R. A. Fisher以後の判別分析の新理論（1）
－遺伝子解析の新手法2（LINGO Program3）の検証－

新　村　秀　一

1. はじめに

新村（2010a, 2011b, 2012, 2016b）の一生の研究テーマは，既存の統計的判別分析の理論に瑕疵があり，それを克服することであった。Fisher（1936, 1956）が確立した分散共分散行列に基ずく線形判別関数（FisherのLDF）と，それ以降に応用発展してきた2次判別関数（QD），正則化判別分析（RDA）（Friedman, 1989）やLASSO（Simin et al., 2013）などの統計的判別分析の主流をなす理論は，Fisherも言うとおりFisherの仮説を満たさないデータに適用してはいけない（田邉, 2011）。しかしFisherの仮説を満たすか満たさないかの検定をFisherは提案しなかった。そのため，多くの分野に適用され研究や産業応用に広く利用され実績をあげてきた。しかし，医学診断や，成績の合否判定を得点合計で行う判別や，各種の格付けのように，判別超平面の近傍に多くのケースが来るような（新村, 1984），明らかにFisherの仮説を満たさないデータに分散共分散行列に基づく判別関数を適用するのは問題である。医学研究で最小誤分類数（MNM）（Miyake and Shinmura, 1979）が0であっても，これらの判別関数の誤分類確率が10％以上もあれば，研究を中断した例も過去にあった思われる。この深刻な事態には救いがある。研究を見直せば，必ず判別結果（誤分類数あるいはそれから計算された誤分類確率（Miyake and Shinmura, 1976））が改善され，悪くならない点である。もし，それほど改善されなければ，研究用に判別超平面の近傍を避け，遠くにある典型例を多く含む研究データが集められたと考えられる。以上のことが多くの判別分析の利用者に知られていないのは，我々統計家の怠慢であろう。あるいは統計家自身が，判別分析の怖さを知らないと考えるべきであろう。この点に関しては，事実を整理して別途検討したい。

昨年末に膨大な実証研究を踏まえてFisher以降の新しい判別理論の確立に成功し，Springerから“New Theory of Disciminant Analysis after R.Fisher”を10月に出版予定である。しかし本来であれば，Cox（1958）がそれを主張するのが一番最適切であったと考える[1]。彼はFisherの仮説を前提としないで，医学診断に用いられるCox回帰やロジスティック回帰などの第2世

[1] また，Glover（1980），Rubin（1997）らがORで数理計画法を用いて種々の判別モデルを提案し，Stam（1997）が米国のOR学会誌の総括論文で幕を引いたが，時期を同じくしてVapnik（1995）が3種類のSupport Vector Machine（SVM）で新世界を開いた。彼は，その意味でCoxに次ぐ，第三世代のFisherの後継者といえる。そして，統計やORでSVMを紹介せず，パターン認識という理解のある研究分野で普及を図ったのは賢明である。

代の判別分析を確立した。FisherやPearsonの誰もが認める後継者である。しかし，Coxはその点を多分明確に指摘しなかった。Fisherの仮説を満たさない場合には，QDが提案されたことに注意が払われず，Fisherの正統な後継者は，Fisherの仮説や，正規分布や，分散共分散行列を不磨の大典として現実を軽視して数式の正確な展開に基づいて統計的判別分析の理論を発展させたと考えている。主流派が，正則化判別関数やLASSOなどの手法を開発した成果はもちろん重要である。しかし，Fisherの著書は哲学的で難解であるが，当初ロザムステッド農場の研究者の職しか得られず，現実の細かい問題にも関心を持っていたことは確かである。そもそもFisherのLDFの検証には，Fisherのアヤメのデータ（Anderson, 1935）として知られている具体的な現実のデータを評価に用いている事実に注目すべきであろう。統計やORの研究者は，数学にあこがれがある。しかし純粋理論は，純粋数学の論理的な世界でしか成立せず，それ以外の学問は現実を観測し，実証研究を行う必要がある。純粋理論という言葉にあこがれず，役に立つ応用理論を開発すべきと考えている。実証研究を行えば，どんな素晴らしく思える理論にも瑕疵があり，研究テーマは簡単に見つけることができる。今日影が薄くなったが，何か検証する必要があるとき，正規乱数などを検証データに用いて評価することが要求された。しかし「正規分布を用いた検証は，正規分布を仮定した理論の裏付けにしかならない」という簡単な事実が理解されていない。この点を十分気を付けて実証研究すべきであろう。筆者はFisherの意図を組み，Andersonが集めたFisherのアイリスデータ，少なくとも3個の多重共線性を持つCPDデータ（Shinmura and Miyake, 1979；新村, 1996；三宅＆新村, 1980），2変数の正規乱数データ（新村＆垂水, 2000），スイス銀行紙幣データ（Flury & Riedel, 1988），学生データ（新村, 2004；2007b；2010a），試験の合否判定データ（新村, 2011a），日本車データ（Shinmura, 2016b）などの現実のデータを用いて8年間実証研究を行い，既存の判別分析が抱える問題に気付いて「小標本のための100重交差検証法（新手法1）（Shinmura, 2014c；2015a-2015c）」と「Best modelによる変数選択法（Shinmura, 2016c）」等の解決策を考え，新しい判別分析の理論（Theory）[2]を確立した。しかし2015年の10月から僅か41日間で，6種類のMicroarrayデータをまず手作業で分析し，その成果を踏まえて「Matroska Feature Selection Method（Method2）」を確立した（Shinmura, 2015d-2015r）。手作業では，分析の継続が困難でまた単純ミスも発生するのでLINGO（Schrage, 2006）で分析プログラム（Program3[3]）を作成し，リサーチゲート（Research Gate）に15編の論文を，人生初めてのフリ

[2] 3種類の最適線形判別関数の改定IP-OLDF（Shinmura, 2007b；2009；2011；2013），改定LP-OLDF，改定IPLP-OLDF（Shinmura, 2014b）を開発した。MNM基準に基づくIP-OLDFは判別分析の重要な2個の新知見を見つけた。改定IP-OLDFと「小標本のための100重交差検証法（Method1）」で，判別分析の4個の深刻な問題を解決した。

[3] 通常のデータの学習標本による6種類のMP-BASED LDFsを実行するプログラムをProgram1とし，Method1で学習標本と検証標本の誤分類確率と判別係数の95％信頼区間を6種類のMP-BASED LDFsで

ーペーパーとして発表した。効果は深く考えていなかったが，6種類のデータを提供するHPを作成し，自らそれを用いてFeature Selectionの論文を発表しているJefry（2006）から早い時期にメールが来たこと，この分野の研究を活発に行っているとみられるカルフォルニア大学のリーダーがTamayoという教授であることが，私の論文の閲覧者として彼の顔写真付きで分かったことなどが大きな成果である。今回用いている6種のデータのうち，Golubら（1999），Shippら（2002）とSinghら（2002）の指導教授であるようだ。また，2015年12月13日時点では2011のRead数と1199の閲覧数と引用文献数が392であったのが，2016年8月21日時点では3237のRead数と2854閲覧数と引用文献数が657になった[4]。そして，2016年の6月にBio関連の国際会議（6月25日から29日）（Shinmura, 2016a[5]）と7月8日㈮と9日㈯に千葉大学の計測自動学会での発表（新村, 2016a）を行い，並行してSpringerから本を出すことにした。

2. Matroska Feature Selection手法（Method2）開発の経緯

2015年中旬に判別分析の4つの問題（Shinmura, 2024a；2015b）の最後まで残っていた問題4に関して，判別係数の95％信頼区間（Confidence Interval, CI）とモデル選択（変数選択，特徴抽出ともいう）（Shinmura, 2016c；2016d）が解決でき，統計学の泰斗のR.A.Fisherが作り上げた統計的判別分析を完全に刷新した判別分析の新理論が完成したと確信した。そして富山市で開催された統計のシンポジューム（2015年10月24日～ 25日）でその成果を発表した。そこで，発表翌日最終日の日曜日に筑波大学の大学院生がMicroarrayデータ（Datasets）を用いた主成分分析（Principal Component Analysis, PCA）の分析例を発表しているのを聞いて，世界的に著名なDatasetsが公開されていることを知った。彼女に分析に用いたDatasetsを入手できるHPのURLを送ってもらうことを依頼し，翌々日の10月28日にメールを受け取り6種のDatasetsをダウンロードした。Alon et al.（1999），Golub et al.とShipp et al.の3種のDatasets

---

実行するプログラムをProgram2として求めるプログラムは，Shinmura（2016f）に公開してある。

[4] 一生の研究環境を整備するのに，RGに研究論文や資料をUPし，世界の研究者と交流を深めることをぜひ若手研究者に勧めたい。資料をPDF化しUPするだけで, それ以上の研究に便利な情報が得られる。筆者は先行研究や類似研究の論文調査の手を抜いてきたが，RGが勝手に情報を提供してくれるので，他人の論文も見るようになった。

[5] 若手研究者におすすめの第2点は，6頁以上のレギュラーペーパーとかフルペーパー・セッションをうたう国際会議が増えてきたことである。5－6人のレフリーに採択され，問題点を訂正してCamera-Readyペーパーを出せば，多くの会議は会議後にDigital Libraryをインターネットに公開している。ICORESという国際会議に初めて参加し，その後に，Springerから内容を⅔以上書き換えた10頁まで無料で書くことができる論文の権利を得た。伝統的な学術誌で，査読者とのやり取りに時間を費やすより効果的である。今回，Biothechnoの国際会議で発表したが聴衆が少なくてがっかりしたが，8月16日にBest Paper Award受賞の通知が来た。これを得ることを目的で発表してきたのは，24頁までの論文をIARIAジャーナルに2017年に投稿できる権利が得られるからである。しかし，参加者は発表が採択された関係者が大半で参加者が少ないために参加料が8万円から10万円前後と高額である。もったいないので，2件の発表を申し込み受理されたが，参加料は発表ごとにとられる場合が多いのでキャンセルするのに手間がかかった。注意が必要である。

は7129個以下と遺伝子数が少ないので32bitのExcelに読み込むことができたが，12625個と遺伝子数の多いChiarettii et al（2004），Shingh et al., Tian et al（2003）は読み込めなかったので，64bit版のOfficeを後日購入して解決できた。試行錯誤の上，Method2を完成し，LINGOでプログラム（Program3）を作成して，表1の結果を得た。RG（新村, 2015a）に2015年中に15本の論文を掲載し，2016年には6月と7月に開催される2つのバイオ関係の国際会議のRegular ペーパーセッションに6頁と7頁の既定の頁数の論文で応募した。2つとも採択されたが参加者の多い7月開催のラスベガスの会議は試験期間中に重なっており断念した。これと並行して，国際会議以外の次の普及の方策を考えたが，Springerから人生初の英語の専門書を出版することにした。

表1.　Summary of six Microarray Data.

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

3. LINGO Program1によるGolubらのデータの分析結果の検証

Theoryでは，6種類の通常データを用いて改定IP-OLDF，改定LP-OLDF，改定IPLP-OLDF，ハードマージン最大化SVM（H-SVM）とソフトマージン最大化SVM（S-SVM）でペナルティｃを10000にしたものをSVM4とし1にしたものをSVM1（SVM0とすべきであった）とした6手法のLDFをLINGOで開発した。そしてJMP（Sall, 2004；新村, 2004）でFisherのLDF，ロジスティック回帰，QDFとRDAの10種の判別関数で比較を行った。ロジスティック回帰を除く3種の統計的判別関数は，分散共分散行列に基ずく判別関数であり，FisherのLDFの延長線上である。これに対し，Coxは，Cox回帰やロジスティック回帰などの医学診断に適した統計的判別関数を発展させたFisherの正当な後継者であろう。CoxがFisherの判別関数では扱えない他の判別分析の世界があることを，筆者に先駆けて明確に問題提起してくれれば良かったと考える。問題5を解決するためにMethod2を開発し，それをLINGOで実行するProgram3を開発した。しかし，41日間の限られた時間で，Method2を開発し，それをProgram3で実行した結果で論文を発表し，書籍まで出版してしまった。ひょっとして，「プログラムにバグがあって計算結果の一部に間違いがないか，あるいは筆者の得意でない7000個以上の判別問題で数値計算上の隠された問題にきずいていないことがないか」と考えたのは，Springerの初

稿の校正を終えた2016年8月11日である。この日から遅ればせながら，Method2とLINGOのProgram1と3の計算結果を見直すことにして昼夜連続で再計算を行い比較検証することにした。しかし暫くして，過去の結果と全く異なる結果が得られ，目の前が真っ暗になった。25日頃に，数日前に「Wordがメモリ管理エラーの表示で終了できなくなり強制終了した」ことを思い出した。そこで多分，これが原因ではないかと考え，8月末まで対策を考えた。しかし，「9月2日に再計算すると思った過去の結果と合致したので，何の通知もなくエラーが修復されたと判断した」。しかし，全く結果をあたかも正常処理したかのように処理することがあるかどうかは断定できない。現在，フォルダーに格納された間違っていると思われる結果の同定と削除も未検討のままである。今後，論文に用いる場合，過去の結果と比較検証する煩雑な十字架を背負った。

3.1　判別係数の検討

Program3の検証の前に，Program1で7129個の遺伝子を6個のMP-based LDFsで判別分析した。この結果，9秒で6個のLDFが，全てGolubらのデータをMNMあるいはNM=0で判別できることが分かった。すなわち，スイス銀行紙幣データより2群は離れていて判別が容易であることが分る。そこで，Program1からExcel上に出力した判別係数をJMPに読み込んで，0になる係数の数を調べることにした。これまでExcel上に出力された判別係数を，不正確な目視で確認していたことを後悔している。また，Excelでは真の0と微小な数値の判定が難しいのが大きな問題であった。そこで，LINGOの分析結果を，JMPという別のソフトで統計的に検証することにした。図1の最初の6列は，Program1でExcelに出力したRIPからSVM1まで6個の判別係数である。それをJMPに読み込んで，左から6列のLDFの判別係数から，右の6列の1/0に変換して度数を求めた。そして分かったことは，LINGO，Excel，JMPという3種の0判定の統一の難しさである。

	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIPC	LPc	IPLPC	HSVMC	SVM4c	SVM1c
1	0	0	0	1.234567881	-2.31699e-8	0	0	0	0	1	1	0
2	0	0	0	1.234567881	-2.31699e-8	0	0	0	0	1	1	0
3	0	0	0.000478653	1.234567881	0.000459164	0.00047866	0	0	1	1	1	1
4	0	0	-0.00000858276	1.234567881	-0.0000470035	-0.00000858819	0	0	1	1	1	1
5	0	0	0	1.234567881	0.000000589121	0	0	0	0	1	1	0
6	0	0	0	1.234567881	-2.31699e-8	0	0	0	0	1	1	0
7	0	0	0.000985976	1.234567881	0.000944805	0.00098597	0	0	1	1	1	1
8	0	0	0	1.234567881	-2.31699e-8	0	0	0	0	1	1	0
9	0	0	-0.000581072	1.234567881	-0.000576105	-0.000581076	0	0	1	1	1	1
10	0	0	-0.000361532	1.234567881	-0.000298082	-0.000361531	0	0	1	1	1	1
11	0	0	0.000105578	1.234567881	0.000116232	0.000105579	0	0	1	1	1	1
12	0	0	0	1.234567881	-2.31699e-8	0	0	0	0	1	1	0
13	0	0	0	1.234567881	-2.31699e-8	0	0	0	0	1	1	0

図1　6個の判別係数とゼロの係数と非ゼロの係数を1/0に変換

LINGOプログラムでは，最初に数千個から数万個の判別係数を初期値として0に設定してい

る。そして最適化計算で判別係数が0でないものをExcelに出力しているが，最適化計算でも初期設定の0で変わらないものと，微小な係数値になったものの区別が難しいし，過去の発表論文にこの読み間違いがあると考えられる。問題は，LINGOやExcelなどの数値計算で0判定に用いている閾値が同じでないことである。それを例えば図2のように，H-SVMの係数値をJMPで0と判定する値をHSVM cの値として0に，それ以外を1に置き換えて，新しい6個の変数を作成した。このとき6個の変数を連続尺度でなく，名義尺度にしておく必要があることが分かった。

$$\left\{\begin{array}{l} \text{If}\left\{\begin{array}{ll} HSVM==0 & \Rightarrow HSVMC=0 \\ \text{else} & \Rightarrow HSVMC=1 \end{array}\right.==0 \Rightarrow HSVMC=0 \\ \text{else} \qquad\qquad \Rightarrow HSVMC=1 \end{array}\right.$$

図2　係数が0のものと0以外の変換

図3は，この度数表である。定数項を含むので合計の7130から1を引いた7129個の判別係数すなわち遺伝子がある。改定IP-OLDFは71個，改定LP-OLDFと改定IPLP-OLDFは26個，H-SVMとSVM4は全係数，SVM1は6225個の判別係数が0でない。即ち改定IP-OLDFは，7129個の遺伝子の中から，僅か71個の遺伝子を用いれば，癌と正常をMNM=0で判別できる。すなわち，この71個は癌を特定できる癌遺伝子と考えられる。1回だけの判別分析では，改定LP-OLDFと改定IPLP-OLDFの方が7129次元を一気に26次元のMatroskaを特徴抽出し，改定IP-OLDFの71個より優れている。しかし，繰り返し判別を行うと改定IP-OLDFの方がより少ない変数を選ぶ傾向がある。SVM1は，通常のデータの分析ではSVM4より判別結果は悪いが，こと遺伝子解析の特徴抽出では904個が0になっている。この理由は今後の課題である。

図3　全データの特徴抽出

3.2　Program1

LINGOは，通常の数式表記とほぼ同じ自然表記と，集合節（SETS），DATA節，CALC節を用いて高度な最適化システムを制御するプログラムを作成できる表記法がある。高校数学で大きな役割を占める微分は，関数の極大と極小値しか得られない。LINGOの自然表記を用

いれば，関数の最大/最小値が簡単に求まるので数理計画法ソフトを使って実際の問題を解くことを大学教育に取り入れるべきであろう。

集合節は，「SETS」で始まり「ENDSETS」で終わる。1次元集合と必要であれば1次元配列を，「集合名：配列名；」で定義する。P，P1，P2が1次元集合で，PとP2は集合だけの定義である。Pは，DATA節の「P=1..7129;」でMicroarrayデータの要素数を指定しているので，7129個の要素を持つ1次元配列になる。P1は定数項を含む7130個の要素を持つ1次元集合を表し，P2はさらに6個の判別結果を表す1から6の識別子を持つ7131個の要素数を持っている。P1は1次元配列VARKとCHOICEを定義している。VARKは6個の判別関数の判別係数が格納され，CHOICEは判別に用いる変数を1/0で指定する。これをExcel上にExcelのセル範囲名CHOICEで定義し，@OLE()関数でLINGOに読み込みLINGOの配列名CHOICEとして用いる。これによって複数の判別関数のモデルが連続実行できる。1次元集合Nはデータ件数のObjectを表す集合で，各ケースに対応したe_iを表す配列Eと，判別スコアを表す配列SCOREを定義している。CONSTANTは，改定IPLP-OLDFが最初に改定LP-OLDFを解いて，正しく判別されたケースを固定し，判別されなかったケースに限定して改定IP-OLDFを解くための1/0の情報インデックスを格納している。1次元集合G2は，6個のLDFの誤分類数をICに，判別境界上の件数をZEROに，正しく判別された件数をNPに出力する。これらの和は，データ件数nになる。最後の1次元集合V2は，6種の判別を行うことを示す。最初の2次元集合Dは，72個のデータを持つ1次元集合Nと，7130個の要素数を持つP1で作られる（72＊7130）の2次元集合を定義し，データが格納されているExcelの配列ISを表す。2次元集合VGはV2とP1すなわち（6*7130）の2次元配列VARK100を定義し，6個のLDFの7130個の判別係数を最適化計算で@OLE()関数でLINGOからExcelに出力する。

```
MODEL:
SETS:
  P; P2; P1:VARK,CHOICE;
  N:E,CONSTANT,SCORE;
  G2:IC,NP,ZERO; V2:;
  D(N,P1):IS;
  VG(V2,P1):VARK100;
ENDSETS
```

DATA節は，集合PからV2までの要素数を，「V2=1..6;」のように指定している。BIGMでBigMを定数10000のように指定する。「CHOICE,IS=@OLE();」は，Excel上のセル範囲のCHOICEに判別モデルの変数の指定を1/0で格納したものと，ISに分析データを与えてそれをLINGOに読み込みLINGOの同名の配列にデータを格納する。

```
DATA:
    P=1..7129; P1=1..7130; P2=1..7131;N=1..72; G2=1..6;V2=1..6;
BIGM=10000;
    CHOICE,IS=@OLE();
ENDDATA
```

SUBMODEL節で，6個のLDFを定義する。最初は，改定IP-OLDFを「RIP」というサブモデル名で定義している。「MIN=@SUM(N(i):E(i));」で目的関数「@SUM(N(i):E(i))」を最小化している。「@SUM」は重要な繰り返し関数の一つで「$\Sigma_{i=1,\cdots,72}$ E(i)」すなわち誤判別を表す1の個数の合計を最小化している。「@FOR(N(i): @SUM(P1(J1): IS(i,J1) *VARK(J1)*CHOICE(J1)) > 1-BIGM*E(i));」の「@FOR」は重要な2番目の繰り返し関数の一つでj1が1から72個のケース数に対して，「$\Sigma_{j1=1,\cdots,72}$ IS(i,J1)*VARK(J1)*CHOICE(J1)) > 1-BIGM*E(i));」の72個の制約式をこの1文で表す。「@FOR(P1(J1):@FREE(VARK(J1)));」は7129個の変数を「:@FREE(VARK(J1)));」で自由変数に指定している。数理計画法は，伝統的に非負の実数を基本にしているので，制約のない正負の実数を自由変数と言っている。「@FOR(N(i):@BIN(E(i)));」で72個のe_iを0/1の2値整数変数としている。この変数は0であれば対応するケースが正しく判別され，1であれば誤判別される。SVM4とSVM1はCALC節でPenalty cの値を10000と1とすることで判別できる。モデルを見れば構造が似ているが，決定変数が0/1の整数であれば整数計画法（IP）で改定IP-OLDF，非負の正の実数であれば線形計画法（LP）で改定LP-OLDF，目的関数が2次式であればSVMになる。

```
SUBMODEL RIP:
  MIN=@SUM(N(i):E(i));
  @FOR(N(i):@SUM(P1(J1):IS(i,J1)*VARK(J1)*CHOICE(J1)) > 1-BIGM*E(i));
  @FOR(P1(J1):@FREE(VARK(J1)));
  @FOR(N(i):@BIN(E(i)));
ENDSUBMODEL
SUBMODEL IPLP:
  MIN=ER; ER=@SUM(N(i):E(i));
  @FOR(N(i):@SUM(P1(J1):IS(i,J1)* VARK(J1)*CHOICE(J1) )>1-BIGM*E(i));
  @FOR(P1(J1):@FREE(VARK(J1)));
       @FOR(N(I)| CONSTANT(i)#NE#0:@BIN(E(I)));
       @FOR(N(I)| CONSTANT(i)#EQ#0:E(I)=0);
ENDSUBMODEL
SUBMODEL LP:
```

```
  MIN=ER;  ER=@SUM(N(i):E(i));
  @FOR(N(i):@SUM(P1(J1):IS(i,J1)*VARK(J1)*CHOICE(J1)) > 1-BIGM*E(i));
  @FOR(P1(J1):@FREE(VARK(J1)));
ENDSUBMODEL
SUBMODEL HSVM:
  MIN=ER; ER=@SUM(P(J):VARK(J)^2)/2;
  @FOR(N(i):@SUM(P1(J1):IS(i,J1)*VARK(J1)*CHOICE(J1)) > 1);
  @FOR(P1(J1):@FREE(VARK(J1)));
ENDSUBMODEL
SUBMODEL SVM:
  MIN=ER; ER=@SUM(P(J):VARK(J)^2)/2+Penalty*@SUM(N(i):E(i));
  @FOR(N(i):@SUM(P1(J1):IS(i,J1)*VARK(J1)*CHOICE(J1)) > 1-E(i));
  @FOR(P1(J1):@FREE(VARK(J1)));
ENDSUBMODEL
```

CALC節は，「@SOLVE(RIP);」でSUBMODELで定義した「RIP」の最適化を行う。「@FOR(n(I): SCORE(I)= @SUM(P1(J1):IS(i,J1)*VARK(J1)*CHOICE(J1)));」でSCORE(I)に72個のケースの判別スコアを格納し，「@FOR(n(I): @IFC(SCORE(I)#EQ#0: Z=Z+1));」のように，0判定を行い，判別超平面上のケース数を配列Zに格納している。ここでは単純に6個のLDFを順次計算しているだけであるが，「@WHILE()」関数で繰り返しなどが簡単にできる。最後にDATA節の「@OLE( )=VARK100,IC,ZERO,NP;」でもって6個のLDFの判別係数と誤分類数，判別超平面上のケース数，正しく判別されたケース数をExcelに出力する。

```
CALC:
@SET('DEFAULT'); @SET('TERSEO',2);
G=1;NM=0;NMP=0; Z=0;
@FOR( P1( J1): VARK( J1) = 0; @RELEASE( VARK( J1)));
@SOLVE(RIP);
  @FOR(P1(J1):VARK100(G,J1)=VARK(J1)*CHOICE(J1) );
  @FOR(n(I):SCORE(I)=@SUM(P1(J1):IS(i,J1)*VARK(J1)*CHOICE(J1)));
  @FOR(n(I):@IFC(SCORE(I)#EQ#0:  Z=Z+1));
  @FOR(n(I):@IFC(SCORE(I)#LT#0:  NM=NM+1));
    @FOR(n(I):@IFC(SCORE(I)#GT#0:  NMp=NMp+1));
  IC(G)=NM; ZERO(G)=Z;NP(G)=NMP;
G=2; NM=0;NMP=0; Z=0;
```

```
@SOLVE(LP);
    @FOR(N(i): @IFC( E(I)#EQ#0: CONSTANT(i)=0; @ELSE CONSTANT(i)=1;));
NM=0;NMP=0; Z=0;
@solve(IPLP);
    @FOR(P1(J1):VARK100(G,J1)=VARK(J1)*CHOICE(J1) );
    @FOR(n(I):SCORE(I)=@SUM(P1(J1):IS(i,J1)*VARK(J1)*CHOICE(J1)));
  @FOR(n(I):@IFC(SCORE(I)#EQ#0:  Z=Z+1));
  @FOR(n(I):@IFC(SCORE(I)#LT#0:  NM=NM+1));
      @FOR(n(I):@IFC(SCORE(I)#GT#0:  NMp=NMp+1));
IC(G)=NM; ZERO(G)=Z;NP(G)=NMP;
G=3;NM=0;NMP=0; Z=0;
@SOLVE(LP);
  @FOR(P1(J1):VARK100(G,J1)=VARK(J1)*CHOICE(J1) );
  @FOR(n(I):SCORE(I)=@SUM(P1(J1):IS(i,J1)*VARK(J1)*CHOICE(J1)));
  @FOR(n(I):@IFC(SCORE(I)#EQ#0:  Z=Z+1));
  @FOR(n(I):@IFC(SCORE(I)#LT#0:  NM=NM+1));
    @FOR(n(I):@IFC(SCORE(I)#GT#0:  NMp=NMp+1));
IC(G)=NM; ZERO(G)=Z;NP(G)=NMP;
G=4;NM=0;NMP=0; Z=0;
@SOLVE(HSVM);
  @FOR(P1(J1):VARK100(G,J1)=VARK(J1)*CHOICE(J1) );
  @FOR(n(I):SCORE(I)=@SUM(P1(J1):IS(i,J1)*VARK(J1)*CHOICE(J1)));
  @FOR(n(I):@IFC(SCORE(I)#EQ#0:  Z=Z+1));
  @FOR(n(I):@IFC(SCORE(I)#LT#0:  NM=NM+1));
      @FOR(n(I):@IFC(SCORE(I)#GT#0:  NMp=NMp+1));
IC(G)=NM; ZERO(G)=Z;NP(G)=NMP;
G=5;PENALTY=10000;NM=0;NMP=0; Z=0;
@SOLVE(SVM);
  @FOR(P1(J1):VARK100(G,J1)=VARK(J1)*CHOICE(J1) );
  @FOR(n(I):SCORE(I)=@SUM(P1(J1):IS(i,J1)*VARK(J1)*CHOICE(J1)));
  @FOR(n(I):@IFC(SCORE(I)#EQ#0:  Z=Z+1));
  @FOR(n(I):@IFC(SCORE(I)#LT#0:  NM=NM+1));
      @FOR(n(I):@IFC(SCORE(I)#GT#0:  NMp=NMp+1));
```

```
IC(G)=NM; ZERO(G)=Z;NP(G)=NMP;
G=6;PENALTY=1;NM=0;NMP=0; Z=0;
@SOLVE(SVM);
  @FOR(P1(J1):VARK100(G,J1)=VARK(J1)*CHOICE(J1) );
  @FOR(n(I):SCORE(I)=@SUM(P1(J1):IS(i,J1)*VARK(J1)*CHOICE(J1)));
  @FOR(n(I):@IFC(SCORE(I)#EQ#0:  Z=Z+1));
  @FOR(n(I):@IFC(SCORE(I)#LT#0:  NM=NM+1));
     @FOR(n(I):@IFC(SCORE(I)#GT#0:  NMp=NMp+1));
IC(G)=NM; ZERO(G)=Z;NP(G)=NMP;
ENDCALC
DATA:
@OLE( )=VARK100,IC,ZERO,NP;
ENDDATA
END
```

3.3　日本車データによるProgram3とProgram1の検証

一気に高次元データの分析をProgram3で行い，論文と書籍を出版した。そして今回，Program1で判別係数が0になる個数を3.2で初めてJMPで確認した。しかし研究方法としては，「通常のデータ」で事前に確認した上で，手に余る高次元データ解析を行うのが研究に重要であることに8月初旬に気づいた。そこで，スイス銀行紙幣データと日本車44車種のデータがLSDであるので，6手法の判別係数の検証を遅ればせながら行うことにした。「試験のデータ」は，100個の変数と4個の変数の判別なので，それを検証することも考えたが，後日に行うことにした。日本車44車種のデータは，以下に示す通り，2個のBGSを見つけてくれた。一方，スイス銀行紙幣データはいわゆる1個のSmall Matroska（SM）[6]を見つけて，BGSと分かっている（X4，X6）を見つけてくれなかった。

図4は，Program1で求まる6個のLDFで6個の判別係数と定数項を表している。6列目までが6個の手法に対応し，最初の6行はX1からX6の判別係数，7行は定数項を表す。JMPで1/0に変換した結果を7列から12列に示す。変数の個数が少ないので度数表でなく，JMPの出力画面を示す。RIPは，X2，X3，X6と定数項が0である。IPLPとLPとSVM1のX2の係数の絶対値が“9.99e-8”以下でもJMPは0と判定していないが，LINGOでは0になる設定をしている。

---

[6] 筆者は遺伝子解析で，癌遺伝子を含むMNM=0の部分空間を全てSMと呼んでいる。その中で最小次元のSMをBasic Gene Space（BGS）と呼ぶことにした。ただし本稿では，MNMが1以上のものもSMを拡大解釈している。

すなわち，LINGO，Excel，JMPという3つのソフトの0判定をどう管理するか今後の課題である。

	RIP	IPLP	LP	HSVM	SVM4	SVM1	RIPc	IPLPc	LPc	HSVMc	SVM4c	SVM1c
1	0.064039589	5.999873227	5.999873227	0.626799308	0.650937451	0.68049469	1	1	1	1	1	1
2	0	-2.79314e-8	-2.79314e-8	-0.000000163815	-0.00000014654	-5.84504e-8	0	1	1	1	1	1
3	0	0	0	1.776756792	1.779111616	1.622019026	0	0	0	1	1	1
4	0.004921196	-0.018700542	-0.018700542	-0.009045659	-0.007373681	-0.002306904	1	1	1	1	1	1
5	0.018511664	-0.132291642	-0.132291642	-0.06074281	-0.052106097	-0.011480242	1	1	1	1	1	1
6	0	-0.0000248452	-0.0000248452	0.00000380573	0.00000346334	-0.000000768695	0	1	1	1	1	1
7	0	0	0	-6.083293104	-6.501028061	-7.366360572	0	0	0	1	1	1

図4　Program1で6個のLDFで6個の判別係数と定数項を1/0に変換

表2は，日本車データの小ループの結果をProgram3Sで，表3は小ループを表示しない大ループだけのProgram3Lで分析した改定IP-OLDFの判別結果である。SM列は，大ループ即ちSM番号を表す。IT列は小ループの繰り返し回数を3に設定した。SM=1でIT=1では，6変数のフルモデルを分析している。SUM列は定数項を除く変数の個数で，NM列はMNMが0であることを示す。この繰り返し判別の結果，X2からX6の係数が0になり，IT=2でX1だけで判別する。しかし，IT=3でX1をさらに0にしないで同じ1変数モデルになり，これがSMになる。しかし1個なのでBGSでもある。すなわちこのBGSに含まれる遺伝子を含む部分空間は全てSMになる。次にSM=2で6個の変数からX1を省いて判別すると，MNM=0である。判別結果は, IT=2でX2とX4からX6の4個の係数が0になる。IT3でこの4個を省いてX3で判別し，当然X3だけがSM2に選ばれる。変数が1個であるので，2個目のBGSでもある。SM=3では，X1とX3を省いた4変数で判別すると，MNM=4でもうSMではなくなる。この分析結果は，X2とX5の係数が0でなく，X4とX6が0になり，IT=3でも同じ結果になり，次の大ループ4（SM4）に進む。X2とX5を省いてX4とX6で判別すると，MNM=9でありここで6個すべての変数を，（X1）と（x3）の2個のBGSと，（X2, X5）と（X4, X6）の2個の変数群に分けた。これらのMNMは，0，0，4，9であり，Microarrayデータでは，癌遺伝子の優先順位を表しているのではないかと予見している。SM=5は，変数はないがプログラムの簡略化のため「IC.GE.15」に設定して誤分類数が15を超えると計算を停止しているのでNM=15を出力している。間違って「IC.GE.8」と設定すれば，SM=4すなわち（X4, X6）の判別は行わないでNM=4が出力される。この結果は，1行目のMatroskaに示されている。X1がSM1，X3がSM2で，これまでMNM=0になるものだけをMatroskaとしてきたが，1以上も拡大解釈すると，X2とX5がSM3，X4とX6がSM4になり，6変数は4個の排他的なSMの和集合になる。これが，筆者が発見した高次元のMicroarrayデータの驚く構造である。これまで10年から20年近く，多くの統計家と医学研究者が，統計的にMicroarrayデータの特徴抽出すなわち癌遺伝子の特定を行ってきた。彼らは，「高次元の遺伝子空間に，幾つかの癌遺伝子を統計手法で探す手法の開

発を試み，どうもうまくいかなかったようだ」。次にIT2=4，IC.GE.15に設定し，Program3Lで判別した。上の小ループの最後の3列目の結果だけが表示される。結局，日本車データは，X1とX3がBGSで，SM3とSM4はMNMが4と9である。この事実から，MNMが癌の優先度を表し，改定IP-OLDFは，IPで「First in First out」で最適解を出力するが，2群の離れ具合で選んでいるのではないかと考えている。

表2　日本車データをProgram3SとProgram3Lで分析した改定IP-OLDFの判別結果

				Matroska	1	3	2	4	3	4	0
SM	IT	T	NM	SUM	X1	X2	X3	X4	X5	X6	c
1	1	6	0	6	1	1	1	1	1	1	1
1	2	1	0	1	1	0	0	0	0	0	1
1	3	1	0	1	1	0	0	0	0	0	1
2	1	5	0	5	0	1	1	1	1	1	1
2	2	1	0	1	0	0	1	0	0	0	1
2	3	1	0	1	0	0	1	0	0	0	1
3	1	4	4	4	0	1	0	1	1	1	1
3	2	2	4	2	0	1	0	0	1	0	1
3	3	2	4	2	0	1	0	0	1	0	1
4	1	2	9	2	0	0	0	1	0	1	1
4	2	2	9	2	0	0	0	1	0	1	1
4	3	2	9	2	0	0	0	1	0	1	1
5	1	0	15	0	0	0	0	0	0	0	1
5	2	0	15	0	0	0	0	0	0	0	1
5	3	0	15	0	0	0	0	0	0	0	1
SM	IT	T	NM	SUM	X1	X2	X3	X4	X5	X6	c
1	4	1	0	1	1	0	0	0	0	0	1
2	4	1	0	1	0	0	1	0	0	0	1
3	4	2	4	2	0	1	0	0	1	0	1
4	4	2	9	2	0	0	0	1	0	1	1
5	4	0	15	0	0	0	0	0	0	0	1

表3は，日本車データをProgram3Sで分析した改定LP-OLDFの判別結果である。SM=2までは表2と同じである。SM=3では，X1とX3を省いた4変数で判別すると，X2とX4とX6の係数が0でなく，X5が0になり，NM=6でもうSMではなくなる。SM=4ではX6で判別すると，「IC.GE.15」の設定で停止する。6個すべての変数を（X1）と（x3）の2個のBGSと，（X2, X4, X6）と（X5）の2個の変数群に分けた。これらのNMは，0，0，6，15以上である。

表3　日本車データをProgram3Sで分析した改定LP-OLDFの判別結果

SM	IT	T	NM	SUM	X1	X2	X3	X4	X5	X6	c
1	1	6	0	6	1	1	1	1	1	1	1
1	2	1	0	1	1	0	0	0	0	0	1
1	3	1	0	1	1	0	0	0	0	0	1
2	1	5	0	5	0	1	1	1	1	1	1
2	2	4	0	4	0	1	1	1	0	1	1
2	3	1	0	1	0	0	1	0	0	0	1
3	1	4	6	4	0	1	0	1	1	1	1
3	2	3	6	3	0	1	0	1	0	1	1
3	3	3	6	3	0	1	0	1	0	1	1
4	1	1	15	1	0	0	0	0	1	0	1
4	2	0	15	0	0	0	0	0	0	0	1
4	3	0	15	0	0	0	0	0	0	0	1

表4は，日本車データをProgram3Lで分析したH-SVMとSVM4/SVM1の判別結果で，変数選択が行えないのでSM1の探索で停止した。

表4　日本車データのProgram3Sで分析したH-SVMとS-SVMの判別結果

SM	IT	T	NM	SUM	X1	X2	X3	X4	X5	X6	c
1	1	6	15	6	1	1	1	1	1	1	1
1	2	6	15	6	1	1	1	1	1	1	1
1	3	6	15	6	1	1	1	1	1	1	1

表5は，日本車データの6手法の表2の5回の判別結果を改定IP-OLDFでMethod2をシミュレーションしている。IC列に誤分類数，ZERO列に判別超平面上のケース数，NPに正しく判別されたケース数を示す。合計は44である。手法列は6手法であり，CHOICE行で1になる変数を各ステップの最初でモデルに選んで判別する。最初6変数で判別すると，3個のSVMの係数は全て0でない。これに対して，3個のOLDFはX3が0，X2とX6が微小な値になっている。筆者は，この扱いに関してLINGOに任せていて見識はないが絶対値で10^{-8}以下を0と判定している。しかし，他にも収束判定条件の設定が色々あるので，その最適な組み合わせは分かっていない。OLDFの定数項は0である。そこでX2とX3と定数項のCHOICE列のように値を0にして省くと，2番目の分析欄のように，変数選択ができない。3番目のように定数項をモデルに含めると，3個のOLDFはX1と定数項の係数が同じで他の5個の変数を省くことができる。これは，これまでの結果とよく合う。しかし，定数項が0になるのは，定数項を含むモデルが冗長と言ってきたが，これほど大きな結果の違いになるとは考えていなかった。多分，この点に関した研究はないと考える。3個のSVMも4番目の分析欄のように0にな

るものもあり変数選択できるが，6変数でできないので，これらの3手法は自然に変数選択できるとは考えていない。4回目の分析で，X1をモデルから省いて5変数で判別すると，3個のOLDFのX3の係数は2で定数項は-9になる。これはすでに書籍にも書いているが判別超平面がX3=9/2=4.5を示す。この点が変数選択に有利になっているようだ。すなわち，小型車は座席が4.5席以下で普通車が4.5席以上という常識的な事実を指摘している。最後の分析で，X1とX3を省いた4変数で判別すると4変数とも変数選択できず終了する。ただし，最後のLPと3種のSVMは誤分類数が3でなく4になっている。

表5　日本車データの6手法の判別結果

IC	ZERO	NP	手法	X1	X2	X3	X4	X5	X6	c
0	0	44	RIP	5.9999	-3E-08	0	-0.02	-0.13	-2E-05	0
0	0	44	IPLP	5.9999	-3E-08	0	-0.02	-0.13	-2E-05	0
0	0	44	LP	5.9999	-3E-08	0	-0.02	-0.13	-2E-05	0
0	0	44	HSVM	0.6268	-2E-07	1.777	-0.01	-0.06	4E-06	-6.08
0	0	44	SVM4	0.6509	-1E-07	1.779	-0.01	-0.05	3E-06	-6.5
0	0	44	SVM1	0.6805	-6E-08	1.622	-0	-0.01	-8E-07	-7.37
	CHOICE			1	1	1	1	1	1	1
0	0	44	RIP	5.9271			-0.02	-0.12	-8E-07	
0	0	44	IPLP	6.0547			-0.02	-0.13	-3E-05	
0	0	44	LP	6.0547			-0.02	-0.13	-3E-05	
0	0	44	HSVM	5.9187			-0.02	-0.12	4E-06	
0	0	44	SVM4	5.9188			-0.02	-0.12	4E-06	
0	0	44	SVM1	2.9584			-0.01	-0.06	-5E-07	
	CHOICE			1	0	0	1	1	1	0
0	0	44	RIP	5.9172			0	0	0	-4.89
0	0	44	IPLP	5.9172			0	0	0	-4.89
0	0	44	LP	5.9172			0	0	0	-4.89
0	0	44	HSVM	5.9172			1E-08	7E-08	0	-4.89
0	0	44	SVM4	5.9175			-0	-0.02	8E-07	-3.97
0	0	44	SVM1	2.9806			4E-06	1E-05	0	-2.96
	CHOICE			1	0	0	1	1	1	1
0	0	44	RIP		0	2	0	0	0	-9
0	0	44	IPLP		0	2	0	0	0	-9
0	0	44	LP		0	2	0	0	0	-9
0	0	44	HSVM		0	2	0	0	0	-9
0	0	44	SVM4		9E-09	2.005	-0	-0	-4E-08	-8.99
0	0	44	SVM1		0	2	0	0	0	-9
	CHOICE			0	1	1	1	1	1	1

3	0	41	RIP		0.0033		-46.4	-199	-0.036	5343
3	0	41	IPLP		0.0033		-46.4	-199	-0.036	5343
4	0	40	LP		6E-06		-0.15	-0.8	-2E-05	25.93
4	0	40	HSVM		6E-06		-0.15	-0.8	-2E-05	25.93
4	0	40	SVM4		6E-06		-0.15	-0.8	-2E-05	25.93
4	0	40	SVM1		6E-06		-0.15	-0.79	-3E-05	25.05
	CHOICE			0	1	0	1	1	1	1

3.4 種々の試行

折角テストデータでプログラムを検証することにしたので，Microarrayデータでは確認できない「DatasetsがSMの和集合になっている構造である」ことを検証することにした。これを日本車データとスイス銀行データを2回コピーし，6変数から18変数のデータを作成した。これで元のデータの3倍のSMが得られるかどうかである（試行1)。また，日本車データで，なぜ最初にX1を含むSMが選ばれ, X3（座席数）を含むSMが選ばれないかである。このため，X3の小型車の座席数の4を1から3の間で変更して分析する（試行2)。

3.4.1 試行1

実は研究の大筋において問題はないが，大きな問題点を見つけた。LINGOは，LP，QP，IP，NLPそして確率計画のソルバーを，ユーザーが意識することなくモデルで指定することによって適切なソルバーが選ばれる。問題点は，各ソルバーで0判定や，収束の打ち切りに設定されている閾値が異なっていることである。筆者は，プログラムをそれと関係なく絶対値が‘9.9…e-8’以下になればという想定でプログラムを作成したが，各ソルバーの収束判定が異なっている。これが原因であると思われるが，少し条件を変えるとことなる結果が得られることがある。元の6変数の分析結果は同じであるので省略して，データを3回コピーして18変数のデータを生成した分析結果を示す。元のデータでX1とX3を含む2個のBGSが得られたが，それを3倍に拡大したデータで6個のBGSが得られることを確認した。

表6は，Program3Sで小ループの繰り返し回数を10回に指定して解いた結果である。2回目から9回目の結果を非表示にして示してある。C1からC6は元の変数，c11からc16は1回目のコピー，c21からc26は2回目のコピーである。最初18変数の判別を改定IP-OLDFで行うと，10回の繰り返し判別の後でC2，C6，c21，c24，c25がSM1に選ばれた。同じ変数の組でなく元のデータからC2とC6，2回目のコピーからc21，c24，c25が選ばれた。SM=2（SM2の探索）のIT=1では，SM1に含まれた5個の変数の値を0にして，選ばれなかった13個の変数を1に設定して10回判別した結果，SM=2，IT=10の結果が得られた。1回目のコピーからc11，c14，c15，c16の4個，2回目からc22の1個が選ばれてX3に対応する3個は省かれた。SM3の探索では，SM1とSM2に含まれる10個を省いた8個の変数を判別し，10回目のループで元の変数

から3個のC1，C4，C5，1回目のコピーデータからc12，2回目のコピーデータからc26の5個を選んだ。いずれにしてもSM1，SM2とSM3はX3に対応する変数を除いた5個の変数でSMになる。SM4では，これらの15個を省いた3個で判別しc23が選ばれた。SM5ではこれまでSMに選ばれた16個を省いた2個で判別しC3が選ばれた。SM6ではこれまでSMに選ばれた17個を省いた1個で判別しc13が選ばれた。すなわち18個の日本車データは，次の6個のSMの排反な和集合になった。

18個の日本車データ＝S1∪S2∪S3∪S4∪S5∪S6＝（C2，C6，c21，c24，c25）∪（c11，c14，c15，c16，c22）∪（C1，C4，C5，c12，c26）∪（c23）∪（C3）∪（c13）

表6　Program3Sで小ループの繰り返し回数を10回に指定して解いた結果

SM	IT	T	NM	SUM	C1	C2	C3	C4	C5	C6	c11	c12	c13	c14	c15	c16	c21	c22	c23	c24	c25	c26	c
1	1	18	0	18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	10	5	0	5	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	1	1	0	1
2	1	13	0	13	1	0	1	1	1	0	1	1	1	1	1	1	0	1	1	0	0	1	1
2	10	5	0	5	0	0	0	0	0	0	1	0	0	1	1	1	0	1	0	0	0	0	1
3	1	8	0	8	1	0	1	1	1	0	0	1	1	0	0	0	0	0	1	0	0	1	1
3	10	5	0	5	1	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	1
4	1	3	0	3	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1
4	10	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
5	1	2	0	2	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
5	10	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
6	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
6	10	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
7	1	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	10	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

表7は18変数のMatroskaである。Matroskaが1から7まで，表6に対応していることが分かる。

表7　18変数のMatroska

SM	C1	C2	C3	C4	C5	C6	c11	c12	c13	c14	c15	c16	c21	c22	c23	c24	c25	c26	c
1	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	1	1	0	0
2	0		0	0	0		2	0	0	2	2	2		2	0			0	0
3	3		0	3	3			3	0						0			3	0
4			0						0						4				0
5			5						0										0
6									6										0
7																			0

表8は，18変数の6個のSMに含まれる判別係数である。18変数の表示が無理なので，6個のSM順に判別係数があるものだけを表示した。各SMは元データ，コピー 1，コピー 2の順に出力された3個の5変数モデルと3個の1変数モデルだけを表示した。

表8 18変数の6個のSMに含まれる判別係数

IT	X2	X6	X21	X24	X25	c
1	-3E-08	-2E-05	5.9999	-0.019	-0.132	0
10	-3E-08	-2E-05	5.9999	-0.019	-0.132	0
	X11	X14	X15	X16	X22	c
1	5.9999	-0.019	-0.132	-2E-05	-3E-08	0
10	5.9999	-0.019	-0.132	-2E-05	-3E-08	0
	X1	X4	X5	X12	X26	c
1	5.9999	-0.019	-0.132	-3E-08	-2E-05	0
10	5.9999	-0.019	-0.132	-3E-08	-2E-05	0
	1	2	X23	4	5	c
1			2			-9
10			2			-9
	1	2	X3	4	5	c
1			2			-9
10			2			-9
	1	2	X13	4	5	c
1			2			-9
10			2			-9

3.4.2 試行2

表9は，小型車の座席数の4を1に変更した18変数の判別結果の判別係数である。X1に代わってX3の3個がSM1からSM3に選ばれX1とX3の逆転が起きた。X3の2群の範囲は1と［5, 8］で，最短の2点の中間は3（=（1+5）/2）であるが，得られた判別関数から判別境界を求めるとX3=3になっている。これらの3変数を省いてSM4からSM6を求めると表示の桁数が少ないのではっきりしないが，残り5変数の同じ係数を持つ判別係数が得られた。小型車の4席を1席に普通車から離すと，X1に代わってX3の3個が最初に選ばれた。

表9　小型車の座席数の4を1に変更した18変数の判別結果の判別係数

SM	IT	X1	X2	X3	X4	X5	X6	X11	X12	X13	X14	X15	X16	X21	X22	X23	X24	X25	X26	c
1	1									0.5										-1.5
1	10									0.5										-1.5
2	1															0.5				-1.5
2	10															0.5				-1.5
3	1			0.5																-1.5
3	10			0.5																-1.5
4	1						-0	6				-0			-0		-0			0
4	10						-0	6				-0			-0		-0			0
5	1								-0		-0		-0	6				-0		0
5	10								-0		-0		-0	6				-0		0
6	1	6	-0		-0	-0													-0	0
6	10	6	-0		-0	-0													-0	0
7	1																			1
7	10																			1

表10は，小型車の座席数の4を2に変更した18変数の判別結果の判別係数である。X1に代わってX3の3個がSM1からSM3に選ばれた。X3の2群の範囲は2と［5, 8］で，最短の2点の中間は3.5（=（2+5）/2）であるが，得られた判別関数から判別境界を求めるとX3=2.33…/0.66…=3.5になっている。これらの3変数を省いてSM4からSM6を求めると表示の桁数が少ないのではっきりしないが，残りの5変数が同じ係数を持つ判別係数が得られた。小型車の4席を2席に普通車から離すと，X1に代わってX3の3個が最初に選ばれた。

表10　小型車の座席数の4を2に変更した18変数の判別結果の判別係数

SM	IT	X1	X2	X3	X4	X5	X6	X11	X12	X13	X14	X15	X16	X21	X22	X23	X24	X25	X26	c
1	1									0.66										-2.33
1	10									0.66										-2.33
2	1															0.66				-2.33
2	10															0.66				-2.33
3	1			0.66																-2.33
3	10			0.66																-2.33
4	1						-0	6				-0			-0		-0			
4	10						-0	6				-0			-0		-0			
5	1								-0		-0		-0	6				-0		
5	10								-0		-0		-0	6				-0		
6	1	6	-0		-0	-0													-0	
6	10	6	-0		-0	-0													-0	
7	1																			1
7	10																			1

表11は，小型車の座席数の4を3に変更した18変数の判別結果の判別係数である。X1に代わってX3の3個がSM1からSM3に選ばれた。X3の2群の範囲は3と［5, 8］で，最短の2点の中間は4（=（3+5）/2）であるが，得られた判別関数から判別境界を求めるとX3=4になっている。これらの3変数を省いてSM4からSM6を求めると表示の桁数が少ないのではっきりしないが，残り5変数は同じ係数を持つ判別係数が得られた。小型車の4席を3席に普通車から離すと，X1に代わってX3の3個が最初に選ばれた。

表11 小型車の座席数の4を3に変更した18変数の判別結果の判別係数

SM	X1	X2	X3	X4	X5	X6	X11	X12	X13	X14	X15	X16	X21	X22	X23	X24	X25	X26	c
1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	-4
1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	-4
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	-4
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	-4
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-4
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-4
4	0	0	0	0	0	-0	6	0	0	0	-0	0	0	-0	0	-0	0	0	0
4	0	0	0	0	0	-0	6	0	0	0	-0	0	0	-0	0	-0	0	0	0
5	0	0	0	0	0	0	0	-0	0	-0	0	-0	6	0	0	0	-0	0	0
5	0	0	0	0	0	0	0	-0	0	-0	0	-0	6	0	0	0	-0	0	0
6	6	-0	0	-0	-0	0	0	0	0	0	0	0	0	0	0	0	0	-0	0
6	6	-0	0	-0	-0	0	0	0	0	0	0	0	0	0	0	0	0	-0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

3.5 スイス銀行データによるProgram3とProgram1の検証

3.5.1 試行1

次にスイス銀行データの変数選択を行った。Program1はうまくいったのに，8月31日まではPorogram3の結果に理解できない点がある。当初はPorogram3にまだ検証の余地があると考えた。しかし，8月中旬から再検証を昼夜連続で実行していて，Wordの終了が行えなくて「Memory管理エラー」が現れたことなどから，Windows10のエラーかも分らないと考えるようになった。当分は様子を見守り計算を中断し，過去の分析例と再検証のために行った結果を比較して，これを確認することを一度決定した。しかし，9月1日になって，もう一度再検証を行ったが問題が見られない。多分オンラインで修復されたのであろうが，結果に最新の注意を払うことにした。

図5は，Program1で6個のLDFで判別した6個の判別係数と定数項を表している。6列目ま

でが6個の手法に対応し，最初の6行はX1からX6の判別係数，7行は定数項を表す。JMPで1/0に変換した結果を7列から12列に示す。変数の個数が少ないので度数表でなく，JMPの出力画面を示すことで個々の状況が把握できる。RIP，IPLP，LPは，X2，X3と定数項が真の0であることを確認した。H-SVMは「0e+0」なので0になっても良いのに，全て1である。JMPの結果にも影響が出ていたか否かは分らない。

	RIP	IPLP	LP	HSVM	SVM4	SVM1	RIPc	IPLPc	LPc	HSVMc	SVM4c	SVM1c
1	-1.09e+0	-1.09e+0	-1.09e+0	0e+0	-1.14e+0	-3.14e-1	1	1	1	1	1	1
2	0e+0	0e+0	0e+0	0e+0	-5.67e-1	6.98e-2	0	0	0	1	1	1
3	0e+0	0e+0	0e+0	0e+0	1.24e-1	-2.15e-1	0	0	0	1	1	1
4	-2.61e+0	-2.61e+0	-2.61e+0	0e+0	-2.3e+0	-1.05e+0	1	1	1	1	1	1
5	-2.83e+0	-2.83e+0	-2.83e+0	0e+0	-2.8e+0	-7.96e-1	1	1	1	1	1	1
6	2.06e+0	2.06e+0	2.06e+0	0e+0	1.8e+0	1.37e+0	1	1	1	1	1	1
7	0e+0	0e+0	0e+0	0e+0	1.02e+2	-8.84e+1	0	0	0	1	1	1

図5　Program1で6個のLDFで6個の判別係数と定数項を1/0に変換

表12は，Program1による6手法の判別結果である。IC列に誤分類数，ZERO列に判別超平面上のケース数，NPに正しく判別されたケース数を示す。合計は200である。手法列は6手法である。次の7列は判別係数で，最後の7列は1が判別モデルにその変数を含み，0は含まないことを表す。最初はフルモデルの6変数で分析すると，OLDFの3手法の判別係数が同じで，しかもX2，X3と定数項が0である。Springerの書籍では，従来の研究を見直しスイス銀行データと日本車データでMethod2を説明したが，6手法の判別係数を同時に比較していなかった。この結果を見れば，通常のデータでは3個のOLDFは自然にX2とX3と定数項を選ぶ必要がないことが分かる。そしてSVMの3手法は，判別係数から変数選択ができないことが分かる。LASSOは高次元データの変数選択を行う理論を研究しているが，まず通常のデータで自然に変数選択を確認することを勧めたい。普通のデータでできないのに，高次元データでできると考える理論的な根拠が分からない。例えば，20年ほど前に遺伝子の高次元データの解析で行われたことは，FisherのLDFやQDFによるアプローチである。この問題は，例えば100件のデータから1万個の変数の分散共分散行列をどう推測するかである。種々の研究が一時行われたようだが，昨年ようやくJMPが高次元データの判別を行うFisherのLDFを開発した。ここで研究者と統計ソフトの開発企業の新手法に対する対応を切り分ける必要がある。私自身，ヘルシンキで開催された1999年のISIの国際会議の後，Tarutoで開かれた国際会議でも発表した。1回の国際会議の出張で2件の発表し，研究費の抑制を考えた最初の学会である。この会議で，多分私の最初のIP-OLDFに関する論文を，その後に発刊される学術誌（Shinmura, 2000a）に推薦してくれたと考えている米国在住のインドの研究者が，このテーマを発表していた。私は遠慮がちに「普通のデータでも分散共分散に基づく判別関数は，今私が始めた研

究で良くないのに，高次元の判別でうまくいくとは思えない」と言った。彼の回答は覚えていないが，横浜でIASC2008の組織委員となり彼を招待したが，同じ研究であったので少なくとも10年以上研究していたことになる。研究のテーマ設定が重要である。これに対してJMPが高次元のFisherのLDFを開発し提供することは，統計ソフトの企業としては望ましいことである。筆者にとっては，わずか数日でFisherのLDFが，1）MicroarrayデータがLSDであるにもかかわらず誤分類数が0でない，2）JMPという世界的な統計ソフトの開発企業のトップ企業には，複数の専門家がいて開発にしのぎを削っていて，個人が開発したソフトよりはるかに信頼性が高い。筆者の私見であるが，判別分析の利用者は分散共分散に基づく判別関数は問題が大きいので利用すべきでないと考えるが，信頼性の高いJMPで実証研究したから断言できる。しかし，統計ソフトの企業はこれらのサポートをやめるべきでないと考えている。なくなると，判別分析の歴史が分からなくなるからである。

この表を見て「定数項が0になる場合を考慮していなかった」ことに気づいた。Program3Lでは，判別係数が0の2変数だけ省略し，判別を繰り返している。そこで，2変数に加え定数項も省いたものと2変数だけの検討をした。定数項も省いた場合，SVM1を除いた5手法の判別係数が等しく誤分類数は0であるのに対し，SVM1は誤分類数が1である。これに対して2変数だけ省くと，3個のOLDFは同じ結果である。H-SVMとSVM4は同じ結果であり，定数項は0にならない。SVM1は誤分類数が1である。この結果から，偶然定数項が0になる場合を考慮していなかったが，例え0になっても定数項は1のまま残して定数項が0になるかそれ以外になるか検討した方が良いことが改めて分かった。

表12 スイス銀行データの6手法

IC	ZERO	NP	手法	X1	X2	X3	X4	X5	X6	c	X1	X2	X3	X4	X5	X6	c
0	0	200	RIP	-1.09	0	0	-2.605	-2.827	2.0618	0	1	1	1	1	1	1	1
0	0	200	IPLP	-1.09	0	0	-2.605	-2.827	2.0618	0							
0	0	200	LP	-1.09	0	0	-2.605	-2.827	2.0618	0							
0	0	200	HSVM	-1.139	-0.567	0.1241	-2.3	-2.796	1.7961	102.38							
0	0	200	SVM4	-1.139	-0.567	0.1241	-2.301	-2.796	1.7961	102.38							
1	0	199	SVM1	-0.314	0.0698	-0.215	-1.046	-0.796	1.3739	-88.37							
0	0	200	RIP	-1.09			-2.605	-2.827	2.0618		1	0	0	1	1	1	0
0	0	200	IPLP	-1.09			-2.605	-2.827	2.0618								
0	0	200	LP	-1.09			-2.605	-2.827	2.0618								
0	0	200	HSVM	-1.09			-2.605	-2.827	2.0618								
0	0	200	SVM4	-1.09			-2.605	-2.827	2.0619								
1	0	199	SVM1	-0.662			-1.074	-1.1	1.1681								
0	0	200	RIP	-1.09			-2.605	-2.827	2.0618	0	1	0	0	1	1	1	1

0	0	200	IPLP	-1.09	-2.605	-2.827	2.0618	0						
0	0	200	LP	-1.09	-2.605	-2.827	2.0618	0						
0	0	200	HSVM	-1.448	-2.275	-2.901	1.7383	120.11						
0	0	200	SVM4	-1.448	-2.275	-2.901	1.7383	120.11						
1	0	199	SVM1	-0.258	-1.098	-0.796	1.4156	-124.6						
0	0	200	RIP		-44	48	-6348		0	0	0	1	0	1	1
0	0	200	IPLP		-44	48	-6348								
0	0	200	LP		-44	48	-6348								
0	0	200	HSVM		-44	48	-6348								
0	0	200	SVM4		-44	48	-6348								
2	0	198	SVM1		-1.173	1.8663	-251.4								
13	0	187	RIP		-5634	373.48			0	0	0	1	0	0	0
13	0	187	IPLP		-5634	373.48									
14	0	186	LP		-1.735	0.1154									
16	0	184	HSVM		-202.9	13.276									
14	0	186	SVM4		-1.734	0.1154									
			SVM1												

3.5.2　間違った分析例と正しい分析結果の一例

スイス銀行紙幣データは，IP-OLDFで（X4, X6）がMNM=0になり，MNMの単調減少性からこの2変数を含む16モデルがMNM=0になることを発見した。しかし，日本車の分析を終わった後，8月25日ごろにProgram3Lを実行すると表13の結果になる。即ち6変数を判別すると，X1，X3，X6の判別係数が0になり，スモールループの2回目からこの3変数が省かれ誤分類数は2である。本データがLSDでないことになる。Program1では，これまでの計算結果と合致しているのに，肝心のProgram3LとProgram3Sが表13の結果になる。データに間違いがないか不安になり，JMPで調べたが問題がないことが分かった。LINGOのバグと考えていろいろ悪戦苦闘し，漸く「メモリー管理異常のメッセージ」でWordが終了しなくなったことを思い出した。

表13　Program3Lの間違った出力例（8月25日ごろ）

NM	SUM	X1	X2	X3	X4	X5	X6	c
2	6	1	1	1	1	1	1	1
2	3	0	1	0	1	1	0	1
43	3	1	0	1	0	0	1	1
43	1	0	0	1	0	0	0	1

表14は，この事実を記録として残しておくことが重要と考え，CHOICEの履歴だけでなく，Matroskaと判別係数の履歴も表示するようにプログラムを修正して2016年9月2日に実行し

た結果である。表12と同じくSM1まで正しく求めている。表12は手作業のシミュレーションであり，表14はProgram3Lでは，（X1，X4，X5，X6）から（X4，X6）というBGSを直接求めることができないことが分かった。日本車データの小型車のX3が4という一定値をとるからBGSを求めることができたが，一般的に変数値がばらつく場合には簡単にBGSを求めることができないことが分かる。この4変数を省いた（X2, X3）で判別すると誤分類数が39である。新しく出力したMatrokaではSM1として（X1，X4，X5，X6）が選ばれ，SM2ではNM=39の（X2，X3）が選ばれて，スイス銀行データは，（X1，X4，X5，X6）∪（X2，X3）であることが分かる。（X1，X4，X5，X6）から（X4，X6）というBGSを求めるのは今のところ手作業になる。CoeffにSM1とSM2で最後に選ばれた判別係数が正しく出力している。これらの検証はMicroarrayデータのような高次元データでは無理で，遅まきながら日本車データとスイス銀行データという小さなテストデータで初めて検証できた。プログラム作成にはそれほど才能がないので，大きな間違いを出さずにProgram3が作成できたことに安堵した。

表14　Program3Lの正しい出力例（2016年9月2日）

					Choice						
SM	IT	T	NM	SUM	X1	X2	X3	X4	X5	X6	c
1	6	4	0	4	1	0	0	1	1	1	1
2	6	2	39	2	0	1	1	0	0	0	1
					Matroska						
SM	IT	T	NM	SUM	X1	X2	X3	X4	X5	X6	c
1	6	4	0	4	1	0	0	1	1	1	0
2	6	2	39	2	1	2	2	1	1	1	0
					Coeff						
SM	IT	T	NM	SUM	X1	X2	X3	X4	X5	X6	c
1	6	4	0	4	1.0904	0	0	2.6051	2.8271	-2.062	0
2	6	2	39	2	0	-991.4	6960.1	0	0	0	-8.00E+05

3.5.3　試行2

表15は，6変数を2回コピーして18変数のデータを作成して判別したCHOICEの表である。SM1としてX2とX3に対応する6変数を除いた12変数がSM1として選ばれMNM=0で，これを省いた6変数はMNM=39である。以上から9月2日以降の計算は正しいと考えられる。

表15　6変数を2回コピーして18変数のデータを作成して判別したCHOICEの表

NM	SUM	C1	C2	C3	C4	C5	C6	c11	c12	c13	c14	c15	c16	c21	c22	c23	c24	c25	c26	c
0	18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	12	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1
0	12	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1
0	12	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1
0	12	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1
39	6	0	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1
39	6	0	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1
39	6	0	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1
39	6	0	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1
39	6	0	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1

4. Program3Lの判別結果の検討

本章では，2015年に行った分析結果をJMPで検証した結果を説明する。

4.1　改定IP-OLDFと改定LP-OLDFによるSMの探索

表16は，Method2をProgram3L（小ループの結果を表示しない）で判別した結果である。図3でH-SVMとSVM4は判別係数が0になるものがなかった。小ループで11回繰り返しProgram3Lで判別を行ったが全係数が0でなかったので，Program1とProgram3Lの結果が同じであり，Program3Lに問題のない一つの検証結果になった。ただし，誤分類数25が表示されているのはプログラムの中で誤分類数を表すICが25以上の指定で計算を打ち切っているからである。

改定LP-OLDFは3分32秒で，66個のSMが得られ，残り5748個以上の遺伝子のMNMが0でないことになる。SN=75はNM = 9 であり，SN=74とSN=76のNM=10であるので，一瞬プログラムの瑕疵かと考えたが，改定LP-OLDFは誤分類されたケースのSVからの距離の和を最小化しているので，これは後日検討することとして多分問題はないであろう。

改定IP-OLDFは4時間29分28秒かかった。MNM=0だけの判別であれば，計算時間は気にならないが，MNM≧1の重なりのあるデータになると一般的に急に計算時間がかかることと，幾つものSheetに大きなデータを変換したSheetを含んでいるためである。69個のSMが求まり，残り5891個の遺伝子空間が線形分離可能（LSD）でない。改定LP-OLDFより計算時間がかかるが，143（=5891 − 5748）個の遺伝子を69個と3個だけ改定LP-OLDFよりSMの数が多い。また, 143個少ないよりコンパクトな癌遺伝子の候補を特定したことになると考えられる。すなわち，繰り返して判別すると，改定IP-OLDFは改定LP-OLDFよりより小さな癌の遺伝子空間を特定できることを示している。この傾向は，多分普遍的であろう。2つの判別手法が

見つけた遺伝子の部分空間が異なるが，BGSは不変であることは間違いがない。このことは，BGSを直接見つける改定Program3ができれば，検証できる。

以上から計算時間の問題はあるが，改定IP-OLDFの見つけたSMが全てLSDであるかをJMPで検証する。しかし，研究の効率性を考えれば，計算時間の少ない改定LP-OLDFで検討するのが効率的である可能性が大きい。

また，MNM=0の遺伝子の部分空間だけに注目したが，MNMが1，2，3以上の部分空間に癌遺伝子が含まれていないという，医学的な判断はできない。しかし，69個ものSMが得られており，その分析を優先すべきであろう。すなわち，「MNMが癌遺伝子の優先度を表す指標」と考えている。その延長線上で，早く見つかったSM1の方が遅く見つかるSM69より，例えば判別スコアで癌と正常が離れているのではないかと仮定している。当初は，各SMをLINGOで判別し，個別に0になることを確認のうえで，改定IP-OLDFの判別スコアから2群の平均などの検定などでこれを検証しようと考えていた。しかし，操作の簡便性と計算時間の短縮から，JMPの方が良いと判断した。

表16 H-SVM, SVM4, 改定LP-OLDFと改定IP-OLDFの探索

	H-SVM(18s)/SVM4(16s)			
	SM	小ループ	Gene	NM
1	1	11	7129	25

LP(3m32s)								RIP(4h29m28s)							
SM	小ループ	Gene	NM	SM	小ループ	Gene	NM	SM	小ループ	Gene	MNM	SM	小ループ	Gene	MNM
1	11	7129	0	51	11	6174	0	1	11	7129	0	51	11	6345	0
2	11	7111	0	52	11	6152	0	2	11	7118	0	52	11	6321	0
3	11	7097	0	53	11	6133	0	3	11	7102	0	53	11	6302	0
4	11	7085	0	54	11	6112	0	4	11	7091	0	54	11	6282	0
5	11	7068	0	55	11	6085	0	5	11	7081	0	55	11	6260	0
6	11	7053	0	56	11	6067	0	6	11	7068	0	56	11	6241	0
7	11	7038	0	57	11	6040	0	7	11	7056	0	57	11	6217	0
8	11	7023	0	58	11	6017	0	8	11	7043	0	58	11	6196	0
9	11	7005	0	59	11	5991	0	9	11	7031	0	59	11	6171	0
10	11	6991	0	60	11	5962	0	10	11	7017	0	60	11	6144	0
11	11	6974	0	61	11	5929	0	11	11	7001	0	61	11	6124	0
12	11	6953	0	62	11	5900	0	12	11	6991	0	62	11	6101	0
13	11	6936	0	63	11	5874	0	13	11	6979	0	63	11	6073	0
14	11	6921	0	64	11	5843	0	14	11	6966	0	64	11	6050	0
15	11	6906	0	65	11	5813	0	15	11	6950	0	65	11	6027	0

16	11	6886	0	66	11	5777	0	16	11	6936	0	66	11	5999	0
17	11	6869	0	67	11	5748	1	17	11	6923	0	67	11	5976	0
18	11	6852	0	68	11	5703	2	18	11	6904	0	68	11	5953	0
19	11	6833	0	69	11	5662	3	19	11	6889	0	69	11	5922	0
20	11	6815	0	70	11	5622	4	20	11	6876	0	70	11	5891	1
21	11	6800	0	71	11	5582	5	21	11	6862	0	71	11	5851	1
22	11	6780	0	72	11	5551	8	22	11	6846	0	72	11	5809	1
23	11	6762	0	73	11	5519	8	23	11	6829	0	73	11	5771	1
24	11	6745	0	74	11	5485	10	24	11	6812	0	74	11	5749	1
25	11	6724	0	75	11	5448	9	25	11	6798	0	75	11	5725	1
26	11	6703	0	76	11	5418	10	26	11	6782	0	76	11	5706	1
27	11	6689	0	77	11	5387	10	27	11	6767	0	77	11	5687	1
28	11	6673	0	78	11	5357	11	28	11	6755	0	78	11	5669	1
29	11	6655	0					29	11	6734	0	79	11	5645	1
30	11	6635	0					30	11	6719	0	80	11	5627	1
31	11	6613	0					31	11	6705	0	81	11	5605	1
32	11	6592	0					32	11	6683	0	82	11	5579	1
33	11	6565	0					33	11	6664	0	83	11	5555	1
34	11	6549	0					34	11	6648	0	84	11	5527	1
35	11	6533	0					35	11	6630	0	85	11	5499	2
36	11	6510	0					36	11	6613	0	86	11	5481	2
37	11	6482	0					37	11	6594	0	87	11	5464	2
38	11	6456	0					38	11	6582	0	88	11	5442	2
39	11	6435	0					39	11	6566	0	89	11	5422	2
40	11	6416	0					40	11	6550	0	90	11	5403	2
41	11	6394	0					41	11	6534	0	91	11	5363	2
42	11	6370	0					42	11	6515	0	92	11	5343	2
43	11	6349	0					43	11	6501	0	93	11	5316	2
44	11	6328	0					44	11	6482	0	94	11	5300	2
45	11	6303	0					45	11	6468	0	95	11	5279	2
46	11	6286	0					46	11	6447	0	96	11	5256	2
47	11	6263	0					47	11	6426	0	97	11	5237	2
48	11	6237	0					48	11	6406	0	98	11	5210	2
49	11	6213	0					49	11	6383	0	99	11	5183	2
50	11	6191	0					50	11	6364	0	100	11	5154	3

4.2　69個のSMのJMPによる検証

LINGOの整数計画法は，分枝限定法を用いている。ユーザーは分岐の方向を横展開と縦展開のいづれかを選ぶことができる。筆者は，この点に関しては経験がないのでデフォルトの横展開のまま利用している。一つの仮説として，「SMの選ばれる順は，場当たり的でなく，2

群の判別スコアが離れているものから選ばれている」のではないかと考えている。このためには，選ばれた69個の改定IP-OLDFの判別スコアを計算し，2群の差のt検定を行えば分かるのではないかと考えてきた。時間があれば実証研究しようと考えていたが，時間がかかるので，簡単にJMPのロジスティック回帰で判別し誤分類が0になることを確認後，FisherのLDFのSM1からSM69の誤分類数が増加傾向を示せば私の仮説が検証できると考えた。

表17のSMはSM1からSM69を表す。「Gene」はProgram3Lが分析対象とした遺伝子の数である。最初は全遺伝子の7129個を判別し，IT2=11に設定し10回判別を小ループで繰り返して，「N_SM」に示す11個の判別係数が0でないSM1を選んだ。SM2はこの11個を除いた7118個の遺伝子を10回判別を繰り返し「N_SM」に示すSM2の遺伝子16個を選んだ。「N_BGS」は2015年末に手作業で求めたBGSの遺伝子数である。複数の数字は，例えばSM12の中に8個と2個のBGSがあることを示す。今回JMPで個々のSMを判別して，「NM_logistic」列のロジスティック回帰によるNMと，「NM_LDF」列のFisherのLDFによるNMを求めた。Firth[5]はLSDの判別で，収束計算は不安定になり，判別係数の95% CIは，0を含む大きな範囲になることを指摘した。通常の推測統計学的判断では，このようなモデルは選ばない。JMPではFisherが提案した最尤推定法を用いて，ヘシアン行列から標準誤差を推定している。このようなコンピュータパワーを利用した方法は，理論分布から紙と鉛筆で標準誤差の計算式が導き出された伝統的な推測統計学と一線を画すべきだと考えている。Method2から100個の判別係数と誤分類確率が計算でき，これらの分布から95% CIを設定しているが，平均を中心として2.5%点と97.5%点が対象でないので標準誤差をあえて計算していない。筆者は，「これらが観測され，ROC曲線上で判別境界を動かしてNM=0になるものがあり，かつMNM=0であれば，ロジスティック回帰はLSDを正しく判別できた」と扱うことにした。69個のSMのNMは全て0であったのでProgram3Lの一番重要な機能は，問題がないと考えられる。

表17　69個のSMのJMPによる検証

SM	Gene	N_SM	N_BGS	NM_logistic	NM_LDF	MNM
1	7129	11	4	0	1	0
2	7118	16	6	0	4	0
3	7102	11	6	0	3	0
4	7091	10	3	0	0	0
5	7081	13	5	0	3	0
6	7068	12	9	0	1	0
7	7056	13		0	3	0
8	7043	12	4	0	1	0
9	7031	14	5	0	3	0

10	7017	16	6	0	3	0
11	7001	10	6	0	2	0
12	6991	12	8, 2	0	3	0
13	6979	13	5	0	2	0
14	6966	16		0	6	0
15	6950	14	7	0	3	0
16	6936	13		0	3	0
17	6923	19	9	0	3	0
18	6904	15	8	0	4	0
19	6889	13	9	0	3	0
20	6876	14	7	0	4	0
21	6862	16		0	1	0
22	6846	17	9	0	3	0
23	6829	17		0	6	0
24	6812	14	10	0	1	0
25	6798	16	10	0	6	0
26	6782	15	12, 3	0	10	0
27	6767	12	6	0	4	0
28	6755	21		0	12	0
29	6734	15	9	0	6	0
30	6719	14	9	0	4	0
31	6705	22	10	0	2	0
32	6683	19	13	0	5	0
33	6664	16		0	7	0
34	6648	18		0	9	0
35	6630	17		0	6	0
36	6613	19		0	8	0
37	6594	12		0	9	0
38	6582	16	12	0	7	0
39	6566	16		0	8	0
40	6550	16		0	5	0
41	6534	19	12	0	6	0
42	6515	14	11	0	11	0
43	6501	19		0	9	0
44	6482	14		0	11	0
45	6468	21		0	8	0
46	6447	21		0	6	0
47	6426	20		0	8	0
48	6406	23		0	11	0
49	6383	19		0	6	0

50	6364	19		0	8	0
51	6345	24		0	6	0
52	6321	19		0	8	0
53	6302	20		0	9	0
54	6282	22		0	10	0
55	6260	19		0	10	0
56	6241	24		0	14	0
57	6217	21		0	8	0
58	6196	25		0	11	0
59	6171	27		0	12	0
60	6144	20		0	9	0
61	6124	23		0	10	0
62	6101	28		0	12	0
63	6073	23		0	12	0
64	6050	23		0	15	0
65	6027	28	28	0	11	0
66	5999	23		0	14	0
67	5976	23		0	14	0
68	5953	31	25	0	9	0
69	5922	31	30	0	12	0

これに対して，FisherのLDFは0から15まで緩やかに増加傾向があることが分かる。図6は，このNMをSNでもって単回帰分析を行った。R^2=0.72，F=1073.15（$p < 0.0001$）であり，定数項は棄却できないけれど，回帰係数のt値は13.16（$p<0.0001$）であるので，本研究の作業仮説は正しいのではないかと考えられる。

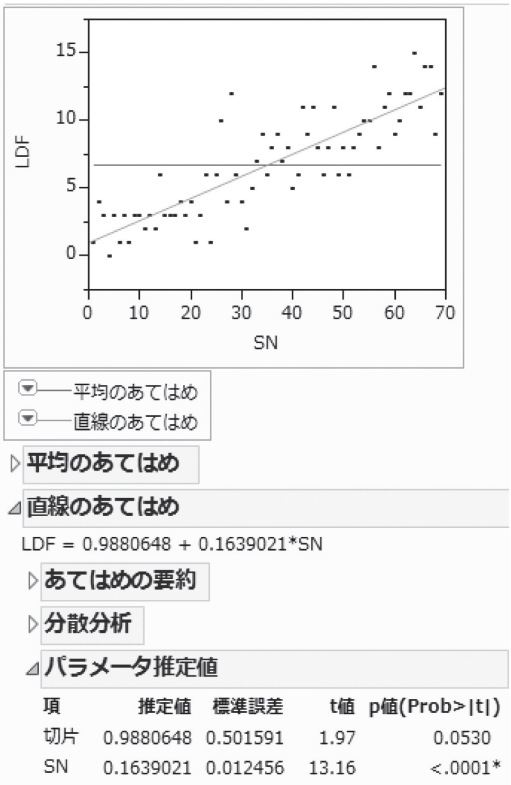

図6　FisherのNMをSNで単回帰分析

4.3　MNMが1以上の15個の部分空間のJMPによる検証

表18は，69個のSMの後で15個のMNM=1の部分空間が選択され，SM=85のMNMが2であることが分かる。

表18　15個のMNM=1のJMPによる検証

Other	Gene	N_SM	MNM	logistic	Other	Gene	N_SM	MNM	logistic
70	5891	40	1	1	78	5669	24	1	4
71	5851	42	1	1	79	5645	18	1	12
72	5809	38	1	3	80	5627	22	1	10
73	5771	22	1	4	81	5605	26	1	10
74	5749	24	1	8	82	5579	24	1	8
75	5725	19	1	9	83	5555	28	1	8
76	5706	19	1	5	84	5527	28	1	5
77	5687	18	1	7	85	5499	18	2	

MNMが1のSMが70から84の15個の部分空間を加えて回帰するとR^2=0.62，F=132.21（p＜0.0001）であり，定数項と回帰係数は棄却できた。しかし，69個のSMの場合に比べ予測は悪くなった。

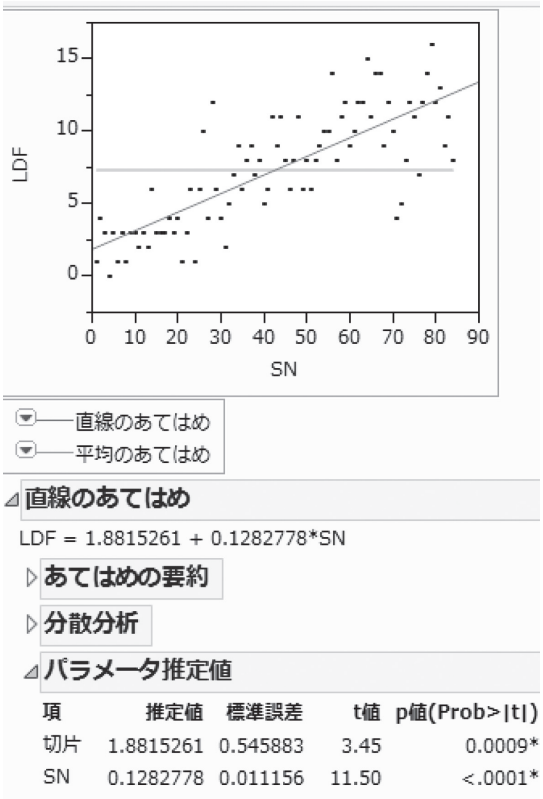

図7　FisherのNMをSNで単回帰分析（MNM=0と1の84個）

5. LINGO Program3によるAlonらのデータの検証

Alonらのデータは，62人の2000個の遺伝子データで，6種の判別の識別子を持つデータ件数が最小のデータである。

5.1　全体の分析

Program1で2000個の遺伝子を6個のMP-based LDFsで判別分析すると，6秒で6個のLDFの全てがNM=0で判別できる。そこで，Excel上のProgram1から出力した判別係数をJMPに読み込んで，判別係数を0かそれ以外になる係数の数を調べるために，名義尺度に変更して調べた。図8は，この度数表である。定数項を含むので合計の2001個から1を引いた2000個の判

別係数がある。改定IP-OLDFは61個，改定IPLP-OLDFと改定LP-OLDFは39個，3個のSVMは全判別係数が0でない。1回だけの判別分析では，改定IP-OLDFと改定LP-OLDFが改定IP-OLDFより優れている。

図8　全データの特徴抽出

5.2　Program3Lの判別結果

表19は，Method2をProgram3Lで判別した結果である。図8でH-SVMとSVM4は判別係数が0になるものがなかったので，小ループで11回繰り返し判別を行ったが全係数が0でなかった。改定LP-OLDFは，58個のSMが得られ，残り867個の遺伝子のMNMが0でないことになる。NMが1，4，5，7，7の後8，9，10がなくIC=11と設定したので打ち切られている。改定IP-OLDFはIC=2と設定し21分45秒かかった。IC=11と設定すると，全遺伝子を6個のLDFで分析しても6秒なのに，SMを順次除外して判別して遺伝子が確実に少なくなっていくのに単にCHOICEで0にしているだけで実際には全遺伝子の判別を行っている。それが誤分類数が増えると計算時間がかかる一つの理由である。これは，Excelファイルがいくつもの変換加工したSheetを含んで最適化を行ったためと計算結果を比較して分かった。結局64個のSMが求まり，残り848個の遺伝子がMNMが1以上であることが分かる。MNM=1が8個の部分空間になり，MNMが2以上の部分空間委649個の遺伝子がある。MNM=2の最初の部分空間には35個の遺伝子が含まれていた。

表19　改定LP-OLDFと改定IP-OLDFの探索

	LP(5m29s)				RIP(21m45s)					
SM	小ループ	Gene	MNM	N_Gene	小ループ	Gene	MNM	N_Gene	NM_logistic	NM_LDF
1	11	2000	0	17	11	2000	0	15	0	0
2	11	1983	0	16	11	1985	0	14	0	2
3	11	1967	0	17	11	1971	0	13	0	4
4	11	1950	0	17	11	1958	0	15	0	3
5	11	1933	0	19	11	1943	0	19	0	2
6	11	1914	0	19	11	1924	0	12	0	3
7	11	1895	0	15	11	1912	0	16	0	2

8	11	1880	0	21	11	1896	0	11	0	5
9	11	1859	0	14	11	1885	0	13	0	3
10	11	1845	0	24	11	1872	0	16	0	1
11	11	1821	0	17	11	1856	0	15	0	4
12	11	1804	0	15	11	1841	0	19	0	2
13	11	1789	0	17	11	1822	0	22	0	3
14	11	1772	0	17	11	1800	0	15	0	4
15	11	1755	0	14	11	1785	0	15	0	2
16	11	1741	0	14	11	1770	0	14	0	3
17	11	1727	0	16	11	1756	0	15	0	3
18	11	1711	0	21	11	1741	0	20	0	4
19	11	1690	0	17	11	1721	0	17	0	1
20	11	1673	0	18	11	1704	0	16	0	19
21	11	1655	0	21	11	1688	0	18	0	1
22	11	1634	0	19	11	1670	0	21	0	4
23	11	1615	0	20	11	1649	0	16	0	3
24	11	1595	0	15	11	1633	0	13	0	3
25	11	1580	0	21	11	1620	0	16	0	1
26	11	1559	0	17	11	1604	0	21	0	7
27	11	1542	0	20	11	1583	0	13	0	3
28	11	1522	0	21	11	1570	0	12	0	4
29	11	1501	0	17	11	1558	0	14	0	4
30	11	1484	0	18	11	1544	0	14	0	1
31	11	1466	0	15	11	1530	0	21	0	6
32	11	1451	0	21	11	1509	0	16	0	5
33	11	1430	0	16	11	1493	0	15	0	4
34	11	1414	0	18	11	1478	0	17	0	4
35	11	1396	0	20	11	1461	0	17	0	0
36	11	1376	0	16	11	1444	0	15	0	7
37	11	1360	0	17	11	1429	0	18	0	5
38	11	1343	0	16	11	1411	0	19	0	5
39	11	1327	0	20	11	1392	0	19	0	5
40	11	1307	0	21	11	1373	0	20	0	3
41	11	1286	0	21	11	1353	0	16	0	3
42	11	1265	0	22	11	1337	0	16	0	6
43	11	1243	0	18	11	1321	0	15	0	7
44	11	1225	0	24	11	1306	0	17	0	3
45	11	1201	0	18	11	1289	0	21	0	4
46	11	1183	0	22	11	1268	0	16	0	6
47	11	1161	0	20	11	1252	0	17	0	5

48	11	1141	0	20	11	1235	0	15	0	6
49	11	1121	0	21	11	1220	0	17	0	6
50	11	1100	0	23	11	1203	0	20	0	5
51	11	1077	0	21	11	1183	0	18	0	5
52	11	1056	0	23	11	1165	0	24	0	6
53	11	1033	0	25	11	1141	0	16	0	4
54	11	1008	0	24	11	1125	0	22	0	6
55	11	984	0	27	11	1103	0	25	0	5
56	11	957	0	28	11	1078	0	20	0	8
57	11	929	0	23	11	1058	0	20	0	9
58	11	906	0	39	11	1038	0	20	0	8
59	11	867	1	38	11	1018	0	24	0	4
60	11	829	4	37	11	994	0	25	0	5
61	11	792	5	37	11	969	0	25	0	8
62	11	755	7	38	11	944	0	29	0	8
63	11	717	7	31	11	915	0	28	0	11
64	11	686	11	686	11	887	0	39	0	7
65					11	848	1	24	8	10
66					11	824	1	28	7	8
67					11	796	1	31	4	6
68					11	765	1	31	4	9
69					11	734	1	20	7	13
70					11	714	1	19	8	9
71					11	695	1	20	9	10
72					11	675	1	26	7	7
73					11	649	2	649(35)	5	6

表26の「NM_logistic」はロジスティック回帰によるNMである。64個のSMのNMは0であり，Program3Lは正しく処理していることが2個のデータで分かった。しかし，MNM=1の8個の部分空間ではNMは4から9まででばらついていて，n = 62と少ないことを考えると判別成績は悪い。NM_Fisherの73個の部分空間を誤分類数をSNの値で回帰して，図9が得られた。R^2=0.35，F=37.74（$p < 0.0001$）であり，定数項と回帰係数のt値は3.54（$p<0.0033$）と6.14（$p<0.0001$）で棄却される。しかし，Golubほど仮説1を支持していないようだ。

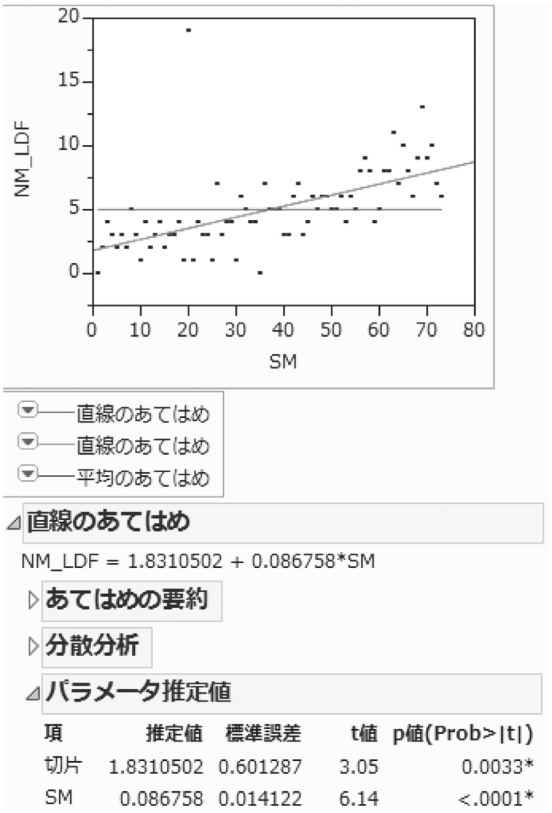

図9 FisherのNMをSNで単回帰分析

6. 終わりに

本研究では，GolubらとAlonらのデータを用いて，Method2のために開発したProgram3の検証を試みた。このデータに焦点を当てたのは，1万個以下の遺伝子数の少ない3個のデータであること，カルフォルニア大学のTamayo教授のグループによる研究であり，医学的な判断を将来お願いしようかと考えているためである。

今回検討した点は次のとおりである。

1) Method2で，Microarrayデータは，排反なSMの和集合であるという驚く構造を発見した。整数計画法でMNM=0になるものに注目しこの研究を行ってきた。そして，正常と癌患者をMNM=0で判別するSMを見つけたので，これらに関係する遺伝子を癌遺伝子と考えた。これが選ばれる順は，多分判別が容易すなわち2群が離れているものから選ばれるのではないかと考えた。そして，ロジスティック回帰がMNM=0の判別をほぼ正しくNM=0で判別することを利用して，選ばれたSMがMNM=0であることを確認した。一方，Fisherの

LDFが一般的にMNM=0のデータを正しく判別できないことを利用して，判別精度は悪いが，誤分類数が多いほど2群の距離が近いことに対応すると考えた。GolubらとAlonらのデータでこれが確認できた。しかし，よく考えてみれば，MNMが1以上のものも計算してみると，どうもSMに含まれる遺伝子数も増加傾向にあるようだ。少ない遺伝子で癌と正常を離れた距離で判別するほど癌であるという優先度が高いと考えることは，間違いでもないのではと考えている。

2）LINGO Program3を短期間で作成し，分析結果に何ら不安を覚えず急いで研究を行ってきた。しかし本来であれば，小さなテストデータでプログラムのテストを行うべきであった。RGでもかなり多くの研究者が筆者の驚く研究を読んでいるのにバグがあり，重要な結果に間違いがあったらと思うと，大いに反省すべきである。幸い「日本車データ」で考えた通りの結果が確認でき，Theory2の骨子は大筋問題が内容である。しかし，追試検証した多くの結果が多分Windows10の不具合で，メモリー管理ミスの後，計算結果が大きく異なっているのに何のエラーも出ず出力されたことが予備のPCで検証して分かった。これらを以前の結果とフォルダーを分けずに格納したので，問題のあると思われる結果の削除を行っていないので，今後間違ってその結果を発表しないか最新の注意を払う必要がある。

（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. Anderson, E. (1935).“The irises of the Gaspe Peninsula.”*Bulletin of the American Iris Society*, 59, 2-5.
3. 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.
4. Cox, D. R. (1958)“The regression analysis of binary sequences (with discussion).”*J Roy Stat Soc* B 20: 215-242
5. Firth, D. (1993).“Bias reduction of maximum likelihood estimates.”*Biometrika*, vol. 80: 27-39
6. Fisher, R. A. (1936).“The Use of Multiple Measurements in Taxonomic problems.”*Annals of Eugenics*, 7, 179-188.
7. Fisher, R. A. (1956). *Statistical methods and statistical inference*. Hafner Publishing Co.
8. Flury, B., Riedel, H. (1988). Multivariate Statistics: A Practical Approach. Cambridge University

Press.

9. Friedman, J. H. (1989). "Regularized Discriminant Analysis." *Journal of the American Statistical Association*, 84/405, 165-175.
10. 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.
11. Glover, F. (1990). "Improved linear programming models for discriminant analysis." *Decision Sciences*, 21, 771-785.
12. Jeffery, IB. Higgins, DG. Culhane, AC. (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/
13. Miyake, A., 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., 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., 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 (1)*, Oct. 28, 2015, 1-4.
35. *_____ (2015e). "Feature Selection of three Microarray data." *Research Gate (2)*, Nov.1, 2015, 1-7.
36. *_____ (2015f). "Feature Selection of Microarray Data (3) - Ship et al. Microarray Data." *Research Gate (3)*, 2015, 1-11.
37. *_____ (2015g). "Validation of Feature Selection (4) - Alon et al. Microarray Data." *Research Gate (4)*, 2015, 1-11.
38. *_____ (2015h). "Repeated Feature Selection Method for Microarray Data (5)." *Research Gate (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 (6)*, Nov. 11, 2015, 1-10.
40. *_____ (2015j). "Matroska Trap of Feature Selection Method (7) -Golub et al. Microarray Data." *Research Gate (7)*, Nov. 18, 2015, 1-14.
41. *_____ (2015k). "Minimum Sets of Genes of Golub et al. Microarray Data (8)." *Research Gate (8)*, Nov. 22, 2015, 1-12.

42. *_____ (2015l). "Complete Lists of Small Matroska in Shipp et al. Microarray Data (9)." *Research Gate (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. et al. Microarray Data (11)." *Research Gate (11)*, Dec. 4, 2015, 1-34.
45. *_____ (2015o). "Feature Selection of Singh et al. Microarray Data (12)." *Research Gate (12)*, Dec. 6, 2015, 1-89.
46. *_____ (2015p). "Final List of Small Matroska in Tian et al. Microarray Data." *Research Gate (13)*, Dec. 7, 1-160.
47. *_____ (2015q). "Final List of Small Matroska in Chiaretti et al. Microarray Data." *Research Gate (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) "Discriminant Analysis of the Linear Separable Data -Japanese automobiles-." *Journal of Statistical Science and Application*, vol. 4, No. 07-08, 165-178. doi : 10. 17265/ 2328-224X/ 2016, 0708, 001.
51. *_____ (2016c). "The Best Model of the Swiss Banknote Data-Validation by the 95% CI of error rates and discriminant coefficients -." *Optimization, and Information Computing*, Vol.3, 322-335, 2015. DOI: 10.19139/soic. 20151202.
52. *_____ (2016d). "The K-fold Cross Validation for Small Sample Method." *Data Analytic 2016*, 1-6.
53. Shinmura, S. (2016f). *The New Theory of Discriminant Analysis after R Fisher*, Springer
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. Simon N, Friedman J, Hastie T, Tibshirani R (2013). "A sparse-group lasso." *J. Comput. Graph. Statist*, 22:231-245
56. Singh, D. et al. (2002). "Gene expression correlates of clinical prostate cancer behavior." *Cancer Cell: March 2002*, Vol. 1, 203-209.
57. Stam, A. (1997). "Nontraditional approaches to statistical classification: Some perspectives on lp-norm methods." *Annals of Operations Research*, 74, 1-36.

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. 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.
61. _____訳著（1986）.『SASによる回帰分析の実践』. 朝倉書店.
62. _____（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-タのモデル決定」。
『成蹊大学経済学部論集』，27/1,180-203.
63. _____（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.
64. 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.
65. 新村秀一（2004）.『JMP活用　統計学とっておき勉強法』. 講談社.
66. _____（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.
67. _____（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.
68. _____（2010a）.『最適線形判別関数』. 日科技連出版.
69. _____（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.
70. _____（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.
71. _____（2011b）.『数理計画法による問題解決法』. 日科技連出版.
72. _____（2012）.「コラム「SAS/JMPとの歩み」，SAS Technical News，春，夏，秋，冬号」.
73. _____（2015a）.「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.
74. **_____（2016a）.「判別分析の新理論と遺伝子解析」,『第9回コンピューテーショナル・インテリジェンス研究会』，77-84.
75. **_____（2016b）.「判別分析の新理論と遺伝子解析のための新手法2」,『成蹊大学経済学部論集』，第47巻第1号，43-77.
76. 田邉國士（2011）.「応用数理の遊歩道（67）帰納という原罪」.『応用数理』，304-309.

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 * or ** before author's name from the Research Gate and Research Map.