならないと考えている。元の標本と擬似母集団は同じ分布なので多くの試行ミスを避けることができる。この手法1で、幾つかのデータで8種のLDFを比較し顕著な成果を得た。また、誤分類確率と判別係数の95% CIを検討できた。

#### 4. The Matroska Feature Selection Method for Microarray Data (新手法2)

##### 4.1 判別分析の新しい問題5

これまでLSDの判別分析の研究はない。多くの統計学者は、

- 1) LSDの判別が非常に容易であると誤解し、
- 2) 判別分析の目的はLSDでなく重複データを判別することと主張することが多い。しかしLSDと重複データは背反である。
- 3) H-SVMと改定 IP-OLDFのみが理論的にLSDを認識でき、「 $MNM \geq 1$ 」という条件で重複データを定義できるが、これができるのはH-SVMと改定IP-OLDFだけである。  
なぜこれまでLSDの判別研究がなかったかは、以下のように考える。
- 1) VapnikがH-SVMでLSDを明確に定義した。しかし、H-SVMはLSDの判別にしか使えないので、実際の判別分析に利用されなかった。
- 2) Vapnikは魅力的なカーネルSVMを提案し、ほとんどの研究者がこのモデルに注目しH-SVMに注目しなかった。
- 3) 研究データがLSDであるか否かは、改定IP-OLDFで全てのモデルを検討する必要がある。
- 4) しかしMicroarrayデータはLSDではあるが、多分H-SVMでMicroarrayデータの判別の研究はできなかった。このため、多くの研究者は10年以上に渡ってMicroarrayデータを研究してきたが明確な結果が得られていない。改定IP-OLDFだけが、現時点でMicroarrayデータの構造を容易に説明できる。すなわち10年以上解決できなかった判別分析の問題5が、問題の提起と同時に突然に解決できた。

##### 4.2 新手法2の概略

近年、研究論文に分析に用いたデータを公開し、他の研究者が検証できる研究分野が増えてきているようだ。筆者自身、それらの情報を正しく把握していなかったため、2010年に基礎研究を終えて応用研究として「試験の合否判定」でLSDの研究を性急に始めてしまった。その時点で、多くの研究者が遺伝子の判別を通常の判別手法で行おうとしていて多分失敗す

ると考えていた。しかし、ケース数 ( $n$ 個) が100前後として、分析する遺伝子 ( $p$ 個) は1万を超えるものがざらである。ケース数が大規模であっても変数が少なければ、計算時間がかかるが従来の統計手法がそのまま利用できる。しかし遺伝子情報は高次元空間のデータとして多くの研究者が研究してきたが、 $n$ に対して $p$ がけた外れに大きい。 $P = 10,000$ とすれば相関係数を求めようとしても  ${}_{1000}C_2 = 10000 * 9999 / 2 = 5000 * 9999$  個ある。それ以前に、わずか1000個のデータのばらつきから10,000個の分散共分散行列や相関行列を求めることは難しく、その研究が活発に行われていた。国際会議で、「FisherのLDFは通常の判別でもうまくいかないのに、さらに不明確な分散共分散行列を求めたうえで、通常の判別分析に持ち込むことは如何なものか?」と質問すると、いやな沈黙の洗礼を受けた。現時点で反省すれば、「なぜ自分で具体的に検証しようとしなくて、試験データの応用研究を優先させたか」大いに後悔している。

2015年の10月27日に富山の科研費シンポジウムから帰った。翌28日に筑波大学の博士課程の石井さんから6個のMicroarrayデータを掲載したHPのメールを受け取った。早速ダウンロードすると32bitのExcelでは3個しか展開できない。そこで一番小さなAlon et al.のデータ[1]を改定IP-OLDF, 改定LP-OLDF, 改定IPLP-OLDF, H-SVM, S-SVM, FisherのLDF, ロジスティック回帰で判別した。MPによるLSDは全てNMが0になり、MicroarrayデータがLSDであり、癌と正常の2群がかなり離れていることが分かった。さらに驚くことに、改定IP-OLDFの判別係数は多くが0であり僅か72個が0でなく自然に2000個の遺伝子から72個の遺伝子でMNM=0と判別できることが分かった。改定LP-OLDFと改定IPLP-OLDFは、0でないものが100変数以上と多い。しかし、H-SVMとS-SVMはほとんどが0でない。分かっていたが、JMPでFisherのLDFとロジスティック回帰で判別すると、Errorで終了した。早速それを論文にまとめた。既存の媒体に発表を試みても掲載に時間がかかる。海外の学者は、フリーペーパーを発表している研究者も多い。そこで日時を明記してRGに人生初のフリーペーパーをUploadした。翌日からShippら[54]とGolubら[8]の7000変数程度のデータを分析し、Alonの結果を再確認した[34-37]。さらに変数選択された遺伝子を用いて再度判別すると、より少ない遺伝子を選ばれた。そこで初めてMicroarrayデータはLSDであり、その中にMNM=0になる部分空間がMatroskaのように詰まっていることを再認識した。2000個のAlonのデータはLSDでありBig Matroskaと呼ぶことにした。その中に  $(2^{2000}-1)$  個のモデルが詰まっているが、MNM=0になるものだけをMatroskaと呼ぶことにする。なぜか2000個の遺伝子を持つBig Matroskaを判別すると、途中のより小さなMatroskaを飛ばして、72個の遺伝子を持つSmaller Matroska (SM) が出てきた。改定IP-OLDFを用いてさらに判別すると、さらに小さなMatroskaが現れる。しかし、3回ほど行くとそれ以上小さなSMが得られない。しかしこれが最小のMatroska (BGS) かどうか分からない。そこでやむを得ず、変数選択法であたりを付けて、全てのモデ

ルで改定IP-OLDFを用いてBGSを見つけた。さらに念のため、元のデータから最初のSMに含まれる遺伝子を除外し、もう一度判別するとさらに他のMatroska構造が見つかった。すなわち遺伝子データは線形分離可能な幾つかのMatroskaの和集合と次元の大きな線形分離可能でない排他的な高次元空間の和集合になっている。10年以上に渡り多くの研究者が高次元データのまま分析していても何も成果が得られないのは、この特徴が理解されていないからである。暫くして、HPを作成したJefferyから遺伝子解析用に開発した自分の製品の使用を薦めるメールがRG経由で届いた。見てみると古色蒼然として役に立たないので婉曲に断った。しかし、マニュアルでこの操作を継続しても幾つのSMがあるかを確認できない。そこでLINGOで汎用モデルを開発した。また、64bitのMS Officeを購入し残りの1万個以上のMicroarrayデータが扱えるようになった。11月10日に六本木でJMPのユーザー会に参加した。尊敬するSall博士が、講演でJMP12でMicroarrayデータが判別できるLDFを開発し、判別結果も報告された。筆者は興奮し、すでにLINGOで分析を行い成功していること。そしてJMP12を借用し検証してよい結果が得られれば購入すると質問した。翌日借用したJMP12で分析してがっかりした。以前に予見していたことであるが、誤分類確率が大きい。しかし一部0になるものもある。そこでそれを記述した論文をUploadするとともに、JMPに送った。ここで初めて、JMPの開発責任者2名が漸く筆者のRGに訪れた。そして借用期限の1か月前に6個すべてを判別すると、誤分類確率が0のものがなくなっていた。筆者の結果を見直して、判別成績が悪くなるが瑕疵を修正したと考えられる。

表1は、LINGOの汎用モデルで分析した結果で、HPからダウンロードした6個のMicroarrayデータの要約である。列のDescriptionは2個のクラスの症例数を示す。Sizeはケース数と遺伝子の数で、「SM: Gene」は「SMの数とそれに含まれる遺伝子総数である。完全な遺伝子名は、参照リストの論文にアップしてある。「Mean, Max, Min」はSMに含まれる遺伝子の平均値、最大値と最小値である。「JMP12」列は、MicroarrayデータのためのFisherのLDFによる判別分析の2×2の分割表である。6個のNMは、5, 3, 8, 3, 10および29である。Alonの最初の判別結果はNM=0であったが、12月8日に再計算すると5に代わっていた。新手法2で幾つかの新しい用語、例えば「Matroska, Matroska系列 (またはMatroska製品), 最小Matroska (SM), 基本的な遺伝子の部分空間 (BGS)」を用いているが、ほとんどの人がこれらの用語を理解することは困難である。従って、5章のスイス銀行紙幣データと6章の日本車データで、これらの用語を説明する。

表1 Summary of six Microarray Data [11]

Data	Description	Size	SM: Gene	Mean	Max	Min	JMP12
Alone et al. [1]	Normal (22) vs. tumour cancer (40)	62 *2000	64 [44]:1152	18	39	11	20:2/3:37
Chiaretti et al. [2]	Bcell (95) vs. Tcell (33)	128*12625	270 [47]:5385	19	62	9	94:1/2:31
Goulb et al. [8]	All (47) vs. AML (25)	72*7129	69 [43]:1238	18	31	10	20:5/3:44
Shipp et al. [54]	Follicular lymphoma (19) vs. DLBCL (58)	77 *7130	213 [42]:3032	14	43	7	17:2/1:57
Singh et al. [55]	Normal (50) vs. tumour prostate (50)	102 *12626	179 [45]:3990	22	47	13	46:4/6:46
Tian et al. [58]	False (36) vs. True (137)	173 *12625	159 [46]:7221	45.4	104	28	16:20/9:128

表2はGolubらの69個のSMの完全なリストである。列 SM1はSMの連続番号で69個のSMがあり、このデータは、69個のSMで構成されていることが分かる。列の「n」は、各SMに含まれる遺伝子の数である。ほとんどの研究者が統計的方法またはLASSOという新しい手法で高次元の遺伝子空間の分析に苦労しているが、68番と69番のSMは高々31個の遺伝子しかないので、各SMを分析することは非常に簡単である。

表2 Small Matroska of Golub et al. Data

SM1	SM2	Gene	n	MNM	35	11	6630	17	0
1	11	7129	11	0	36	11	6613	19	0
2	11	7118	16	0	37	11	6594	12	0
3	11	7102	11	0	38	11	6582	16	0
4	11	7091	10	0	39	11	6566	16	0
5	11	7081	13	0	40	11	6550	16	0
6	11	7068	12	0	41	11	6534	19	0
7	11	7056	13	0	42	11	6515	14	0
8	11	7043	12	0	43	11	6501	19	0
9	11	7031	14	0	44	11	6482	14	0
10	11	7017	16	0	45	11	6468	21	0
11	11	7001	10	0	46	11	6447	21	0
12	11	6991	12	0	47	11	6426	20	0
13	11	6979	13	0	48	11	6406	23	0
14	11	6966	16	0	49	11	6383	19	0
15	11	6950	14	0	50	11	6364	19	0
16	11	6936	13	0	51	11	6345	24	0
17	11	6923	19	0	52	11	6321	19	0
18	11	6904	15	0	53	11	6302	20	0
19	11	6889	13	0	54	11	6282	22	0
20	11	6876	14	0	55	11	6260	19	0
21	11	6862	16	0	56	11	6241	24	0

22	11	6846	17	0	57	11	6217	21	0
23	11	6829	17	0	58	11	6196	25	0
24	11	6812	14	0	59	11	6171	27	0
25	11	6798	16	0	60	11	6144	20	0
26	11	6782	15	0	61	11	6124	23	0
27	11	6767	12	0	62	11	6101	28	0
28	11	6755	21	0	63	11	6073	23	0
29	11	6734	15	0	64	11	6050	23	0
30	11	6719	14	0	65	11	6027	28	0
31	11	6705	22	0	66	11	5999	23	0
32	11	6683	19	0	67	11	5976	23	0
33	11	6664	16	0	68	11	5953	31	0
34	11	6648	18	0	69	11	5922	31	0

## 5. スイス銀行データによる新手法2の解説1

### 5.1 Matroska構造と1個のBGS

IP-OLDFで6変数をもつ200ケースのスイス銀行紙幣データを判別すると、表3のように2変数モデル (X4, X6) でMNM = 0であることを見つけた。これで (X4, X6) を含む63 (=2<sup>6</sup>-1=63) 個のモデルのうち、16個は線形分離可能なモデルで、残り47個は線形分離可能でない。16個のモデルのうち6変数の最大のMatroskaに、残り15個のMatroskaを組み合わせることでMatroska製品を製造することができる。Matroska製品には、最後に必ずBGSの (X4,X6) が唯一1個含まれる。このBGSは、MNMの単調減少性とMNM=0の場合、このBGSを含む全てのモデルがMatroskaすなわち線形分離可能なモデルでMatroska製品の部品になる。これが新手法2の骨子である。癌治療で、BGSになる遺伝子を直接修復すれば良いであろう。残りの線形分離可能でない高次元の部分空間の遺伝子は、癌との関連性はわからないが、修復の優先度が低いことは確かである。

表3 16個の線形分離可能なモデル

SN	p	var.	RIP	logistic	SVM4	SVM1	LDf	QDF	RDA
1	6	1-6	0	0	0	0	1	1	1
2	5	2-6	0	0	0	0	1	1	1
3	5	1,3-6	0	0	0	0	1	1	1
4	5	1,2,4-6	0	0	0	0	1	1	1
5	5	1-4,8	0	0	0	0	1	1	1
8	4	3-6	0	0	0	0	1	1	1
9	4	2,4-6	0	0	0	0	1	1	1
10	4	1,4-6	0	0	0	0	1	1	1
11	4	2-4,6	0	0	0	0	1	1	1
12	4	1,3,4,6	0	0	0	0	1	1	1
13	4	1,2,4,6	0	0	0	0	2	1	1
23	3	4-6	0	0	0	0	1	1	1

24	3	3,4,6	0	0	0	0	1	1	1
25	3	1,4,6	0	0	0	0	2	2	1
26	3	2,4,6	0	0	0	0	1	1	1
27	2	4,6	0	0	0	0	3	1	1

表4は、Matroskaの生産業者の観点から、スイス銀行紙幣データの構造を示す。列SNはMatroskaの製品番号である。5つの列「6,5,4,3,2」は、Matroskaに含まれる変数（遺伝子）である。6変数の大きなMatroskaに、4個の5変数のMatroska (X2, X3, X4, X5, X6), (X1, X3, X4, X5, X6), (X1, X2, X4, X5, X6), (X1, X2, X3, X4, X6) を含み、それらはMNM=0である。2個のモデルの (X1-X3, X5, X6) と (X1-X5) はMatroskaではない。5変数のMatroskaには3個の4変数のMatroska, 各4変数のMatroskaには2個の3変数のMatroskaを含んでいる。最後に、各3変数のMatroskaは、新手法2では最小のMatroska (X4, X6) をBGSと呼ぶ。このBGSでスイス銀行紙幣データの構造を記述できる。Matroskaの生産者は、16個のMatroskaの組み合わせで24個のMatroska製品を作ることができる。各Matroskaの製品は、以下のMatroska系列で仕様を定義できる。例えば、最初のMatroska製品をSN = 1で表すと以下のMatroska系列を持っている：(1-6)  $\ni$  (2-6)  $\ni$  (3-6)  $\ni$  (4-6)  $\ni$  (4,6)

表4 16個の線形分離可能なモデル

SN	6	5	4	3	2
1	1-6	2-6	3-6	4-6	4, 6
2				3, 4, 6	4, 6
3			2, 4-6	4-6	4, 6
4				2, 4, 6	4, 6
5			2-4,6	3, 4, 6	4, 6
6				2, 4, 6	4, 6
7	1, 3-6	3-6	4-6	4, 6	
8				3, 4, 6	4, 6
9			1, 4-6	4-6	4, 6
10				1, 4, 6	4, 6
11			1, 3, 4, 6	3, 4, 6	4, 6
12				1, 4, 6	4, 6
13	1, 2, 4-6	2, 4-6	4-6	4, 6	
14			2, 4, 6	4, 6	
15		1, 4-6	4-6	4, 6	
16			1, 4, 6	4, 6	
17		1, 2, 4, 6	2, 4, 6	4, 6	
18			1, 4, 6	4, 6	
19	1-4, 6	2-4, 6	3, 4, 6	4, 6	
20			2, 4, 6	4, 6	
21		1, 3, 4, 6	3, 4, 6	4, 6	
22			1, 4, 6	4, 6	
23		1, 2, 4, 6	2, 4, 6	4, 6	
24			1, 4, 6	4, 6	



5.2 新手法2の解説

スイス銀行紙幣データで新手法2を説明する。表5は、6個のLDFのNMと改定IP-OLDFの判別係数を示す。最初の16個のモデルは(X4, X6)を含むため、SVM1以外の5個のLDFのモデルはMNM=0である。多くの研究者は、

- 1) S-SVMが線形分離可能なモデルを判別できると考えているが、間違いであることが分かる。
- 2) またペナルティ cとしてc = 1のような小さな値を選択することを好むが、SVM1 (C = 1の場合)は16個を正しく判別できない。多くの分析でSVM4はSVM1よりも優れていることを他の分析でも確認している。

X1からCの7列は、改定IP-OLDFの判別係数である。改定IP-OLDFでデータを判別するとX2およびX3の2個の係数は自然にゼロになる。従って6変数から4変数に特徴選択を自然に行うことができる。そこで「SN = 8の4変数モデル (1, 4-6)」を判別すると、より小さなモデルに変数選択できないので4変数モデルで変数選択を停止し、この4変数モデルをSMと呼ぶ。このステップの後、統計的アプローチでBGSの(X4, X6)を探す必要がある。フルモデルからBGSを削除した後、改めて改定IP-OLDFで4変数モデル(1-3, 5)のサイズの小さなモデルを判別する。このモデルのMNMが18であるため、変数選択を停止し(X4, X6)が1個のBGSになる。すなわちスイス銀行紙幣データは、フルモデルを大きなMatroskaとし、その中に5変数から2変数の小さなMatroskaを含む製品が24個含まれるが、2変数のBGSでこの構造を全て記述できる。

表5 6個のNMと改定IP-OLDFの判別係数

SN	Var.	RIP	SVM4	SVM1	LP	IPLP	HSVM	X1	X2	X3	X4	X5	X6	c
1	1-6	0	0	1	0	0	0	-1.09	0	0	-2.605	-2.827	2.0618	0
2	2-6	0	0	1	0	0	0		0.4079	1.8425	-4.177	-4.627	2.1941	-513
3	1,3-6	0	0	1	0	0	0	-1.09		0	-2.605	-2.827	2.0618	0
4	1,2,4-6	0	0	1	0	0	0	-1.09	0		-2.605	-2.827	2.0618	0
5	1-4,6	0	0	2	0	0	0	7.2219	-5.243	2.331	-11.12		10.907	-2606
6	3-6	0	0	1	0	0	0			1.684	-3.935	-4.308	2.1904	-444.3
7	2,4-6	0	0	1	0	0	0		-2.937		-2.473	-2.704	2.2947	113.14
8	1,4-6	0	0	1	0	0	0	-1.09			-2.605	-2.827	2.0618	0
9	2-4,6	0	0	2	0	0	0		0	6.8966	-21.52		23.724	-3408
10	1,3,4,6	0	0	2	0	0	0	13.663		-9.759	-27.99		27.701	-5308
11	1,2,4,6	0	0	2	0	0	0	8.4638	-4.232		-14.38		14.145	-3126
12	4-6	0	0	1	0	0	0				-4.804	-6.477	2.5979	-250.7
13	3,4,6	0	0	2	0	0	0			0	-44		48	-6348
14	1,4,6	0	0	2	0	0	0				-44		48	-6348
15	2,4,6	0	0	2	0	0	0		0		-44		48	-6348
16	4,6	0	0	2	0	0	0				-44		48	-6348
17	1-3, 5	18	22	22	22	18	1.23	9506.8	-4625	-9990	0	-7071	0	-67712

表6は、SVM4 (左) とH-SVM (右) の係数を示す。フルモデルの全ての係数がゼロでないので、H-SVMとSVM4は自然に特徴選択を行うことができないことが分かる。

表6 SVM4 (左) とH-SVM (右) の判別係数

SN	var	X1	X2	X3	X4	X5	X6	c	X1	X2	X3	X4	X5	X6	c
1	1-6	-1.138	-0.569	0.1248	-2.301	-2.796	1.7967	102.2	-1.14	-0.57	0.124	-2.3	-2.8	1.796	102
2	2-6	0	-2.084	0.738	-2.627	-2.489	2.2758	-92.6	0	-2.08	0.746	-2.62	-2.49	2.275	-94
3	1,3-6	-1.48	0	-0.105	-2.266	-2.886	1.6987	146.09	-1.48	0	-0.1	-2.27	-2.89	1.699	146
4	1,2,4-6	-1.293	-0.313	0	-2.293	-2.837	1.7578	124.33	-1.3	-0.29	0	-2.3	-2.83	1.758	123
5	1-4,8	7.6939	-5.537	3.745	-10.26	0	10.187	-2759	7.694	-5.54	3.746	-10.3	0	10.19	-2759
6	3-6	0	0	0.6575	-3.306	-2.862	3.0182	-447.4	0	0	0.657	-3.31	-2.86	3.018	-447
7	2,4-6	0	-1.923	0	-3.173	-2.115	2.4519	-41.83	0	-1.92	0	-3.17	-2.12	2.452	-42
8	1,4-6	-1.448	0	0	-2.275	-2.901	1.7383	120.11	-1.45	0	0	-2.27	-2.9	1.738	120
9	2-4,6	0	-4.828	6.8995	-21.52	0	23.725	-3409	0	-4.83	6.897	-21.5	0	23.72	-3408
10	1,3,4,6	13.663	0	-9.759	-27.99	0	27.701	-5308	13.66	0	-9.76	-28	0	27.7	-5308
11	1,2,4,6	8.4638	-4.233	0	-14.38	0	14.145	-3126	8.464	-4.23	0	-14.4	0	14.14	-3126
12	4-6	0	0	0	-3.75	-2.5	3.125	-377.3	0	0	0	-3.75	-2.5	3.125	-377
13	3,4,6	0	0	-1E-05	-44	0	48	-6348	0	0	0	-44	0	48	-6348
14	1,4,6	24.746	0	0	-29.15	0	30.678	-9366	24.75	0	0	-29.2	0	30.68	-9366
15	2,4,6	0	0	0	-44	0	48	-6348	0	-0	0	-44	0	48	-6348
16	4,6	0	0	0	-44	0	48	-6348	0	0	0	-44	0	48	-6348
17	1-3,5	1.9142	-1.219	-2.704	0	-1.528	0	115.03							

表7は、改定LP-OLD (左) と改定IPLP-OLD (右) の係数を示す。フルモデルの X2 と X3 の2個の係数はゼロである。従って、6変数から4変数に自然に特徴選択を行うことができる。「SN=8の4変数モデル (1, 4-6)」を判別すると4変数モデルをより小さなモデルに減らすことはできないので変数選択を停止する。両方の結果は改定IP-OLDFと同じであるが、改定IP-OLDFは改定LP-OLDFと改定IPLP-OLDFよりも高次元の遺伝子空間をより小さなSMに減らすことができることを確認している。さらに改定LP-OLDFと改定IPLP-OLDFは、理論的にはLSDを判別できる保証はないので、Microarrayデータの判別にこれらのLDFを使用しない。MPによる改定IP-OLDFなどが特徴選択を自然に行うことができる理由を理論的に説明できない。しかし、H-SVMが特徴選択を行うことができない理由を恐らく次の点と考えられる。

- 1) SVMは2個のサポートベクトルにケースを固定するので、ケースの有効桁数が大きい場合、係数が0すなわち座標軸と平行にならない。逆に有効桁数が小さい場合には幾つかの係数が0になるかもしれない。
- 2) Microarrayデータの有効桁数が大きいので、H-SVMは特徴選択を行うことはできない。

表7 改定LP-OLDF (左) と改定IPLP-OLDF (右) の判別係数

SN	var	X1	X2	X3	X4	X5	X6	c	X1	X2	X3	X4	X5	X6	c
1	1-6	-1.09	0	0	-2.61	-2.83	2.06	0	-1.09	0	0	-2.61	-2.83	2.062	0
2	2-6	0	-2.94	0	-2.47	-2.7	2.30	113.135	0	-2.94	0	-2.47	-2.7	2.30	113.135
3	1,3-6	-1.09	0	0	-2.61	-2.83	2.06	0	-1.09	0	0	-2.61	-2.83	2.06	0
4	1,2,4-6	-1.09	0	0	-2.61	-2.83	2.06	0	-1.09	0	0	-2.61	-2.83	2.06	0
5	1-4,8	7.222	-5.24	2.331	-11.1	0	10.91	-2605.6	7.222	-5.24	2.331	-11.1	0	10.91	-2605.6
6	3-6	0	0	0	-4.8	-6.48	2.60	-250.69	0	0	0	-4.8	-6.48	2.60	-250.69
7	2,4-6	0	-2.94	0	-2.47	-2.7	2.30	113.135	0	-2.94	0	-2.47	-2.7	2.30	113.135
8	1,4-6	-1.09	0	0	-2.61	-2.83	2.06	0	-1.09	0	0	-2.61	-2.83	2.06	0
9	2-4,6	0	-4.83	6.897	-21.5	0	23.72	-3408.1	0	-4.83	6.8966	-21.5	0	23.72	-3408.1
10	1,3,4,6	13.66	0	-9.76	-28	0	27.70	-5307.6	13.66	0	-9.759	-28	0	27.70	-5307.6
11	1,2,4,6	8.464	-4.23	0	-14.4	0	14.14	-3126.5	8.464	-4.23	0	-14.4	0	14.14	-3126.5
12	4-6	0	0	0	-4.8	-6.48	2.60	-250.69	0	0	0	-4.8	-6.48	2.60	-250.69
13	3,4,6	0	0	0	-44	0	48	-6347.8	0	0	0	-44	0	48	-6347.8
14	1,4,6	0	0	0	-44	0	48	-6347.8	0	0	0	-44	0	48	-6347.8
15	2,4,6	0	0	0	-44	0	48	-6347.8	0	0	0	-44	0	48	-6347.8
16	4,6	0	0	0	-44	0	48	-6347.8	0	0	0	-44	0	48	-6347.8
17	1-3,5	1.914	-1.22	-2.7	0	-1.53	0	115.031	9559	-4629	-10151	0	-7089	0	-57272

### 5.3 100重交差検証法 (新手法1)

新手法1でスイス銀行紙幣データからリサンプリング標本を生成し、8個のLDFを評価する。表8は、16個の線形分離可能なモデルを示す。「M1とM2」は、学習および検証標本の平均誤分類確率である。改定IP-OLDF, H-SVM, SVM4, LP, IPLPとロジスティック回帰の全ての16個のNMはゼロである。SVM1とFisherのLDFは、全ての線形分離可能なモデルを認識できない。他のデータでは、SVM4, LP, IPLPとロジスティック回帰も全ての線形分離可能なモデルを認識できない場合があることを観察している。改定IP-OLDFは、最適モデルとして3番目のモデルを選択し、M2は0.26パーセントである。H-SVM, SVM4, 改定IPLP-OLDFおよび改定LP-OLDFは最適モデルとして8番目のモデルを選択し、M2はそれぞれ0.38, 0.37, 0.41および0.27パーセントである。SVM1とロジスティック回帰は12番目のモデルを選択し、それらのM2は0.52と0.41パーセントである。FisherのLDFのM2は0.54パーセントで7番目のモデルを選択する。改定IP-OLDFの最適モデルは8個のLDFの間でM2が最小である。3番目のモデルの7個のM2Diffは、それぞれ、0.21, 0.21, 0.28, 0.23, 0.01, 0.26, および0.29パーセントである。アイリスデータと並んで、スイス銀行紙幣データはFisherの仮説を満たし、FisherのLDFのNMはMNMに収斂する傾向がある。次に、線形分離ではない11個のモデルを検証すると、11モデルの中で最適モデルとして改定IPLP-OLDFは第23モデルを選択する。6個のM2Diffは0.08, 0.84, -0.03, 0.12, 0.33および1.75%であるため、改定IPLP-OLDFと改定IP-OLDFは5個のLDFよりもわずかに優れている。5変数モデル (X1, X3-X6) のM2は16個のモデルの中で最小であるが4変数モデル (X1, X4-X6) のM2は2番目に最小である。従

って、この改定IP-OLDFはフルモデルが4変数モデル (X1, X4-X6) に減少した理由の一つかもしれないと思われる。

表8 新手法1

RIP		M1	M2	t	Diff.	Model	
53m42s	1	0	<u>0.30</u>	<u>453</u>	0.30	1-6	
	2	0	0.77	307	0.77	2-6	
	3	0	<u>0.26</u>	456	0.26	1,3-6	
	4	0	<u>0.30</u>	453	0.30	1,2,4-6	
	5	0	0.70	243	0.70	1-4,6	
	6	0	0.74	409	0.74	3-6	
	7	0	0.75	419	0.75	2,4-6	
	8	0	<u>0.27</u>	<u>454</u>	0.27	1,4-6	
	9	0	0.77	362	0.77	2-4,6	
	10	0	0.63	379	0.63	1,3,4,6	
	11	0	0.62	379	0.62	1,2,4,6	
	12	0	<u>0.69</u>	<u>402</u>	0.69	4-6	
	13	0	0.67	353	0.67	3,4,6	
	14	0	0.60	379	0.60	1,4,6	
	15	0	0.66	366	0.66	2,4,6	
	16	0	<u>0.47</u>	<u>359</u>	0.47	4,6	
HSVM		M1	M2	t	Diff1	M1Diff	M2Diff
35m6s	1	0	0.53	-147	0.53	0.00	0.23
	2	0	0.46	182	0.46	0.00	-0.30
	3	0	0.46	-163	0.46	0.00	<u>0.21</u>
	4	0	0.45	-158	0.45	0.00	0.15
	5	0	0.72	141	0.72	0.00	0.02
	6	0	0.46	192	0.46	0.00	-0.28
	7	0	0.43	-185	0.43	0.00	-0.32
	8	0	<u>0.38</u>	-164	0.38	0.00	<u>0.11</u>
	9	0	0.70	149	0.70	0.00	-0.06
	10	0	0.66	147	0.66	0.00	0.03
	11	0	0.65	143	0.65	0.00	0.03
	12	0	0.39	184	0.39	0.00	-0.30
	13	0	0.63	147	0.63	0.00	-0.04
	14	0	0.60	142	0.60	0.00	-0.01
	15	0	0.59	142	0.59	0.00	-0.07
	16	0	0.46	140	0.46	0.00	-0.01
SVM4		M1	M2		Diff1	M1Diff	M2Diff
44m46s	3	0	0.464		0.46	0.00	<u>0.21</u>
	8	0	<u>0.374</u>		0.37	0.00	<u>0.10</u>
SVM1		M1	M2		Diff1	M1Diff	M2Diff
46m17s	3	0.26	0.54		0.28	0.26	<u>0.28</u>
	12	0.32	<u>0.52</u>		0.21	0.32	-0.17
IPLP		M1	M2		Diff1	M1Diff	M2Diff
47m31s	3	0	0.49		0.49	0.00	<u>0.23</u>
	8	0	<u>0.41</u>		0.41	0.00	<u>0.14</u>

LP		M1	M2	Diff1	M1Diff	M2Diff
19m58s	3	0.00	0.27	0.27	0.00	<u>0.01</u>
	8	0.00	<u>0.27</u>	0.27	0.00	<u>0.00</u>
Logistic		M1	M2	Diff1	M1Diff	M2Diff
46m	3	0.00	0.52	0.52	0.00	<u>0.26</u>
	12	0.00	<u>0.41</u>	0.41	0.00	<u>-0.27</u>
LDF		M1	M2	Diff1	M1Diff	M2Diff
55m	3	0.53	0.55	0.02	0.53	<u>0.29</u>
	7	0.51	<u>0.54</u>	0.03	0.51	<u>-0.20</u>

## 6. 日本車44車種の判別による新手法2の解説2

### 6.1 問題3の説明

6章では、日本車44車種のデータを使用して、問題3を説明する。小型車15車種と普通車29車種を表9の6変数で判別する。小型車の排出量（X1）と座席数（X3）は上限が普通車以下に制限されている。小型車と普通車の排出率は、それぞれ[0.657, 0.658]と[0.996, 3.456]の範囲である。座席数は、4人と5人から8人用である。従ってX1とX3の2個の1変数モデルで線形分離可能であり、2個のBGSがあることが簡単に分かる。「P」は変数増加法で選択された変数であり、排出量（X1）、価格（X2）、座席数（X3）、CO2（X4）、燃費（X5）と販売台数（X6）の順に変数選択される。列tは2個のクラスの平均値の差の検定のt値である。LDFとQDFは、FisherのLDFとQDFのNMである。MNMは改定IP-LDFのMNMである。QDFおよび改定IP-OLDFは、1変数モデルX1で線形分離可能である。小型車の座席数は4であるため、3変数モデルのQDFのNMは、全ての普通車を小型車に誤分類する。小型車の座席数の4に小さな乱数を加えるだけで、QDFのNM=29は全て0になる。最後の2列は、RDAのNMである。2012年以前に、QDFがデータに問題を発見したときに、JMPはQDFをRDAに自動的に切り替えた。小型車の座席数は4であるため、QDFとRDAの両方が3変数モデル（X1, X2, X3）で全ての普通車を小型車に誤分類した。この事実がQDF（およびRDA）に実装された一般化逆行列法の欠陥である。この事実をJMPに指摘した後、修正RDAがリリースされたが利用者は $\lambda$ と $\gamma$ の2個のパラメータを[0, 1]の範囲で選択する必要がある。そこで11 \* 11個のグリッド探索で $\lambda = \gamma = 0.1$ が良いことが分かった。この値はデータごとに調べる必要があり、とても実用として勧められない。せっかくの最適化手法を使いながら、このような恣意的な選択を行う手法は、問題であろう。

表9 改定IP-OLDF, LDF, QDFとRDAのMNMとNMの比較

p	Var.	t	LDF	QDF	MNM	$\lambda = \gamma = 0.8$	0.1
1	Emission	11.37	2	0	0	2	0
2	Price	5.42	1	0	0	4	0
3	Capacity	8.93	1	29	0	3	0
4	CO <sub>2</sub>	4.27	1	29	0	4	0
5	Fuel	-4.00	0	29	0	5	0
6	Sales	-0.82	0	29	0	5	0

## 6.2 新手法2

表10は63個のモデルである。排出量から販売台数の6列の1 / 0は、1であればモデルにその変数を含み0なら含むないことを表す。最初の32個のモデルはX1を含んでいる。次の16個のモデルはX3を含みX1を含んでいない。最後の15個のモデルはX1とX3の両方を含んでいないので線形分離可能でないので表から省く。LDFとQDF列はNMを表す。QDFの29は乱数を加えることで0になる。この改善案を提案したが、JMPは頑として採用しない道を選んでいる。

表10 MNM=0の48個のモデル

SN	Emission	Price	Capacity	CO2	Fuel	Sales	LDF	QDF
1	1	0	1	0	0	0	2	29
2	1	1	1	0	0	0	1	29
3	1	0	1	1	0	0	1	29
4	1	0	1	0	1	0	1	29
5	1	0	1	0	0	1	2	29
6	1	1	1	1	0	0	1	29
7	1	1	1	0	1	0	1	29
8	1	1	1	0	0	1	1	29
9	1	0	1	1	1	0	1	29
10	1	0	1	1	0	1	1	29
11	1	0	1	0	1	1	1	29
12	1	1	1	1	1	0	0	29
13	1	1	1	1	0	1	0	29
14	1	1	1	0	1	1	1	29
15	1	0	1	1	1	1	1	29
16	1	1	1	1	1	1	0	29
17	1	0	0	0	0	0	2	0
18	1	1	0	0	0	0	1	0
19	1	0	0	1	0	0	1	0
20	1	0	0	0	0	1	2	0
21	1	0	0	0	1	0	2	0
22	1	1	0	1	0	0	1	0
23	1	1	0	0	1	0	1	0
24	1	1	0	0	0	1	1	0

25	1	0	0	1	1	0	1	0
26	1	0	0	1	0	1	1	0
27	1	0	0	0	1	1	4	0
28	1	1	0	1	1	0	0	0
29	1	1	0	1	0	1	1	0
30	1	1	0	0	1	1	1	0
31	1	0	0	1	1	1	2	0
32	1	1	0	1	1	1	0	0
33	0	0	1	0	0	0	0	29
34	0	1	1	0	0	0	5	29
35	0	0	1	0	1	0	3	29
36	0	0	1	0	0	1	1	29
37	0	0	1	1	0	0	0	29
38	0	1	1	1	0	0	5	29
39	0	1	1	0	1	0	6	29
40	0	1	1	0	0	1	6	29
41	0	0	1	1	1	0	3	29
42	0	0	1	0	1	1	3	29
43	0	0	1	1	0	1	1	29
44	0	1	1	1	1	0	4	29
45	0	1	1	1	0	1	5	29
46	0	1	1	0	1	1	6	29
47	0	0	1	1	1	1	4	29
48	0	1	1	1	1	1	5	29

表11の最初の6列はMPによるLDFのNMである。線形分離可能な48個のモデルで、SVM1が2個のモデルで線形分離可能でないだけで、容易に線形分離可能であることを示す。これに対して表10に示したようにLDFとQDFの判別結果が非常に悪いことが分かる。15個の線形分離可能でないモデルでは、僅かに改定IP-OLDFが優位であることが分かる。その後の7列は、改定IP-OLDの係数である。この表で新手法2をシミュレートする。フルモデル (SN = 1) を判別した場合、X1の係数のみが5.917であり、他の5個の係数がゼロであるので、自然に6次元空間を1次元の部分空間に減らすことができる。さらに判別超平面は、 $X1=0.8652$ であり、小型車の最大値と普通車の最小値の平均と等しい。このことから改定IP-OLDFは証明はできないがOCPの重心を選ぶようだ。またLSDでは、線形分離可能なモデルを先行的に選び他の変数を0にする構図がうかがわれる。6変数の大きなMatroskaには5変数から1変数の小さなMatroskaが含まれていて、X1を含む120 ( $= 5 * 4 * 3 * 2$ ) 個のMatroska製品ができる。そして(X1)がBGSである。次に、フルモデルからX1を削除し、改定IP-OLDFで5変数モデル(X2-X6)を判別すると、X3の係数のみが2で他の4変数の係数がゼロである。また判別超平面は $X3=4.5$ 台で、小型車の2席と普通車の5席の平均になる。従って、5次元空間を1次元の部分空間に減らすことができる。5次元の大きなMatroskaには4変数から1変数のBGSであるX3を含み、24 ( $= 4 * 3 * 2$ ) 個のMatroska製品ができる。5変数モデルからX3を削除し4変数モデル(X2, X4-X6)を改定IP-OLDFで判別すると4個の係数がゼロではない。従って、この

データの構造は、2個のBGSのX1とX3と他の線形分離可能でない4変数で構成されていることが理解できる。これまで10年間以上、多くの統計学者は高次元のデータを分析するために苦勞してきたが、これらの2個の変数はMatroska構造を非常に簡単に説明できる。

表11 63モデルの6個のNMと改定IP-OLDFの判別係数

SN	Var.	RIP	SVM4	SVM1	LP	IPLP	HSVM	x1	x2	x3	x4	x5	x6	C
1	1-6	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
2	1-5	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
3	1-4,6	0	0	15	0	0	0	5.92	0	0	0	0	0	-4.893
4	1-3,5,6	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
5	1,3-6	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
6	1,2,4-6	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
7	1-4	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
8	1-3,5	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
9	1-3,6	0	0	19	0	0	0	5.92	0	0	0	0	0	-4.893
10	1,3-5	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
11	1,3,4,6	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
12	1,3,5,6	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
13	1,2,4,5	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
14	1,2,4,6	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
15	1,2,5,6	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
16	1,4-6	0	0	0	0	0	0	5.92	0	0	0	0	0	-9
17	1-3	0	0	0	0	0	0	0	0	2	0	0	0	-9
18	1,3,4	0	0	0	0	0	0	0	0	2	0	0	0	-9
19	1,3,5	0	0	0	0	0	0	0	0	2	0	0	0	-4.893
20	1,3,6	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
21	1,2,4	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
22	1,2,5	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
23	1,2,6	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
24	1,4,5	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
25	1,4,6	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
26	1,5,6	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
27	1,3	0	0	0	0	0	0	0	0	2	0	0	0	-9
28	1,2	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
29	1,4	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
30	1,6	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
31	1,5	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
32	1	0	0	0	0	0	0	5.92	0	0	0	0	0	-4.893
33	2-6	0	0	0	0	0	0	0	0	2	0	0	0	-9
34	2-4	0	0	0	0	0	0	0	0	2	0	0	0	-9
35	2-4,6	0	0	0	0	0	0	0	0	2	0	0	0	-9
36	2,3,5,6	0	0	0	0	0	0	0	0	2	0	0	0	-9
37	3-6	0	0	0	0	0	0	0	0	2	0	0	0	-9
38	2-4	0	0	0	0	0	0	0	0	2	0	0	0	-9
39	2,3,5	0	0	0	0	0	0	0	0	2	0	0	0	-9
40	2,3,6	0	0	0	0	0	0	0	0	2	0	0	0	-9
41	3-5	0	0	0	0	0	0	0	0	2	0	0	0	-9
42	3,5,6	0	0	0	0	0	0	0	0	2	0	0	0	-9



43	3,4,6	0	0	0	0	0	0	0	0	2	0	0	0	-9
44	2,3	0	0	0	0	0	0	0	0	2	0	0	0	-9
45	3,5	0	0	0	0	0	0	0	0	2	0	0	0	-9
46	3,6	0	0	0	0	0	0	0	0	2	0	0	0	-9
47	3,4	0	0	0	0	0	0	0	0	2	0	0	0	-9
48	3	0	0	0	0	0	0	0	0	2	0	0	0	-9
49	2	5	6	6	6	5		0	0	0	0	0	0	-134.3
50	2,4-6	3	4	4	4	3		0	0	0	-46	-199	-0	5342.8
51	2,5,6	4	6	6	6	4		0	0	0	0	4.03	-0	-782.7
52	2,4,5	4	4	4	4	4		0	0	0	-0.3	-1.7	0	45.1
53	2,4,6	4	6	6	6	4		0	0.03	0	-121	0	-0.6	-28515
54	4-6	8	15	15	15	8		0	0	0	-96	-809	-0.2	29809
55	2,5	4	6	6	6	4		0	0.03	0	0	45.4	0	-40747
56	2,4	4	6	6	6	4		0	0	0	-0.1	0	0	-461.8
57	5,6	8	14	14	14	9		0	0	0	0	-685	-0.2	17748
58	2,6	4	6	6	6	4		0	0.03	0	0	0	0.09	-40125
59	4,5	10	12	12	11	10		0	0	0	3.54	10.8	0	-601.3
60	4,6	8	11	11	11	8		0	0	0	160	0	-0.3	-14026
61	5	10	11	11	11	10		0	0	0	0	-2.5	0	59.5
62	4	10	11	11	11	10		0	0	0	90.7	0	0	-8980
63	6	13	15	15	15	15		0	0	0	0	0	-0.7	6773.5

表12は、SVM4（左）とH-SVM（右）の係数を示す。X2、X4とX6の3つの係数が非常に小さいが、SVM4及びH-SVMのフルモデルの全ての係数はゼロでないで、これらの結果は次のことを暗示している：

- 1) SVM4及びH-SVMは、全ての遺伝子データで自然に特徴選択を行うことはできない。X2、X4およびX6の3つの係数が非常に小さいので、データが特定の条件にある場合はゼロになる可能性は否定できない。LASSOは、このような係数を見つけて、0にすることを試みる手法と理解すればよいが、完ぺきではないようだ。
- 2) 表11は、X1の係数は5.92かゼロであることを示し、X3の係数は2かゼロである。この事実は、このデータは非常に単純な構造であることを意味する。またこのようなデータでなければ、新手法2でBGSを必ず見つけることができないようだ。
- 3) 有効数字の桁数が少ないなどのデータ構造がシンプルである場合に、偶然にH-SVMとSVM4は特徴選択する能力を有する疑いがある。

表12 SVM4(左)とH-SVM(右)の判別係数

SN	x1	x2	x3	x4	x5	x6	x7	x1	x2	x3	x4	x5	x6	x7
1	0.87	-2.E-08	1.74	-1.E-02	-0.1	3.E-06	-5.1	0.63	-2.E-07	1.78	-9.E-03	-0.1	4.E-06	-6.1
2	0.87	-4.E-08	1.73	-1.E-02	-0.1	0.E+00	-5.3	0.62	-2.E-07	1.79	-4.E-03	-0	0.E+00	-7.1
3	0.89	-7.E-08	1.8	2.E-03	0	-3.E-06	-9	0.61	-2.E-07	1.79	2.E-03	0	1.E-06	-8.6
4	0.86	-1.E-07	1.77	0.E+00	-0	7.E-07	-8.3	0.62	-2.E-07	1.78	0.E+00	-0	2.E-06	-8.1
5	0.61	0.E+00	1.79	-3.E-03	-0	5.E-07	-7.9	0.61	0.E+00	1.79	-2.E-04	-0	4.E-08	-8.5
6	5.62	-1.E-06	0	-8.E-02	-0.5	3.E-05	17.1	5.62	-1.E-06	0	-8.E-02	-0.5	3.E-05	17.1

7	0.82	-2.E-07	1.73	2.E-03	0	0.E+00	-8.4	0.61	0.E+00	1.79	0.E+00	0	0.E+00	-8.6
8	0.83	-2.E-07	1.73	0.E+00	-0	0.E+00	-8	0.61	-2.E-07	1.79	0.E+00	-0	0.E+00	-8.1
9	0.81	-2.E-07	1.74	0.E+00	0	-6.E-06	-8.3	0.61	0.E+00	1.79	0.E+00	0	0.E+00	-8.6
10	0.61	0.E+00	1.79	-4.E-06	-0	0.E+00	-8.6	0.61	0.E+00	1.79	-5.E-08	-0	0.E+00	-8.6
11	0.61	0.E+00	1.79	0.E+00	0	0.E+00	-8.6	0.61	0.E+00	1.79	0.E+00	0	0.E+00	-8.6
12	0.61	0.E+00	1.79	0.E+00	0	-5.E-09	-8.6	0.61	0.E+00	1.79	0.E+00	0	0.E+00	-8.6
13	5.77	-1.E-06	0	-4.E-02	-0.4	0.E+00	8.95	5.77	-1.E-06	0	-4.E-02	-0.4	0.E+00	8.94
14	5.83	-2.E-06	0	2.E-02	0	1.E-05	-4.7	5.83	-2.E-06	0	2.E-02	0	1.E-05	-4.7
15	5.74	-2.E-06	0	0.E+00	-0.1	2.E-05	-0.3	5.74	-2.E-06	0	0.E+00	-0.1	2.E-05	-0.3
16	5.92	0.E+00	0	-4.E-04	-0	7.E-08	-4.8	5.92	0.E+00	0	-1.E-06	-0	0.E+00	-4.9
17	0.8	-2.E-07	1.74	0.E+00	0	0.E+00	-8.3	0.61	0.E+00	1.79	0.E+00	0	0.E+00	-8.6
18	0.61	0.E+00	1.79	0.E+00	0	0.E+00	-8.6	0.61	0.E+00	1.79	0.E+00	0	0.E+00	-8.6
19	0.61	0.E+00	1.79	0.E+00	-0	0.E+00	-8.6	0.61	0.E+00	1.79	0.E+00	0	0.E+00	-8.6
20	0.61	0.E+00	1.79	0.E+00	0	-2.E-08	-8.6	0.61	0.E+00	1.79	0.E+00	0	0.E+00	-8.6
21	5.9	-1.E-07	0	9.E-04	0	0.E+00	-4.9	5.9	-8.E-08	0	6.E-04	0	0.E+00	-4.9
22	5.86	-2.E-06	0	0.E+00	-0.1	0.E+00	-0.2	5.86	-2.E-06	0	0.E+00	-0.1	0.E+00	-0.2
23	5.91	-8.E-08	0	0.E+00	0	-6.E-07	-4.8	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9
24	5.92	0.E+00	0	-2.E-06	-0	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9
25	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9
26	5.92	0.E+00	0	0.E+00	-0	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9
27	0.61	0.E+00	1.79	0.E+00	0	0.E+00	-8.6	0.61	0.E+00	1.79	0.E+00	0	0.E+00	-8.6
28	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9
29	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9
30	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9
31	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9
32	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0	0.E+00	-4.9
33	0	0.E+00	2	-7.E-08	-0	0.E+00	-9	0	0.E+00	2	0.E+00	0	0.E+00	-9
34	0	4.E-09	2	-9.E-05	-0	0.E+00	-9	0	0.E+00	2	0.E+00	0	0.E+00	-9
35	0	2.E-09	2	4.E-06	0	-4.E-08	-9	0	0.E+00	2	0.E+00	0	0.E+00	-9
36	0	2.E-07	2.1	0.E+00	-0	-2.E-06	-9.6	0	0.E+00	2	0.E+00	0	0.E+00	-9
37	0	0.E+00	2	7.E-09	-0	0.E+00	-9	0	0.E+00	2	0.E+00	0	0.E+00	-9
38	0	1.E-08	2.01	3.E-05	0	0.E+00	-9	0	0.E+00	2	0.E+00	0	0.E+00	-9
39	0	1.E-08	2.01	0.E+00	-0	0.E+00	-9	0	0.E+00	2	0.E+00	0	0.E+00	-9
40	0	3.E-07	2.17	0.E+00	0	-6.E-06	-10	0	0.E+00	2	0.E+00	0	0.E+00	-9
41	0	0.E+00	2	3.E-07	0	0.E+00	-9	0	0.E+00	2	0.E+00	0	0.E+00	-9
42	0	0.E+00	2	0.E+00	-0	0.E+00	-9	0	0.E+00	2	0.E+00	0	0.E+00	-9
43	0	0.E+00	2	4.E-07	0	0.E+00	-9	0	0.E+00	2	0.E+00	0	0.E+00	-9
44	0	4.E-07	2.23	0.E+00	0	0.E+00	-11	0	0.E+00	2	0.E+00	0	0.E+00	-9
45	0	0.E+00	2	0.E+00	-0	0.E+00	-9	0	0.E+00	2	0.E+00	0	0.E+00	-9
46	0	0.E+00	2	0.E+00	0	-1.E-09	-9	0	0.E+00	2	0.E+00	0	0.E+00	-9
47	0	0.E+00	2	2.E-07	0	0.E+00	-9	0	0.E+00	2	0.E+00	0	0.E+00	-9
48	0	0.E+00	2	0.E+00	0	0.E+00	-9	0	0.E+00	2	0.E+00	0	0.E+00	-9
49	0	6.E-06	0	0.E+00	0	0.E+00	-8.6							
50	0	6.E-06	0	-1.E-01	-0.8	-2.E-05	25.9							
51	0	5.E-06	0	0.E+00	-0.1	8.E-06	-5.9							
52	0	6.E-06	0	-2.E-01	-0.9	0.E+00	27.9							
53	0	5.E-06	0	1.E-02	0	6.E-06	-8.9							
54	0	0.E+00	0	-9.E-03	-0.2	1.E-04	6.2							
55	0	6.E-06	0	0.E+00	-0.1	0.E+00	-6.2							
56	0	6.E-06	0	1.E-02	0	0.E+00	-9.1							
57	0	0.E+00	0	0.E+00	-0.2	1.E-04	4.31							

58	0	6.E-06	0	0.E+00	0	-3.E-05	-7.8
59	0	0.E+00	0	0.E+00	-0.2	0.E+00	5
60	0	0.E+00	0	4.E-02	0	1.E-04	-5
61	0	0.E+00	0	0.E+00	-0.2	0.E+00	5
62	0	0.E+00	0	5.E-02	0	0.E+00	-5
63	0	0.E+00	0	0.E+00	0	0.E+00	1

表13は、改定LP-OLDF（左）と改定IPLP-OLD（右）の係数を示す。フルモデル（SN = 1）を判別すると、X3の係数がゼロである。次にSN = 6の5変数モデル（X1, X2, X4-X6）を判別するとX3を除く5個の係数がゼロではないので停止する。結局これらの判別関数は、新手法2で今後検討する必要が認められない。しかし、X1を省くと16個のモデル全てで $f=2 \times X3-9$ という判別関数になる。判別超平面は $X3=4.5$ になる。2群の分散共分散を考えないで、座席数が4と[5, 8]であるので、小型車と普通車は4.5台で容易に判別できるという単純な事実を示している。あえてその事実に注目しないで6変数で判別すれば、LDFとQDFはもっともらしい統計量を出力し最もらしいレポートが書ける。この議論は判別分析では牽強付会に感じられるが、重回帰分析を行えば牽強付会とは言えない。この時、最初からX1とX3を重回帰分析から省くべきか否かはすぐに結論を出せない問題である。

表13 改定LP-OLDF（左）と改定IPLP-OLDF（右）の判別係数

SN	x1	x2	x3	x4	x5	x6	x7	x1	x2	x3	x4	x5	x6	x7
1	6	-3.E-08	0	-2.E-02	-1.E-01	-2.E-05	0	6	-3.E-08	0	-2.E-02	-1.E-01	-2.E-05	0
2	5.86	-6.E-07	0	-1.E-02	-1.E-01	0.E+00	0	5.86	-6.E-07	0	-1.E-02	-1.E-01	0.E+00	0
3	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9
4	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9
5	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9
6	6	-3.E-08	0	-2.E-02	-1.E-01	-2.E-05	0	6	-3.E-08	0	-2.E-02	-1.E-01	-2.E-05	0
7	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9
8	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9
9	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9
10	13	0.E+00	-2.4	0.E+00	0.E+00	0.E+00	0	5.93	0.E+00	0	-2.E-02	-1.E-01	0.E+00	0
11	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9
12	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.9	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9
13	5.86	-6.E-07	0	-1.E-02	-1.E-01	0.E+00	0	5.86	-6.E-07	0	-1.E-02	-1.E-01	0.E+00	0
14	6.08	-2.E-06	0	-4.E-03	0.E+00	-8.E-06	-2.5	6.08	-2.E-06	0	-4.E-03	0.E+00	-8.E-06	-2.5256
15	5.74	-2.E-06	0	0.E+00	-1.E-01	2.E-05	-0.3	5.74	-2.E-06	0	0.E+00	-1.E-01	2.E-05	-0.3388
16	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.8935
17	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9
18	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.8935
19	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9
20	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.9	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9
21	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.8935
22	5.86	-2.E-06	0	0.E+00	-1.E-01	0.E+00	-0.2	5.86	-2.E-06	0	0.E+00	-1.E-01	0.E+00	-0.2061
23	6.04	-2.E-06	0	0.E+00	0.E+00	-5.E-06	-2.9	6.04	-2.E-06	0	0.E+00	0.E+00	-5.E-06	-2.9004

24	6.38	0.E+00	0	-3.E-02	-1.E-01	0.E+00	0	5.93	0.E+00	0	-2.E-02	-1.E-01	0.E+00	0
25	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.8935
26	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.8935
27	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9	0	0.E+00	2	0.E+00	0.E+00	0.E+00	-9
28	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.8935
29	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.8935
30	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.8935
31	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.8935
32	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.9	5.92	0.E+00	0	0.E+00	0.E+00	0.E+00	-4.8935
33	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
34	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
35	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
36	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
37	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
38	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
39	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
40	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
41	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
42	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
43	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
44	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
45	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
46	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
47	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
48	0	0	2	0	0	0	-9	0	0	2	0	0	0	-9
49	0	6E-06	0	0	0	0	-8.6	0	0.0286	0	0	0	0	-38570
50	0	6E-06	0	-0.1492	-0.8023	-2E-05	25.9	0	0.0368	0	-495	-2140	-0.408	54848
51	0	5E-06	0	0	-0.0727	8.4E-06	-5.9	0	0.0005	0	0	4.0305	-0.003	-783
52	0	6E-06	0	-0.1559	-0.8729	0	27.9	0	0.0005	0	-6.705	-27.8	0	711
53	0	5E-06	0	0.01364	0	6.5E-06	-8.9	0	0.0292	0	8.404	0	0.1545	-40783
54	0	0	0	-0.0089	-0.2475	0.00013	6.2	0	0	0	-1.139	-27.98	0.0215	590
55	0	6E-06	0	0	-0.069	0	-6.2	0	0.0303	0	0	75.742	0	-41871
56	0	6E-06	0	0.0139	0	0	-9.1	0	0.0004	0	-0.146	0	0	-462
57	0	0	0	0	-0.2058	0.00013	4.31	0	0	0	0	-17.36	0.0175	291
58	0	6E-06	0	0	0	-3E-05	-7.8	0	0.0009	0	0	0	0.0026	-1152
59	0	0	0	0	-0.2222	0	5	0	0	0	-200.9	-1788	0	62557
60	0	0	0	0.04364	0	0.00011	-5	0	0	0	160.48	0	-0.249	-14169
61	0	0	0	0	-0.2222	0	5	0	0	0	0	-443.6	0	9848
62	0	0	0	0.04651	0	0	-5	0	0	0	175.44	0	0	-16843
63	0	0	0	0	0	0	1	0	0	0	0	0	0	1

### 6.3 手法1

日本車データからリサンプリング標本を生成し100重交差検証法で8個のLDFを評価する。表9でみたようにX4からX6は判別に重要ではないので、X1, X2, X3の3変数による7個のモデルを調べる。X1またはX3含む6個の線形分離可能なモデルと線形分離可能でない1変数モデル(X2)に分かれる。表14は新手法1の結果を示す。最初の7行は改定IP-OLDF(RIP)の7個の判別モデルである。「モデル」欄は独立変数を示す。M1とM2は、学習および検証標本の

平均誤分類確率である。改定LP-OLDFの6個のM2はゼロである。改定IP-OLDF, 改定IPLP-OLDFの5個のM2はゼロである。H-SVMとSVM4の2個のM2はゼロである。SVM1の1個のM2はゼロである。FisherのLDFのすべてのM2はゼロでない。この結果で、7個のLDFを評価することができる。M2に注目すると、改定IP-OLDF, 改定LP-OLDFだけが6モデルとも0である。モデル選択の原理原則は、「オッカムのカミソリ」あるいは「ケチの原理」であるので、5番と6番の1変数モデルを選ぶことになる。5番目の7個のLDFと改定IP-OLDFのM2の差をとると、0, 0, 0.84, 0, 0, 0, 6.09%であり、FisherのLDFだけが非常に悪いことが分かる。

表14 8個のLDFのM1とM2

MNM		M1	M2	Diff2	モデル	
1m28s	1	0.00	0.00	0.00	X1, X2, X3	
	2	0.00	0.00	0.00	X1, X2	
	3	0.00	0.07	0.07	X1, X3	
	4	0.00	0.00	0.00	X2, X3	
	5	0.00	<u>0.00</u>	0.00	X1	
	6	0.00	<u>0.00</u>	0.00	X3	
	7	9.55	12.68	3.14	X2	
HSVM		M1	M2	Diff2	M1Diff	M2Diff
1m11s	1	0.000	0.114	0.11	0.000	0.11
	2	0.000	0.205	0.20	0.000	0.20
	3	0.000	0.000	0.00	0.000	-0.07
	4	0.000	0.114	0.11	0.000	0.11
	5	0.000	0.000	0.00	0.000	0.00
	6	0.000	0.000	0.00	0.000	0.00
SVM4		M1	M2	Diff2	M1Diff	M2Diff
59s	1	0.000	0.114	0.11	0.000	0.11
	2	0.000	0.227	0.23	0.000	0.23
	3	0.000	0.000	0.00	0.000	-0.07
	4	0.000	0.114	0.11	0.000	0.11
	5	0.000	0.000	0.00	0.000	0.00
	6	0.000	0.000	0.00	0.000	0.00
7	12.386	12.977	0.59	2.841	0.30	
SVM1		M1	M2	Diff2	M1Diff	M2Diff
1m	1	1.48	1.59	0.11	1.477	1.59
	2	2.02	2.68	0.66	2.023	2.68
	3	0.00	0.00	0.00	0.000	-0.07
	4	0.00	0.16	0.16	0.000	0.16
	5	0.73	0.84	0.11	0.727	0.84
	6	0.00	0.00	0.00	0.000	0.00
7	12.39	12.98	0.59	2.841	0.30	
IPLP		M1	M2	Diff2	M1Diff	M2Diff
59s	1	0.00	0.27	0.27	0.000	0.27
	2	0.00	0.00	0.00	0.000	0.00
	3	0.00	0.00	0.00	0.000	-0.07
	4	0.00	0.00	0.00	0.000	0.00

5		0.00	0.00	0.00	0.000	0.00
6		0.00	0.00	0.00	0.000	0.00
7		9.55	12.66	3.11	0.000	-0.02
LP		M1	M2	Diff2	M1Diff	M2Diff
42s	1	0.00	0.00	0.00	0.000	0.00
	2	0.00	0.00	0.00	0.000	0.00
	3	0.00	0.00	0.00	0.000	-0.07
	4	0.00	0.00	0.00	0.000	0.00
	5	0.00	0.00	0.00	0.000	0.00
	6	0.00	0.00	0.00	0.000	0.00
7		12.39	12.98	0.59	2.841	0.30
logistic		M1	M2	Diff2	M1Diff	M2Diff
6m	1	0.00	0.36	0.00	0.000	0.36
	2	0.00	0.05	0.00	0.000	0.05
	3	0.00	0.02	0.00	0.000	-0.05
	4	0.00	0.00	0.00	0.000	0.00
	5	0.00	0.00	0.00	0.000	0.00
	6	0.00	0.00	0.00	0.000	0.00
7		12.30	13.01	0.00	2.750	0.33
LDF		M1	M2	Diff2	M1Diff	M2Diff
7m	1	1.500	2.349	0.00	1.500	2.35
	2	1.886	2.913	0.00	1.886	2.91
	3	4.523	4.750	0.00	4.523	4.68
	4	10.432	12.828	0.00	10.432	12.83
	5	5.364	5.742	0.00	5.364	5.74
	6	4.705	6.089	0.00	4.705	6.09
7		26.886	27.027	0.00	17.341	14.34

## 7. 終わりに

1997年から2010年に、筆者は判別分析の新しい理論を確立した。改定IP-OLDFは問題1と問題2を解決した。IP-OLDFは判別分析の新たな事実を示した。問題4を解決するために新手法1を提案した。小標本の研究データを分析したい研究者には新手法1は非常に簡単で強力である。他の判別関数と改定IP-OLDFを比較し優位性を示すことができた。この利点に加えて、平均誤分類確率で最適なモデル選択が簡単に行える。さらにFisherは判別係数および誤分類確率のためのSEの式を定式化していないが判別係数の95% CIが求まる。2010年から応用研究として、試験の合否判定、スイス銀行紙幣データと日本車データを用いてLSDの判別に焦点を当てた。しかし2015年に成果を得て終了したがインパクトが弱かった。2015年10月の終わりに6個のMicroarrayデータで改定IP-OLDFが自然に特徴選択を行うことが分かった。そこで、「MicroarrayデータのためのMatroska特徴選択法」の新手法2を開発した。この手法は、Microarrayデータが線形分離可能な幾つかの小さなMatroskaと、線形分離可能でない高次元の遺伝子空間で構成されていることを明らかにした。多くの研究者は高次元遺伝子空間の解析に四苦八苦してきたが(問題5)、新手法2で通常の統計的手法を用いて簡単に分析できる

ので問題5は簡単に解決できる。

MPを用いた判別分析の研究は筆者が研究を開始した1997年以前に数多く行われていたが [10,15], Stamの総括論文[56]で終焉したことを知らなかった。それに代わって, Vapnikが SVMという新しい衣を被って登場したが, ORや統計の研究分野でなくパターン認識の分野を選んだのは賢明であった。また筆者が事前に先行研究を調査するタイプの研究者であれば, MPを用いた判別分析の研究を行わなかったであろう。

(成蹊大学経済学部教授)

## References

1. Alon, A. et al. (1999). "Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays." *Proc. Natl. Acad. Sci. USA*, 96, 6745-6750.
2. Chiaretti, S. et al. (2004). "Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival." *Blood*. April 1, 2004, 103/7, pp. 2771-2778.
3. Edger, A. (1935). "The irises of the Gaspé Peninsula." *Bulletin of the American Iris Society*, 59, 2-5.
4. Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic problems." *Annals of Eugenics*, 7, 179-188.
5. \_\_\_\_\_ (1956). *Statistical methods and statistical inference*. Hafner Publishing Co.
6. Flury, B. and Riedel, H. (1988). *Multivariate Statistics: A Practical Approach*. Cambridge University Press.
7. Friedman, J. H. (1989). "Regularized Discriminant Analysis." *Journal of the American Statistical Association*, 84/405, 165-175.
8. Golub, T.R. et al. (1999). "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring." *Science*. 1999 Oct 15; 286 (5439) : pp. 531-537.
9. Goodnight, J. H. (1978). "SAS Technical Report – The Sweep Operator: Its Importance in Statistical Computing – (R-100) ." *SAS Institute Inc. Technical Report*.
10. Glover, F. (1990). "Improved linear programming models for discriminant analysis." *Decision Sciences*, 21, 771-785.
11. Jeffery, I.B. Higgins, D.G. and Culhane, A.C. (2006). "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data." *BMC Bioinformatics*. Jul

- 26; pp. 7:359. <http://www.bioinf.ucd.ie/people/ian/>
12. Lachenbruch, P. A. and Mickey, M. R. (1968). "Estimation of error rates in discriminant analysis." *Technometrics*, 10, 1-11.
  13. Miyake, A. and Shinmura, S. (1976). *Error rate of linear discriminant function*, F.T. de Dombal & F. Gremy editors 435 - 445, North-Holland Publishing Company.
  14. \_\_\_\_\_ (1979). "An algorithm for the optimal linear discriminant functions." *Proceedings of the International Conference on Cybernetics and Society*, 1447-1450.
  15. Rubin, P. A. (1997). "Solving mixed integer classification problems by decomposition." *Annals of Operations Research*, 74, 51-64.
  16. Sall, J. P. Creighton, L. and Lehman, A. (2004). *JMP Start Statistics, Third Edition*. SAS Institute Inc. (Shinmura, S. edited Japanese version)
  17. Schrage, L. (2006). *Optimization Modeling with LINGO*. LINDO Systems Inc. (Shinmura, S. translated Japanese version)
  18. \* Shinmura, S. and Miyake, A. (1979). "Optimal linear discriminant functions and their application." *COMPSAC*, 79, 167-172.
  19. \* Shinmura, S. (2000a). "A new algorithm of the linear discriminant function using integer programming." *New Trends in Probability and Statistics*, 5, 133-142.
  20. \* \_\_\_\_\_ (2000b). "Optimal Linear Discriminant Function using Mathematical Programming." *Dissertation*, March 200, 1-101, Okayama Univ.
  21. \* \_\_\_\_\_ (2003). "Enhanced Algorithm of IP-OLDF." *ISI2003 CD-ROM*, 428-429.
  22. \* \_\_\_\_\_ (2004). "New Algorithm of Discriminant Analysis using Integer Programming." *IPSI 2004 Pescara VIP Conference CD-ROM*, 1-18.
  23. \* \_\_\_\_\_ (2005). "New Age of Discriminant Analysis by IP-OLDF –Beyond Fisher's Linear Discriminant Functions." *ISI2005*, 1-2.
  24. \* \_\_\_\_\_ (2007b). "Comparison of Revised IP-OLDF and SVM." *ISI2009*, 1-4.
  25. \* \_\_\_\_\_ (2009). "Practical discriminant analysis by IP-OLDF and IPLP-OLDF." *IPSI 2009 Belgrade VIPSI Conference CD-ROM*, 1-17.
  26. \* \_\_\_\_\_ (2011b). "Beyond Fisher's Linear Discriminant Analysis - New World of Discriminant Analysis -." *ISI2011 CD-ROM*, 1-6.
  27. \* \_\_\_\_\_ (2013). "Evaluation of Optimal Linear Discriminant Function by 100-fold Cross Validation." *ISI2013 CD-ROM*, 1-6.
  28. \* \_\_\_\_\_ (2014a). "End of Discriminant Functions based on Variance-Covariance Matrices." *ICORES*, 5-14.



29. \*\_\_\_\_\_ (2014b). "Improvement of CPU time of Linear Discriminant Function." *Statistics, Optimization and Information Computing*, vol. 2, 114-129.
30. \*\_\_\_\_\_ (2014c). "Comparison of Linear Discriminant Functions by K-fold Cross Validation." *Data Analytics 2014*, 1-6.
31. \*\_\_\_\_\_ (2015a). "The 95% confidence intervals of error rates and discriminant coefficients." *Statistics, Optimization and Information Computing*, vol. 3, 66-78.
32. \*\_\_\_\_\_ (2015b). "Four Serious problems and New Facts of the Discriminant Analysis." E. Pinson et al. (Eds.) *ICORES 2014 Revised and Selected Papers*, CCIS 509, 15-30, Springer.
33. \*\_\_\_\_\_ (2015c). "A Trivial Linear Discriminant Function." *Statistics, Optimization, and Information Computing*, Vol.3, December 2015, 322-335. DOI: 10.19139/soic.20151202.
34. \*\_\_\_\_\_ (2015d). "The Discrimination of the microarray data (Ver. 1) ." *Research Gate Free Paper (1)* , Oct. 28, 2015, 1-4.
35. \*\_\_\_\_\_ (2015e). "Feature Selection of three Microarray data." *Research Gate Free Paper (2)* , Nov.1, 2015, 1-7.
36. \*\_\_\_\_\_ (2015f). "Feature Selection of Microarray Data (3) – Shipp et al. Microarray Data." *Research Gate Free Paper (3)* , 2015, 1-11.
37. \*\_\_\_\_\_ (2015g). "Validation of Feature Selection (4) – Alon et al. Microarray Data." *Research Gate Free Paper (4)* , 2015, 1-11.
38. \*\_\_\_\_\_ (2015h). "Repeated Feature Selection Method for Microarray Data (5) ." *Research Gate Free Paper (5)* , Nov. 9, 2015, 1-12.
39. \*\_\_\_\_\_ (2015i). "Comparison Fisher's LDF by JMP and Revised IP-OLDF by LINGO for Microarray Data (6) ." *Research Gate Free Paper (6)* , Nov. 11, 2015, 1-10.
40. \*\_\_\_\_\_ (2015j). "Matroska Trap of Feature Selection Method (7) –Golub et al. Microarray Data." *Research Gate Free Paper (7)* . Nov. 18, 2015, 1-14.
41. \*\_\_\_\_\_ (2015k). "Minimum Sets of Genes of Golub et al. Microarray Data (8) ." *Research Gate Free Paper (8)* , Nov. 22, 2015, 1-12.
42. \*\_\_\_\_\_ (2015l). "Complete Lists of Small Matroska in Shipp et al. Microarray Data (9) ." *Research Gate Free Paper (9)* , Dec. 4, 2015,1-81.
43. \*\_\_\_\_\_ (2015m). "Sixty-nine Small Matroska in Golub et al. Microarray Data (10) ." *Research Gate (10)* , Dec. 4, 1-58.
44. \*\_\_\_\_\_ (2015n). "Simple Structure of Alon et al. Microarray Data (11) ." *Research Gate Free Paper (11)* , Dec. 4, 2015, 1-34.
45. \*\_\_\_\_\_ (2015o). "Feature Selection of Singh et al. Microarray Data (12) ." *Research Gate*

- Free Paper (12)*, Dec. 6, 2015, 1-89.
46. \* \_\_\_\_\_ (2015p). "Final List of Small Matroska in Tian et al. Microarray Data." *Research Gate Free Paper (13)*, Dec. 7, 1-160.
  47. \* \_\_\_\_\_ (2015q). "Final List of Small Matroska in Chiaretti et al. Microarray Data." *Research Gate Free Paper (14)*, Dec. 20, 2015, 1-16.
  48. \* \_\_\_\_\_ (2015r). "Matroska Feature Selection Methods for Microarray Data." *Research Gate Free Paper (15)*, 1-16.
  49. \* \_\_\_\_\_ (2016a). "Matroska Feature Selection Method for Microarray Data." *Biotechno 2016*, 1-6.
  50. \_\_\_\_\_ (2016b). *New Theory of Discriminant Analysis after R. Fisher*. Springer (unpublished).
  51. \* \_\_\_\_\_ (2016c). "The Best Model of the Swiss Banknote Data-Validation by the 95% CI of error rates and discriminant coefficients -." *Statistics, Optimization, and Information Computing*, Vol.3, 322-335, 2015. DOI: 10.19139/soic.20151202.
  52. \* \_\_\_\_\_ (2016d). "Discriminant Analysis of the Linear Separable Data." *Journal of Statistical Science and Application, 2016*, (in Press) .
  53. \* \_\_\_\_\_ (2016e). "The Discriminant Analysis of the Iris Data by New Theory." *Data Analytic 2016*, 1-6. (unpublished).
  54. Shipp, M.A. et al. (2002). "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning." *Nature Medicine*, 8, 68-74.
  55. Singh, D. et al. (2002). "Gene expression correlates of clinical prostate cancer behavior." *Cancer Cell: March 2002*, Vol.1, 203-209.
  56. Stam, A. (1997). "Nontraditional approaches to statistical classification: Some perspectives on lp-norm methods." *Annals of Operations Research*, 74, 1-36.
  57. Taguchi, G. and Jugulum, R. (2002). *The Mahalanobis-Taguchi Strategy – A Pattern Technology System*. John Wiley & Sons.
  58. Tian, E. et al. (2003). "The Role of the Wnt-Signaling Antagonist DKK1 in the Development of Osteolytic Lesions in Multiple Myeloma." *The New England Journal of Medicine*, Vol. 349, 26, 2483-2494.
  59. Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.
  60. 小西貞則, 本田正幸 (1992). 「判別分析における誤判別率推定とブートストラップ法」. 『応用統計学, 21/2』, 67-100.
  61. 清水忠彦, 常俊義三, 河野洋, 新村秀一 (1975). 「光化学スモッグによる自覚症状の分類 (共著)」. 『大気汚染研究』, 9 (4) 734-741.

62. 新村秀一, 北川護, 高木義人, 野村裕, (1973). 「2段階重みづけによるスペクトル診断」. 『第12回日本ME学会大会論文集』, 107-108.
63. 新村秀一, 北川護, 野村裕, (1974). 「スペクトル診断 (第2報)」. 『第13回日本ME学会大会論文集』, 414-415.
64. 新村秀一 (1984). 「医療データ解析, モデル主義そしてOR」. 『オペレーションズ・リサーチ, 29/7』, 415-421.
65. 新村秀一訳著 (1986). 『SASによる回帰分析の実践』. 朝倉書店。
66. 新村秀一 (1996). 「重回帰分析と判別分析のモデル決定 (2) : 19変数をもつC.P.D.データのモデル決定」. 『成蹊大学経済学部論集, 27/1』, 180-203.
67. \_\_\_\_\_ (1998). 「数理計画法を用いた最適線形判別関数」. 『計算機統計学, 11/2』, 89-101.
68. 新村秀一, 垂水共之 (2000). 「乱数データを用いた最適線形判別関数の評価」. 『計算機統計学, 12/2』, 107-123.
69. 新村秀一 (2007a). 「改定IP-OLDFによるIP-OLDFの問題点の解消」. 『計算機統計学, 19/1』, 1-16.
70. \_\_\_\_\_ (2007b). 「数理計画法による判別分析の10年」. 『計算機統計学, 20/1 & 2』, 59-94.
71. \_\_\_\_\_ (2010a). 『最適線形判別関数』. 日科技連出版.
72. \_\_\_\_\_ (2010b). 「線形計画法による改定IP-OLDFの計算時間の改善」. 『計算機統計学, 22/1』, 37-57.
73. \_\_\_\_\_ (2011a). 「合否判定データによる判別分析の問題点」. 『応用統計学, 40/3』, 157-172.
74. \_\_\_\_\_ (2011b). 『数理計画法による問題解決法』. 日科技連出版.
75. \* \_\_\_\_\_ (2012). 「コラム SAS/JMPとの歩み」, 『SAS Technical News, 春, 夏, 秋, 冬号』.
76. \_\_\_\_\_ (2015). 「いかに研究成果を世界に発信するか—判別分析の4つの問題と新事実—」. 『SASユーザー会』, 484 - 493.
77. 三宅章彦, 新村秀一 (1980). 「最適線形判別関数のアルゴリズムとその応用」, 『医用電子と生体工学, 18/6』, 452-454.

Researchers can download author's papers with \* before author's name from the Research Gate.