「判別分析の新理論」と応用研究としての「癌の遺伝子解析」

－なぜ1970年から行われた研究が成功しなかったのか？－

新　村　秀　一

1. はじめに

筆者は1971年に京都大学の数学科の大学院を落ちて，できたばかりの住商情報システム㈱（現SCSK）とNECの両社とも学部の採用が終わっていたが採用された。SCSKは役員面接で成績が悪いといわれながら採用された。NECは学部の試験問題を作るのが面倒ということで大学院生と同じ試験を受け白紙であったが不思議なことに採用されたので，SCSKに断りに行くと，「優が少ない」といっていたNECからきていた専務が「秀才きらめくNECにいっても目が出ない。これからの若い産業が面白い」と説得され，できたばかりのSCSKに入ることにした。大学院に落ちたのは，面倒なので水泳部の部活のやりすぎということにしてきたが，Gaussにあこがれ複素関数論を選んだことが大きい。岡潔が複素関数論の多くの問題を解決し，難問しか残っていないという状況にあることを知らなかった。すなわち教員も業績が出ない状況であったと考える。3年次の指導教授は岡潔の弟子ということで選んだが，最初のゼミの日に行くと自分が読んでいたフランス語の論文をコピーして，来週から輪読するということである。辞書を引いたが18世紀のフランス語は古語であるらしく載っていない。翌週ゼミで先生に「辞書にも載っていなく読めませんでした」というと，「将来研究者になるのなら，過去の論文から何を述べていて，何が問題かを読み通れないようであれば，あきらめろ」といわれた記憶がある。ハーバード大学で活躍していた広中平祐氏の代数幾何の分野に人気があり，九大の教授になった私の友人は永田ゼミにいったが学部の3年生にそのような指導を受けていないことを確認した。結局複素関数論で業績がないので教養部の数学の助教授であった親切な藤家ゼミに移った。そこでは辻正次先生の複素函数論（1968版）で具体的な等角写像などの585頁の日本語のテキストを2人のゼミ生で輪読し，自分の能力にあっていて有意義であった。しかし社会人になって数年もなぜ数学で研究者になれなかったかを自問し続けた結果，「すでに人が完成した立派な学問体系を理解し追従することは研究者の使命でなく，新しい分野を開拓するということを理解していなかった」ことに落ち着いた。

1971年にSCSKに入って10月ごろ迄の研修の途中の6月ごろにNECの大阪に出向になり，暫くして課長に連れられて大阪府立成人病センターの循環器医長の野村裕先生のところに連れていかれた。そこで課長からNECとの共同プロジェクトの「心電図の自動診断システム解

析」の手足に使ってくれということである。暫くして，先生から数冊の心電図の本と東大出版会の高橋晄正（編）『計量診断学』を渡された。彼は判別分析を医学応用し計量診断学の先導役であり，計量経済，計量心理学などの当時はやりの『計量…学』の先導役であった。先生から暫くして「理解できましたか？」と聞かれ，何と答えたか覚えていない。その後で，心電図の自動診断で集めた多分2000症例100個以上の計測値の入ったMTを渡された。そして「NECからは手足のように使ってくださいということですが，数学科も出たので統計手法を使って正常所見と10個以上の異常所見を分ける診断論理を開発してください。良ければ商業化を予定している心電図自動診断システムに組み込んであげる」といわれた。多分不安な表情を見せたのであろうが，「できなくてもすでに自分は枝分かれ論理で開発済み」ということである。偉い研究者は大きな目標を立て，途中経過を毎年報告するようだ。私もその洗礼を受け，1973年に「二段階重みづけによるスペクトル診断」を第12回日本ME学会大会で発表しているし，1972年には関西ME学会で発表済みである。1年後にすでに判別分析に疑問を抱いて，ロジスティック回帰の基本的な考えである「ある計測値が連続的に大きくあるいは小さくなれば，正常から異常になる確率が0から1へ連続的に近づく」と同じことをベイズの定理で試みた。阪大出身の中村博士（工学）からはベイズを拡大解釈していて発表には反対であったが，彼の言い分をじっと聞いたうえで野村先生は共著者にしてくれた。

以上のような私的なことに紙面を割いたのは，「なぜ筆者が分かるような判別分析の4つの深刻な問題」を世界中の誰も指摘しなかったかに大きな疑問を持っているからである。さらに第5の新理論の応用問題は公的な研究費を世界中で使いながら，周辺の統計研究家が「いつになったらBigデータの応用としてのMicroarrayデータの分析に関して成果を出すのか？」と冷ややかに見られていた研究テーマである。それが僅か54日で統計的には「高次元の遺伝子空間が簡単に線形分離可能（LSD）な部分空間に分割できる」ことを示した。癌と正常が高次元空間で完全に分離しているという基本的な認識がないため混迷を極めたのであろう。さらに10個から40個前後の少数の遺伝子の組のSmall Matryoshka（SM）で完全に2群で判別できる。このことは，血液で癌の検査を10万円前後で行っている現在の診断制度を改善できることを示す。さらに癌の悪性度指標と考えられるものを見つけたが，症例による検証が必要であり，Microarrayデータを公開している米国の6研究グループに共同研究をResearch Gateで呼びかけたが，対応がなく医学応用での実証研究が行えず苦戦している。

現在私の考える結論は以下の通りである。筆者は院に入る前に学部で数学者への道を閉ざされた。それでも社会人になってその後遺症を克服するのに数年かかった。多くの統計やオペレーションズ・リサーチ（経営科学）の研究者は数学に憧れがあり，自らは純粋理論でなく応用研究の学徒であるのに対して，無意識に実証研究が重要という認識を下に見て避けていることが現実の問題が見つけられない理由ではないかと考える。非常に幸運なことに医学

診断で統計分析を用いた研究者への窓口が開かれた。統計はデータが異なるだけでいろいろな分野に応用できる「学際的な学問」である。参入障壁は高いが，医学データは競争の激しい医学研究で確立された信頼性が他の分野に比べて著しく高くて信頼性があるデータという利点がある。そこで，理論研究をあきらめて応用研究ということに後ろめたさはなく，実証研究を積極的に行い，既存の理論が現実のデータに合わない点を素直に観察し考察したことが幸いしたと考える。統計やオペレーションズ・リサーチ（経営科学）は応用研究であるにもかかわらず，現実を軽視して「データが正規分布であるという一方的な仮説を前提とし，数式を展開することで理論研究を行っているという錯覚が問題である」と考える。その結果「統計の新しい研究対象として注目され，1970年以降七転八倒し多くの研究論文が生産されたが，革新的な結果が全く得られなかった」と考えるべきであろう。

2. Fisherと判別分析

2.1 推測統計学

Fisherと同世代の統計学者は，統計学を記述統計から何とか数学を目標として科学的な学問にするため「推測統計学」を考えた。そのからくりは，2つある。

⑴ 分析対象のデータをGauss分布と仮定することである。Gaussが2点間の測量を数回繰り返して得られる観測値の分布がGauss分布という指数関数であることを示した。統計でも私が社会人になって勉強した書籍の一部ではGauss分布といって筋を通す書籍もあったが，いつの間にか正規分布に置き換わった。研究者は，先行研究の成果を尊重すべきであるのに，正規分布と呼ぶことは今でも統計の横暴と考えている。Fisher以後の後継者は「正規分布と呼ぶことで，「普遍的というイメージを定着させたい」のであろうが，これが判別分析の大きな問題である。重回帰分析などでは，誤差分布が正規分布を仮定することに大きな欠点は見つけられない。しかし，後述するが「判別規則が非常に簡単であり，問題点が容易に顕在化し，判別分析にはいくつも問題がある」のに世界中で誰も気づかなかった点が問題である。またResearch Gateでは，私の論文が斬新でスマートというコメントがあるが日本で褒められたことはない。

⑵ 背理法

統計学を数学に近づけるための仕組みが背理法という分かりにくい論理である。高校数学で背理法を習うが，多くの高校では受験に関係がないので教えないことが多いと考えられる。「Aを仮定すればBが導かれる。しかし，Bは二元論の「有(1)/無(0)」で有り得ない事象である場合，Aの仮説は間違っていると考える。統計ではさらにデータから事象の生じる確率を計算し，一つの目安として農事試験などでは小標本しか得られないので，5％以下を有りえない事象とみなし，棄却（否定）する。しかしt 値や相関係数が同じであっ

てもデータ件数が増えていけばp値は限りなく小さくなり，Bigデータの帰無仮説は棄却される運命にある。田邉（2011）によれば，棄却の論理的な意味は「仮説は真ではあるがその下で非常にまれな事象が起きたか，あるいは仮説自体が真でないかのどちらかである」とFisher（1936，1956）自身が述べていると紹介している[1]。すなわち，彼は対象とする現象が正規分布でないことがあることを認めたうえで，その場合は推測統計学的結論が正しくないことを指摘している。しかし，この点を明確に記述したテキストは少ないし，統計教育でもはっきりと教えていない。判別分析は，判別する2群が正規分布と仮定（Fisherの仮説）して導かれている。このため，重回帰分析と同じく推測統計学と勘違いしている人もいるが，Fisherは誤分類確率や判別係数の標準誤差を定義していないので，伝統的な推測統計学ではない（問題4）。

⑶　今日の高野山が具現化する高野山ワールドは，空海ブランドでその後の後継者によって体系化された部分が大きい。統計でも「Fisherの仮説」や「Fisherのアイリスデータ」などもFisherブランドと考えられるが，今となっては誰が言い出したことか事実関係を調べることが難しい。

2.2　Fisherの線形判別分析

Fisherは「Fisherの仮説」に基づいてFisherの線形判別関数（LDF）を導いた。これは説明変数で表されるp次元空間でデータ全体の全分散に対して2群間の群間分散との比すなわち「相関比」を最大化して導かれたという説明が行われている。そして多くのテキストが全く意味のない偏微分で最適解を求める説明がなされている。この基準は多分後で考えられたのではないかと考えるが，詳しい資料の入手が困難で推測の域を出ない。英明なFisherは，Gauss分布が ê(－(x－m)^2/2s^2)/(SQRT(2*pai)*s^2) で表される指数関数であり，2群の平均値のm_1とm_2だけが異なる正規分布f_1とf_2と考えれば，その比の対数が簡単に次のような1次式になることをすぐに閃いたと推測できる。

$$\log(f_1/f_2)=\log[\hat{e}\{-(x-m_1)^2/2s^2+(x-m_2)^2/2s^2)\}=[\{(x-m_2)^2-(x-m_1)^2\}/2s^2]$$
$$=(m_1-m_2)/s^2*x+(m_2^2-m_1^2)/(2*s^2) \quad (1)$$

もし$m_1=-m_2$で 原点を判別境界に取れば$f_1=f_2$なので$\log(f_1/f_2)=0$になり，原点の判別スコアは（m_1-m_2）/s^2*x+($m_2^2-m_12$)/(2*s^2)=0 という1次式になる。判別境界を動かせば0

[1] 同書の日本語訳（岩波書店）は読んだがこのような記述は見つけれなかった。多分原著が何版も改定されているためと考える。

に代わって負から正の値が右辺に代入され判別スコアを表す。Fisherは今日のように便利な計算機のない時代に指数関数である正規分布を考えれば，簡単に1次式で判別関数を定義できることから定式化したと考える。

これを相関比最大化で同じ1次式が得られたというのは，後で考えられたことと考える。そして，これを偏微分で説明を行うのは不思議である。統計の研究者や利用者は，不思議なことに数理計画法による最適化に最も遠い存在である。実データに合わせて偏微分で最適解を求めることを避けて，簡単に1次式を計算機環境のない時代に判別分析の世界を開いたことに意味がある。Fisherは最尤推定法の提案者でもある。彼はデータに適合するFisherのLDF（F-LDF）を決して最尤推定法で求めることはしなかった。多次元の正規分布を想定した場合，単にp変数の分散共分散を求めることで，簡単に1次式のLDFが求まる。このため統計ユーザーは，最大値/最小値と極大値/極小値の違いを知ることなく，その煩わしさから解放された。そして統計は数理計画法に比べて優位性が享受できた。この他定義域で唯一の解しか求めないことが重要である。最適解が部分空間を含めて多数あっても，決してそれを見つけることができないので「癌の遺伝子解析にこれらの判別関数は全く役に立たない」[2]。

さらに「Fisherの仮説」を満たすか否かの良い検定統計量がないのが実情である。FisherはAnderson（1945）が集めて今日判別関数の評価に用いられている「Fisherのアイリスデータ」で検証を行っている。しかし，4変数であり他の判別関数と優劣を比較するには適していないので今後は用いるべきでない。また，彼か彼の同世代の誰かが，平均の他，分散共分散行列が異なる場合に2次判別関数（QDF）を提案していることである。このことは，創業者らは実際のデータは「Fisherの仮説」を満たさないデータもあることに注意を払っていたことを示す。しかし理論でもソフトでも使いやすいので，これ以降は分散共分散行列に基づく判別理論が発展し主流と考えられている。変数が一定値を取れば，逆行列が計算できないので，それを可能とする一般化逆行列の技術が完成した。そして正則化判別関数（RDA, Freidman;1989）や，重回帰分析や判別分析の係数を0にすることで「癌の遺伝子解析」で癌遺伝子を特定することを目的としたと思われるLASSO（Simon et al., 2013）が注目されている。ある世界的に有名な出版社からLASSOに関する解説書が出ている。Springerから出版に際して参考にしようとして読んだが，100頁前後で行き詰った。600頁の大著であり，厳密な数式の展開であるが，実データの検証が一つもない。共著者は，ものすごく強靭な精神と意思の持ち主である。自分たちの理論が現実に適用できるかできないか不安に思わないだろうかと

[2] 相関比最大化基準による判別関数は，線形分離可能なデータ（LSD）の代表であるMicroarrayデータの誤分類数（NM）を必ず0にできないので，全く役に立たない。現在判別分析の主流であるSVMはLSDを正しく判別できるが，2次計画法（QP）を用いているので，部分空間の最適解を見つけることができない。

考えた。2015年から2年以上たち，彼らの目的の一つは筆者が簡単に解決したが，LASSOはいまだに良い結果が出たという報告はないようだ。

2.3 Fisherの仮説を満たすデータは少ない

Fisherの仮説を満たすデータは，F-LDFの誤分類数（NM）が最小誤分類数（MNM）に収束する。しかし現実のデータではFisherの仮説を満たすデータは少なく，その場合NMはMNMより大きく乖離する。また，Fisherの仮説を満たす良い検定法がないので，統計ソフトの使いやすさから適/不適にかかわらず利用されている。大きな間違いを避ける意味で，QDFやロジスティック回帰と比較検討する方が良いであろう。あるいは，NMとMNMの乖離の程度で判断できる。

これに対して，式(2)で表されるロジスティック回帰は，F-LDFとQDF,その後のRDAやLASSOのように分散共分散行列をもとに発展した判別分析と異なり，医学分野でよく用いられている。Pは例えば正常と比較する疾病の確率で，オッヅ比の対数が線形式になるというモデルである。Fisherの後継者の一人であるCox（1958）が開発したCox回帰の範疇と考えればよいであろう。この手法は，米国のフラミンガム心臓研究（Framingham Heart Study）で開発された手法である。東大医学部の開原先生の研究室で開催された勉強会で，「Walker & Duncan」のFortranのプログラム付きの循環器疾患のコホート分析の論文の輪読会で知った。しかし実際の開発者の引用できる文献はないようだ。右のように変換するとPはマーケッティングなどで知られた成長曲線になる。

$$\log(P/(1-P)) = f(x) \quad \text{あるいは} \quad P = 1/(1+\mathrm{Exp}(-f(x))) \tag{2}$$

Fisherの提案した判別分析は，2群は平均が異なり，各群に属する症例は平均を中心にばらついているという点[3]である。平均から離れるにしたがって出現頻度は少なくなる。そして2群の頻度が同じ判別境界で2群を判別する。このため，1）判別境界の近傍には症例数が少ない，2）正常と疾病の典型例は各群の平均であると考えている，ことが現実に適合していない。以上の点が，全く医学診断に適していないことに気づいて，1973年にベイズの定理で実データの度数を調べ，計測値が連続的に大きく（あるいは小さく）なるにつれ事後確率を0から1になるような試みを行った。そして，1948年にOR誌の編集委員として医療特集号の担当になり，その際に執筆し「地球モデル」として紹介した（新村，1984）。すなわち，正常を地

[3] ガウスが2地点間の距離の測定を繰り返した測定値の分布であるということは理解しておくべきである。

球と考え，各疾病群は山と考える。山頂が疾病の典型例であり，正常からの乖離の程度で表される。重要な点は，地平線が判別超平面で，この近傍に疾病の症例が多いことである。このようなモデルはTaguchi & Jugular（2002）が品質管理で正常状態を基底空間と考え，異常状態を正常の分散共分散行列で計算したマハラノビスの汎距離が大きいほど異常と定義したことと同じである。しかしスペクトル診断は，プログラムの作成能力に劣っていて大変な手作業であり，発展させることはできなかった。後でロジスティック回帰を知ることで，外国人研究者の汎用的にまとめるスマートなアプローチに脱帽した。

ここで重要なのは，「地球モデル」が適しているのは「医学診断」だけでなく「株や債券あるいは不動産などの格付け」や「試験の合否判定（新村，2011a）」などが「Fisherの仮説」に基づく判別よりも現実データに適している点である。ししてLSD判別が正しく行えないと，「癌の遺伝子解析」に役立たないという事実である。ビッグデータ解析の重要な研究テーマで，まったく成果がでなかった理由である。

3. 判別分析の5つの問題

判別分析の5つの問題は何度か取り上げているが，同じテーマで異なった知見を加えて発展させている。本研究では特に癌の遺伝子解析と問題1に関して大きな進展があった。

3.1 判別分析の問題1 ―誤分類確率の多くの欠陥―

⑴ 判別規則と判別超平面上のケースの扱い

判別分析は重回帰分析と異なり，次のように単純である。

判別規則：LDFを$f(x)$とし1群の外的基準を$y_i=-1$, 2群を$y_i=1$とする。判別規則は$y_i * f(x)>0$であれば両群に正しく判別され，$y_i * f(x)<0$であれば両群のいづれかに誤判別される。判別超平面上にくる患者（$f(x)=0$）は2群のいずれに判別するかは判定できない未解決の問題1であるが，統計研究では理由なく1群に正しく判別されたとする研究者が多い。

これは大学卒業後に取り組んだ心電図の診断論理を開発中に見つけたのでこだわりがあった。なぜ多くの研究者は論文に$f(x)\geqq 0$であれば1群に，$f(x)<0$であれば2群と表記するか疑問であった。きっと未解決の問題で対応できないので仕方なしにこのように対応しているとも考えた。しかし，医学論文の中には$f(x)>0$であれば1群に，$f(x)<0$であれば2群と表記し$f(x)=0$に触れないものもある。すぐに確認すればよいのに，2010年に日科議連出版から『最適線形判別関数（新村，2010）』の本を出した後で，2～3人の主要な統計研究者に聞いたが，

1）統計は5%間違うことを前提にしているので問題にするのがおかしい，

2) これからのBigデータの時代に判別境界上のデータで多少間違えても問題でない,

3) 確かにカテゴリカル・データの場合は判別超平面上に多くのケースが来る可能性が高い,

4) 分からない,

5) この問題は未解決であるので,統計は確率の学問なのでサイコロで帰属を決める,

などである。1) と2) は論外である。3) と4) は救いがある。5) はこの問題を決定不能と理解しているが医学診断で医師は判別超平面上の近辺の症例の診断に心血を注いでいるので,無理に現実を無視した間違った解釈をする必要がない。そこで「医者は医学診断で博打をやっているのでないと反論」すると,「それもそうだね」と納得された。彼らはこの点を判別分析の問題1として扱った筆者の論文を見る環境にあったが,誰にも真剣に読まれていないことが分かった。一方,ORに提出した論文で,査読者の一人から「問題を作っているし,連続空間の1点を判別関数が選ぶ確率は0である」という驚く理由で却下された。研究は「新しい問題を見つけてそれを解決する繰り返しであり,統計分析は数学で考える純粋な連続空間で考えているわけではない」ので,今だこの棄却理由に納得していない。

⑵ 判別係数と誤分類数の関係

しかし,これほど重要な問題が放置されたままであったのは驚きである。これを解決できたのは1997年から整数計画法 (IP) でMNM基準による最適線形判別関数 (IP-OLDF) を研究した際,それを説明する図1で初めて解決できた。データがクラス1に1例,クラス2に2例の3例で2個の変数 (X_{1i}, X_{2i}) の値を持つ次のデータの判別を考える。

Class1: ケース1:$(X_{1i}, X_{2i})_{i=1} = (-2, -3)$

Class2: ケース2:$(X_{1i}, X_{2i})_{i=2} = (-2, 1)$, ケース3:$(X_{1i}, X_{2i})_{i=3} = (1, -3)$

このデータで,IP-OLDFは⑶のように定式化される。

$$\text{MIN} = \Sigma e_i; \qquad (3)$$

$$\text{H1:} \quad y_1 * (-2b1 - 3b2 + 1) >= -M * e_1;$$

$$\text{H2:} \quad y_2 * (-2b1 + 1b2 + 1) >= -M * e_2;$$

$$\text{H3:} \quad y_3 * (\ \ 1b1 - 3b2 + 1) >= -M * e_3;$$

今 f = b1*X1 + b2*X2 + c というLDFを考える。説明変数の値を代入した判別スコアが f > 0 であればclass1, f < 0 であればclass2に判別されると当初は考えていた。しかし,不等号の向きは解析者が指定できないことと,異なった不等号の向きを制約式で使い分けるのはわずらわしい。そこでclass1であれば $y_i = 1$, class2であれば $y_i = -1$ という識別子を導入する。こ

れは重回帰分析では目的変数の値として利用すれば，重回帰分析でF-LDFになることは知られている。これによって拡張された判別スコアが $y_i * f > 0$ であれば正しく2群の何れかに判別され，$y_i * f < 0$ であれば2群に誤判別されると統一できる。しかし数値計算上の配慮から拡大された判別スコア $y_i * f >= 0$ であれば正しく2群の何れかに判別され，$y_i * f < 0$ であれば2群に誤判別されると考え，$y_i * f > 0$ と $y_i * f = 0$ と $y_i * f < 0$ の例数をカウントし出力してチェックする。ここで上で与えられた3件の計測値を変数に代入し，MNM基準で逆に判別係数（b1，b2）を求めるわけである。判別関数の定数項cは任意の実数であるが1に固定する。ここで問題になるのは，正しく判別される場合は $y_i * f >= 0$ でよいが，誤判別されるケースがどれになるかは事前にわからないので $y_i * f < 0$ と指定できない。そこで0/1の2値を取る整数変数 e_i を用い，正しく判別されれば $e_i=0$，誤判別されれば $e_i=1$ とする。判別境界からどれだけの距離で誤判別されるかわからないので $M=10000$ のような大きなBigM定数を用いて $y_i * f >= -M * e_i$ とする。これで正しく判別されるケースでは $y_i * f >= 0$，誤判別されるケースでは $y_i * f >= -10000$ になる。要は誤判別されるケースの判別スコアが $0 > y_i * f >= -10000$ の範囲になることを期待している。すなわち2値整数変数は，あれを選ぶかこれを選ぶかの選択モデルに利用される。今回の場合，判別超平面を $f = 0$ にするか，代替案として $f = -10000$ を選ぶかの選択問題に置き換えたことになる。もし -10000 より判別スコアが小さくなれば不等号は成立しないのでエラーになる。MはIPで収束計算を確実に行うために経験的に用いられている定数であり，一般に10000程度が良いとされている。これを10や100のように小さくとれば，誤判別されるケースを正しくとらえることができない（新村，2010）。一方，100,000のように大きくとれば，数理計画法では係数で他の係数を割る演算が多いので，絶対値が例えば 10^{-8} 以下であればディジタル計算では0と判定すると，それ以降の計算が0になって影響を及ぼす。例えば旧東京三菱銀行でSAS/IMLを用いて投資分析システムを開発したが，分析結果がおかしいので企業人の時代にコンサルタント依頼を受けた。株や債券の係数の最大値と最小値の比が 10^8 でおかしくなっていた。そして，IP-OLDFの目的関数“$MIN=\Sigma e_i$;”で e_i が1になる個数の和を最小化している。これでMNM基準のLDFが定義できた。

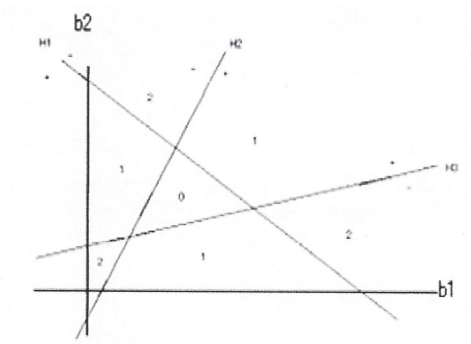

図1　判別係数の空間で，誤分類数とLDFの関係を説明

これを2次元の判別係数の（b1，b2）平面にプロットしたのが図1である。H_i はケース x_i で得られる線形超平面 $y_i * f(x_i)=0$ であり，$y_i * f(x_i)>0$ であれば＋半平面，$y_i * f(x_i)<0$ であれば－半平面と呼ぶことにする。＋半平面に含まれる点を判別係数に選べばケース x_i は正しく判別され，－半平面に含まれる点を判別係数に選べば，ケース x_i は誤判別され，$y_i * f(x_i)=0$ であれば判定不能になる。3つの線形超平面から7個の凸体が作られる。各凸体の内点に対応する判別係数はそれを囲む線形超平面の－半平面の数を数えればNMになる。そして，内点の無限にある判別係数は－半平面に対応した同じケース x_i を誤分類する。しかし判別超平面で作られる頂点や辺上にケースがあり，x_i をいずれに判別するかは決定できない。また凸体は有限個であるので，NMが最小の凸体すなわち図の三角形を最適凸体と呼ぶことにする。このNMがMNM=0になる。IP-OLDFはデータが一般位置にあればこの最適凸体の3つの頂点の一つを選ぶ。p個の説明変数がある場合，頂点が丁度p個の線形超平面で作られる場合，データは一般位置にあり，頂点に接する 2^p 個の凸体に必ず最適凸体がある。最適凸体は必ず＋半平面で囲まれているが，p個の0/1の 2^p 個の組み合わせには必ずすべて＋のものが1個あることで分かる。しかし頂点が（P+1）個の線形超平面で作られる場合はデータが一般位置にないといい，$2^{(p+1)}$ 個の凸体に必ずしも最適体凸体があるか否かは一般的に言えない。以上のことを前提とすれば，通常の判別関数はどの判別係数を求めるか分からないので，誤分類数は得られたNMに線形超平面のケースh個をカウントし，（NM+h）まで増える可能性がある。またこの図から容易に分かるが，隣り合った凸体のNMは1だけ異なる関係がある。

⑶　統計的判別関数がSMを見つける困難な2点

2017年11月17日にJMPが世界4都市で開催しているDiscovey Summitで50分の発表枠をもらったので，退官記念講演に変えて「横長データの代表であるMicroarrayデータによる癌の遺伝子診断」の発表を行った（新村，2017a）。そこで図1を使って，「癌の遺伝子解析」を行うための2つのハードルを示した。スイス銀行紙幣データでも改定IP-OLDFは自然に特徴選択

が行えることを確認した。それ以外の判別関数は，1）まず6変数の判別係数の（X1-X3, X5）の4個の判別係数を 0 にする必要がある。そして図1のように（X4, X6）の2変数の判別係数の部分空間を特定する必要がある。その上で，最適凸体の内点を求める必要がある。これで他の判別関数が癌の遺伝子を特定できないことが分かる。翌日新潟大学の「多様な分野における統計科学の総合的研究」と12月3日の筑波大学での「大規模複雑データの理論と方法論，及び，関連分野の応用」シンポジュームで発表したが（新村2017b; 2017c），それ以上に有効な情報を得た（石井，2017）。筆者が分析した6種類のMicroarrayデータを含む10種類以上で，1）高次元空間の分布を調べるソフトで2群が完全に分離していること，2）高次元のPCAで第1固有値が第2固有値以降に比べスパイク状に大きい，の2点である。筆者が用いているJMP（Sall, 2004）では，高次元のF-LDFは判別できるが主成分分析や分布を検証できない。しかし，筆者の「癌の遺伝子解析」の驚く結果が，他の手法でも確認できたことの意義は大きい。筆者の知る限り，「癌の遺伝子解析」で2群がLSDであるとはっきり結論を出しているのは，日本のこの2研究だけである。しかし石井の指導教官である青嶋と矢田らは，2群が異なった球面に張り付いていると結論しているが，もし満遍なく布置していればFisherの仮説を満たすはずであり，F-LDFのNMが0にならない事実と反している。筆者の研究はさらに，彼らが見つけた信号空間が分割され小さな遺伝子空間に分割されることを見つけた。

⑷ 定数項を1にした意義とIP-OLDFの定式化

定数項をC=1に固定したことで，p個の変数の判別係数の空間で判別係数とNMの関係を図1のように世界で初めて示した。パターン認識の研究では定数項Cを固定しないで，定数項を含む（p + 1）次元空間で考えている。この場合3個の線形超平面は全て原点を通り，判別係数とNMの画期的な関係は分からない。

しかし，最初はこれで良いと考えたが求まったMNMが他のモデルに比べて極端に悪いものがあることが分かった。少し考えて，C=－1 と C=0 にした3通りのモデルを解く必要があることが分かった。またC=2に固定した場合，これはC=1 の場合と相似な関係にあり同じNMと判別係数の関係になる。

以上から，式⑷でIP-OLDFを定義する。bは，定数項を1に固定しているので，p次元の判別係数の係数空間の任意の値である。y_i は癌を1に，正常を－1とする目的変数である。Mは10000に設定したBigM定数と呼ぶ一定値である。e_i は 0/1 の2値整数変数である。目的関数はこの和すなわちMNMを求める。即ち3件が正しく判別されればMNM=0になる。そして2次元の判別係数の空間を（${}^t x_i b+1$）=0 という線形超平面で分割すると空間は有限個の凸体に分割される。任意の凸体の内点が，3個の超平面の－半平面（${}^t x_i b+1<0$）にくる個数が図に示してある。ある凸体の任意の無限個のLDFは，このNM個の同じケースを誤判別するので判別分析では等値と考えられる。有限個のケースから有限個の凸体しか作られないので，必

ずNMが最小のMNMになる凸体があり，これを最適凸体と呼ぶことにした。

$$MIN=\Sigma e_i;\qquad y_i\times({}^t x_i b+1)>=-M\times e_i; \tag{4}$$

(5)　改定IP-OLDF，改定LP-OLDFと改定IPLP-OLDFの定義

IP-OLDFは最適凸体の内点を直接求めることができなかった。そこで，悪戦苦闘し式(5)の改定IP-OLDFを定義した。定数項をb_0とし，右辺に1を挿入した。ケースが正しく判別されれば$e_i=0$とし，拡張された判別スコアが$f=y_i\times({}^t x_i b+b_0)>=1$になり，判別スコアが1以上になる。誤判別されるケースは，$e_i=1$とし，拡張された判別スコア$f=y_i\times({}^t x_i b+b_0)>=1-10000=-9999$になる。すなわち正しく判別されるケースはSV=1で判別し，誤判別されるケースはSV=−9999というSVの代替案を選ぶ。MはBigM定数で10000に設定したので，誤判別されるケースの拡張された判別スコアはSV=−9999に引っ張られて判別スコアは大きな負の値になる。これによって誤判別されたケースの判別スコアは［−1，1］の範囲に入らない。すなわちf=0にないので，図1で説明した判別超平面上に来ることを避けることができる。M=1，10，100，1000，10000で検証すると100以下で［−1，1］の範囲に入る事例も出てくることを確認している（新村，2010）。筆者は，BigM定数のようにデータに無関係に適した値を設定できない手法，すなわちソフトマージン最大化SVM（Vapnik,1995）のペナルティ cや，2つのパラメータの設定をユーザーに求めるRDAは使うべきでないと考えている。データごとに個別に適した値を求めたとしても，他のデータに適用できないからである。この点BigM定数はIPで広く検証された値であり問題がない。以上の理由と多くの実証研究で，多くの欠陥のあるNMに代わってMNMが判別関数の重要な統計量であると断言できる。

$$MIN=\Sigma e_i;\qquad y_i\times({}^t x_i b+b_0)>=1-M\times e_i; \tag{5}$$

b_0: free decision variable.

e_i: 0/1 integer variable

M: BigM定数＝10000

式(6)は改定LP-OLDFである。2値の整数変数から非負の実数に変更しただけであるが，これによって計算速度の速いLPで解くことができる。正しく判別されるケースはSV＝1より大きく制約され，誤判別されるケースはSV=1から小さくなる距離を表す。目的関数を“$MIN=\Sigma M\times e_i;$”とすればその距離の和の最小化基準になるが，実際には“$MIN=\Sigma e_i;$”のまま分析している。この判別関数は重なりの有るデータでは，判別超平面上にケースが来ることが多く使用しない方が良いことが分かっている（問題1）。しかしIP-OLDFは，LSDを理論的

に判別する保証はないが，ロジスティック回帰と同じく判別結果は意外と良いことが分かった。

$$MIN=\Sigma e_i;\qquad y_i\times({}^t x_i b+b_0)>=1-M\times e_i; \tag{6}$$
$$e_i:\text{非負の実数.}$$

改定IPLP-OLDFは，第1ステップで改定LP-OLDFで判別する。そして$e_i=0$で正しく判別されたケースのe_iを0に固定する。第2ステップでは，固定されていないケースだけに改定IP-OLDFを適用する。将来計算時間のかかる改定IP-OLDFに代わって，改定IPLP-OLDFを用いることを考えて開発した。しかしIPソルバーの計算速度の改善で，2010年以降普通のデータでは改定IP-OLDFより計算を2段階で行うので高速ではなくなった。

(6) 誤分類数NMの欠点

改定IP-OLDF以外の判別関数は，判別結果の評価に用いる誤分類数に多くの欠陥がある。

1) 研究で比較のために用いているF-LDF, QDF, ロジスティック回帰，改定LP-OLDF, 改定IPLP-OLDF, H-SVM, S-SVMでペナルティcを10000と1としたSVM4とSVM1でNMが異なる。

2) 判別超平面を動かすとNMが異なってくる。このため，幾つかの水準でNMを求めてROC曲線で比較することを提案した。ROC曲線を医学診断に導入したのはLustedであるが，彼の本を野村先生と中村博士が翻訳を行った。また，筆者が初めて国際会議に参加したトロントで開催されたMedinfo77でLustedが座長のセッションに割り振られ，会議前に打ち合わせを行った。急に彼から何分発表したいと聞かれ，15分といったつもりが50分といったようで，隣にいた開原先生から小声で間違いを指摘された。JMP（新村，2004）のロジスティック回帰はROC曲線で判別結果を表しているので，NMが最小の結果を用いている。そしてNM=0かつMNM=0である場合，ロジスティック回帰はLSDを正しく判別したと解釈している。

3) 2群のケース数に比例した事前確率で判別境界を変更した場合と，事前確率を1:1にした場合では結果が異なる。元々2群が正規分布であると仮定しているので後者の方が正当な対応である。しかし，判別分析の初期のユーザーは医学の比重が高いので，症例数を反映したことが要求された。さらに正常症例より，異常症例を間違って診断することをできるだけ修正するため，リスクで判別境界を動かすこともある。これらはNMが判別境界の変更で容易に変わるので，できるだけ良い結果を得たいという希望を反映したものと考えられる。また，これらを考慮していないSVMの結果と比較するためにも，ケース数に比例した判別結果を用いた方が良い。

これに対してMNMはデータに対して一意に決まり，全てのNMの下限値である。それに加えて，「癌の遺伝子解析」という重要な問題や，LSD判別に適している点である。

3.2 線形部分離可能なデータ（LSD）の判別

(1) ハードマージン最大化SVM

VapnikはLSDの判別を式(7)のハードマージン最大化基準で明確に示した（H-SVM）。すなわち2つのSVでクラス1とクラス2を完全に判別し，SV間にケースが来ないように空間を3つの領域に分割する。その上で，SVの距離を最大化すれば汎化能力が高まると主張した。SV間の距離最大化基準にすると非線形最適化になり極大/極小値から真の最大値/最小値を求める困難な問題が発生する。そこで式(7)のように逆数を取ればQPにできる。そして拡張された判別スコアが“$y_i \times ({}^t x_i b + b_0) > = 1$”を満たす制約式を解くことになる。ただしQPは遺伝子空間全体の定義域で，唯一の極小値かつ最小値を求めNM=0であってもLSDであることが分かる。しかし，部分空間のNM＝0である最小値を求めることができないので，全ての遺伝子の組み合わせ判別モデルを探索しなければ部分空間のSMを見つけることはできない。

$$MIN = \|b\|^2/2; \qquad y_i \times ({}^t x_i b + b_0) > = 1 \tag{7}$$

筆者がスイス銀行紙幣データ（Flury & Rieduyle,1988）の真札と偽札各100枚の6変数の63個の全てのモデルを改定IP-OLDFで判別し，（X4, X6）の2変数モデルでMNM＝0であることを発見した。そしてMNMの単調減少性（$MNM_k \geqq MNM_{(k+1)}$）を見つけた。すなわち，k個の変数のモデルでMNM_kが得られた場合，それに残りの変数から1個選んで追加した（k+1）変数のモデルの$MNM_{(K+1)}$は必ず単調減少するという事実である。この研究を行って，多くの場合は数式で示さず言葉で説明できる。すなわちk次元の部分空間は（k+1）次元の部分空間に含まれるので，$MNM_{(k+1)}$は必ずMNM_kより等しいか小さくなる。「数式で厳密に定義しないのは論文でなくエッセイである」というコメントで棄却されたこともある。言葉で間違いなく説明できるのになぜ数式展開しなければいけないのか今もって分からない。そして重要なことに，$MNM_k = 0$であればこのk変数を含むすべてのモデルはMNMが0になる。スイス銀行紙幣データでは，（X4, X6）の2変数を含む16個のモデルのMNMが0になり，残り47個が1以上になる。すなわち6変数のモデルを大きなMatryoshakaと考える。この中に5変数から2変数までの15個の小さなMatryoshkaが含まれる。すなわち，LSDはMatryoshka構造という特殊な構造を持っている。Golubら（1999）が論文で30年以上この研究を行っていると述べているので，少なくとも筆者が大学3年の1970年からMicroarrayデータから癌と正常を分ける研究が医学研究者に加えて統計研究者にも取り上げられてきたと考

えられる。医学研究グループはそれなりに症例との検討を行っているが，統計の格好の新しいテーマとして行われてきた多くの統計研究は一つも芳しい成果を出していない。少なくとも筆者が分析した6つの研究は，研究に用いたMicroarrayデータを広く公開しているが，どの研究もMNM=0と指摘したものがない。その上何を基準に癌の遺伝子を特定しようとしているかの基準も明確ではない。筆者の基準は2クラスがMNM=0であるので，これを癌遺伝子を特定する遺伝子の組と考えている。その上で，MicroarrayデータはMNM=0になる小さなMatryoshka（Small Matryoshka, SM）に分割できる。しかもSMに含まれる遺伝子数は少ないので，比較的妥当な価格で血液から癌の遺伝子診断が正確に行える。世界で初めてLSDの判別研究を行っていても，種々の無理解があった。日本と海外の学術誌や国際会議で，「判別分析はLSDのような簡単な判別が重要でなく，重なりのあるデータの判別が目的である」という指摘である。確かに重なりの有るデータを判別する需要が多い。しかし理論的に正しくLSDを判別しMNM＝0であるとできるのはH-SVMと改定IP-OLDFだけである。H-SVMはNM＝0を出力するが，それはMNM＝0と等価である。しかし，H-SVMは重なりのあるデータに適用するとエラーになるので，LSDの判別の定義だけで実際にLSDの研究されなかったようだ。統計研究者の一部で，LSD判別が重要でなく重なりの有るデータの判別が重要であるという指摘は論理的でない。MNM＝0であればLSD,MNM＞＝1で重なりの有るデータと初めて定義できる。彼らはMNMという統計量を知らないので彼らの主張が論理的でないことは明らかである。その上で，LSDというはっきりした結果で判別結果を正確に評価できる重要性を全く理解していないことを示す。

⑵ ソフトマージン最大化SVM（S-SVM）

H-SVMに続いて幾つかのケースがSVで判別できないケースを許すソフトマージン最大化SVM（S-SVM）が式⑻で定式化された。制約式のSVで判別スコアが1以上で判別できないケースがあるの場合，非負の実数e_iでSVを（$1-e_i$）に変更して，目的関数でこの距離の和を最小化する第2項を付け加える。2目的最適化を解くアルゴリズムはないので，ハードマージン最大化の逆数の第1項と荷重和で単目的化するためにペナルティcという重みを導入した。問題は2つあり，最適なcについての研究がないことである。そこで種々の研究である程度値が大きいことが良い場合が多いことが分かった。しかし断定できないのでc=10000をSVM4し，C＝1をSVM1として両方を比較に用いることにした。次の問題は，普通2目的最適化はMarkovitzeのポートフォーリオ分析のように（新村，2007;2011b），2次式で表されるリスクを最小化し，利益を最大化する第2項を制約式である利益以上とし，その水準を変えて効率フロンティアを描く方法が重みづけより一般的である。Vapnikはそれを知っていたと考えられるが重みで単目的化する方法を選んだ。この後Kernel SVMという多くの研究者を魅力した非線形判別分析と呼ばれる方法を提案した。多くの研究者は，こちらの方に注目して

LSDの判別に注目しなかったようだ。

$$MIN = \|b\|^2/2 + c \times \Sigma e_i; \qquad y_i \times ({}^t x_i b + b_0) >= 1 - e_i \tag{8}$$

c: penalty c for combining two objectives. e_i: non − negative value.

⑶ SVM研究の判別分析に占める意義

一般的には判別分析の主流は，1）正規分布を仮定し，2）相関比最大化基準をよりどころとし，3）それを分散共分散行列というコンピュータ処理が容易である情報を利用する，という3点セットでF-LDFとQDFの後，分散共分散行列の技術，RDAやLASSOが開発された。しかし筆者は，Fisherを第1世代とし，異なった判別像のCox回帰やロジスティック回帰を第2世代，そして数理計画法（Mathematical Programming, MP）のQPで特定の理論分布を仮定しないVapnikによるSVMの研究を第3世代と考えるのが適切であると考えている。ロジスティック回帰は，対数尤度がLDFで表されるので正規分布を前提としていると紹介するインターネットの解説頁もあるが間違いである。Fisherの提案した最尤推定法で与えられた判別データから収束計算を行う。このため計算に用いているヘシアン行列からロジスティック回帰の標準誤差を求めている。これは，正規分布から標準誤差を求める伝統的な推測統計学と区別すべきであろう。またFirth（1993）は，LSD判別を行うと収束計算が不安定になり，標準誤差は異常に大きくなると指摘している。一般的にこのような判別モデルは考慮すべきではない。しかし，改定IP-OLDFでMNM＝0であることを確認し，ROC上でロジスティック回帰のNMが0になる場合，筆者はロジスティック回帰がLSDを正しく判別できると拡大解釈している。そして，改定IP-OLDFでMNM＝0であることを確認した全てのSMに適用し，ロジスティック回帰でもそれらがLSDであることを確認検証に用いている。

MPによる判別分析は，数理計画法でも研究されStam（1997）が伝統ある米国のOR学会誌に総括論文を発表し第1次の研究は終焉したと考えている。新村（2011b）はL.Schrage（1992）のテキストで，多くの重回帰モデルがMPで定式化できることを知って感銘を受けた。判別分析の紹介は無かったが統計と数理計画法を融合した判別分析の研究を1997年に始めた。数理計画法の世界では，Stamの報告でこのテーマは終焉したことを知らなかったが，かなり後になってSchrage教授から関連文献が送られてきて初めて知った。しかし，Vapnikは1995年にSVMの解説書を出版している。彼は発表の場を統計やORを避けて，パターン認識などの工学分野で普及に努めたのは賢明である。これらの気難しい分野で発表していれば，筆者以上に多くの障壁に遭遇したであろう。統計やORの研究は，SVM研究者を取り込むことに失敗したわけである。結局判別は，F-LDFを第1世代と，そしてロジスティック回帰やCox回帰を第2世代とし，SVMで第3世代に入ったと認識すべきである。筆者の研究もSVMに連な

っている。

3.3 一般化逆行列の瑕疵（問題3）

重回帰分析や判別分析などの多変量解析や主成分分析では，対象とする現象のばらつきをとらえる分散共分散行列が重要である。変数が一定値を取る場合，逆行列が求まらない問題がある。技術力のない統計ソフトの会社は，それらの変数を分析前に省いて処理すればよい。しかしSAS社は一般化逆行列の研究で，それを解消する技術を確立した。特にSAS社の社長のGoodnight氏の代表的な研究業績は，1）一般化逆行列，2）分散共分散行列を基本に重回帰分析で全ての回帰モデルを見つける研究（特許取得済み），が重要でありSASの技術遺産になっている。一応, 2010年までに問題1, 問題2と問題3を解決し, 2010年に日科技連から『最適線形判別関数』を上梓した。そして次に何を応用研究のテーマにしようかと考えた。その時，「癌の遺伝子解析」もかすかに脳裏を横切ったが，スイス銀行紙幣データでLSD判別に成功していた。それ以外のLSDデータを探すことは偶然のめぐり逢いであると考えていたが，「試験の合否判定」がLSDデータであり容易に手に入ることが閃いた。大学入試センター試験であれば，LSD判別の結果が試験の難易度や大門ごとの難易での年次比較ができるのではないかと考えてアプローチした。そして大学入試センターから大学生で実施した3年間のセンター試験の研究用データを借り受けることに成功した。ある程度の分析結果をセンターの研究員に報告し，成蹊大学で応用統計学会と入試センター共催のシンポジュームでも発表させてもらった。そこでの結果は驚くもので，数学の大門4問で10%，50%，90%の3水準を合否判定すると，すべての判別でF-LDFのNM＝0になるものはなかったうえに，数学で3割近くの誤分類確率になる例を確認した。これは，「地球モデル」で述べたように合否の2群が正規分布でないのに正規分布を仮定して求めた判別超平面がその近傍に多くの合格学生がいて誤分類確率が高くなるためである。さらに90%以上を合格とし未満を不合格とすると，QDFとRDAの組み合わせで合格群が全て不合格群に誤判別された。QDFは「データにおかしなものがあり無条件でRDAに切り替える」というメッセージを出す。この解決に3年以上かかり，東大で開催された日本計算機統計学会で4年ゼミの黒岩さんに「東京都27市の公立図書館の経営効率性」の発表と，私は人生初の「試験の合否判定」の終了報告を行ない，「誰か，この問題を解決してほしい」と述べたが無関心であった。数日後の深夜に「多変量的な検討ばかりで，各得点分布の1変数の分析を省いていた」という初歩的なミスに気づいた。分布を調べると，特定の設問で合格群の成績の良い学生全員が正答し，不合格群がバラついてることが簡単に分かった。重回帰分析やF-LDFでは分散共分散行列は2群に関係なくプールしたものが用いられるので影響を受けない。しかしQDFの場合，2群で別々の分散共分散行列を用いる。2群で同じ変数が一定値の場合は検討しているが，一方だけが一定値の場合を検討

していなかったようだ。悪いことに，RDAという筆者の熟知していない手法に切り替わることが原因の特定を困難にした。筆者の報告で，半年以上JMPの担当者が色々試みを行っていることが判別結果がゴロゴロと変わることで確認できた。このことを知らないユーザーが利用していたら迷惑な話である。半年ほどして，RDAは2つのパラメータを［0，1］の範囲でチューニングしてほしいといわれた。S-SVMのPenaruty cと同様であるが，多くのデータで一般に利用できる値が固定できなければ，このような手法は使うべきではない。RDAも当初は，最適な値を検証しユーザーが指定できないようにしていたがそれが破綻したわけである。もう5年以上になるが，QDFでは解決策が示されていないが，それに代わってデータの不備に関する情報が出力されるが対応方法が分からない。

3.4 判別分析は推測統計学でない（問題4）

判別分析は推測統計学でないと指摘すると統計研究者の中には怪訝な顔をする人がいる。統計の利用者の場合，「判別係数や誤分類確率に標準誤差が出ていないでしょう」というとすぐに理解してもらえる。このため判別分析のモデル選択は工夫がいる。一つは，2群を識別する y_i を目的変数として重回帰分析を行いモデル選択をすることである。しかし癌の遺伝子解析の米国の多くの論文では，「一つ取って置（LOO）法」が用いられている。n件のデータから1個を取り去り（n－1）個を教師データとして判別モデルを求め，1個の検証データで評価することをn回繰り返す。しかし，筆者は検証標本は一定であり，教師データより件数が多くあるべきと考える。

そこで分析に用いるデータが小標本の場合，それを100回コピーして検証標本とする。乱数で大小順に並べ替えて，上から順に100分割し学習標本として，これら100個の学習標本で100倍に膨れ上がった疑似標本を検証標本とする「小標本のための100重交差検証法（新手法1）」を考えた。当初は馬鹿正直に学習標本を乱数でサンプリングしていたが，手間暇がかかって問題であった。また後での再利用や検証を考えると新手法1の方が優れている。小西・本田ら（1992）は，Bootstrap法で標準誤差を求めることを提案しているが，コンピューターインテンシブな方法をとるのであれば，直接研究対象のデータで役に立つ方が便利である。

この方法を用いて，Springerでは6種類のデータで，全ての組み合わせモデルで検証標本の平均誤分類確率が一番小さいモデルをBestモデルと命名した。比較する8種類のLDFでこれらのBestモデルの平均誤分類確率でもって簡単に8種類のLDFの評価が行える。Fisherのアイリスデータではそれほどの違いはないが，他のデータでは圧倒的に改定IP-OLDFのBestモデルがよく，ロジスティック回帰や，改定LP-OLDFやSVM4が次に良く，多くの場合F-LDFやSVM1の成績が悪かった。MNM基準に関しては，長らく編集委員も務めた「行動計量学会誌」で「学習標本を過推定するMNMは愚かな判別基準で統計のイロハも知らない。正規分

布を仮定するF-LDFが一番検証結果が良いに決まっている」というレフリーコメントと共に数回の改定の後で論文がリジェクトされた。確かに「Fisherの仮説を満たすアイリスデータでF-LDFはMNMに収斂しそれほど違いはない。しかし他のデータでは検証結果は非常に悪い」ことが35年かかって実証できた。

3.5 Bigデータ分析として注目され失敗した癌の遺伝子解析

2015年10月25日㈰に富山県民会館で開催された統計シンポジュームで手法1で求めた判別係数がF − LDF以外が自明な判別係数になったので，判別分析の新理論が完成したと考えて終了報告を行った（新村，2015）。翌日の午後1番の石井（2015）の発表で，米国の6研究グループが研究論文に使用したMicroayyayデータを公開していることを知った。2000年以前にインターネットで調べた際は，データの形式が面倒でありあきらめたが，Excelに容易に展開できる。しかも，6個のデータが不思議なことにアイルランドの医学部のHPからダウンロードできる。28日にJeffryら（2006）のHPからExcelデータをダウンロードし，77症例7129遺伝子をもつShipp他（2002）のデータを改定IP-OLDFで判別するとMNMが0である上に，僅か32個の遺伝子の係数だけが0でないことが分かった。すなわち，高次元の遺伝子空間がLSDであり，その32次元の部分空間もまたLSDであるという，遺伝子空間が特殊なMatryoshka構造を持つことを知った。筆者の知る限りでは，LSDの判別の研究を行っているのは筆者だけである。それに対して，日本と英語の論文誌のレフリーから，判別分析の目的はLSDのような簡単な判別でなく，Overlapデータの判別が重要であると指摘された。この指摘は間違いで，LSD判別は奥が深く，また結果が明らかで検証結果の評価が明らかになる。例えば，理論的にLSDを正しく判別できるのは，H-SVMと改定IP-OLDFだけである。「改定LP-OLDFも多くの場合にLSDでNM = 0になり，多くの判別係数を自然に0にする点が，S-SVMと異なる」。ロジスティック回帰は，LSDのデータを判別するとFirthが指摘するように，最尤推定の収束計算は不安定になり得られた判別係数の標準誤差は大きくなる。本来であれば，このような結果は採択しない。しかし，筆者は判別スコアを表すROC曲線上で判別境界を動かしてNMが最小のものを選んでNM=0になり，かつ改定IP-OLDFでMNM = 0であることを確認できた場合，ロジスティック回帰はLSDを認識したと判定している。SVM4は，ほぼLSDを正しく判別できる。即ち，F-LDF，QDF，RDAそしてLASSOなどは，この比較から「LSDを正しく判別できない重大な問題がある」ことが明らかになる。それができない判別関数が，さらに部分空間のMNM = 0になるSMを見つけることができないのは自明である。

3.6 Springerの概略

筆者の不確かな記憶では, 2000年以前に統計の国際会議で“Small n Large p”のデータから, 分散共分散行列を推定し, 癌の遺伝子解析などに適用しようという試みがあった。今日これらの研究を検索しても探し出すことは難しい。最近ではLASSOを含め, 多くの研究で種々のFeature Selection法が研究されているが, 筆者は何もしないで「自然にSMで癌遺伝子の選択」が行えたことになる。さらにこれらの遺伝子の組のSMを全体から省いて判別すると, また別のMNM = 0であるSMが見つかった。最終的にこれを繰り返すことで, 遺伝子空間はSMと呼ぶ排他的な和集合と, MNMが1以上の残りの部分空間に分離されることが分かった。すなわち, 高次元の遺伝子空間は複数個の癌遺伝子を特定できるSMの和集合である信号の部分空間と, 癌遺伝子を特定できない雑音である部分空間に自然に分かれた。これまでの研究では, Microarrayデータから特定したい「癌遺伝子」の定義が明らかでないことも問題である。そこで数理計画法ソフトのLINGO (Schrage, 2006; 2017) でMatryoshka Feature Selection Method (新手法2) を開発し, 6種類のMicroarrayデータの全てのSMを2015年12月20日までに見つけた。すなわち, この解決困難とされてきた問題5を僅か54日で簡単に解決できた。癌の遺伝子空間は高次元であり信号と雑音が混じっているためBig Dataの統計分析は困難といわれている。これを分離するための工学的な種々のフィルタリング手法が提案されている。新手法2をLINGOで実行すれば, このフィルタリングも自然に簡単に行える。

何故, 癌の遺伝子解析は非常に容易であるにもかかわらず, 1970年から良い結果がでず, 当事者以外の研究者が永遠に無理と考えるようになっていたのか？筆者は, それは単に統計的判別関数が全く役に立たなかったからと考える。この理由を, 1) Small N Large Pデータ, 2) NP-Hard, 3) 雑音を含む高次元データの困難さ, 等のバズワードを取り上げて説明したい。筆者は2010年に, 大学入試センター試験の大学生のアルバイトで実施した研究用データで得点を説明変数として, 合格水準を10%という緩い合否判定, 50%, そして90%という難関試験を想定し, 3年間の本試験と予備試験の合否判定を行った。その時, 数学で90%を合否判定に選んだ場合, JMPのQDFとRDAが合格群を全て不合格群に誤判別する問題3にであった。この解決に3年要したが, 理由は, 10%の合格者全員がある設問に正しく回答し, 90%の不合格群の学生の回答パターンがばらついている場合に, 一般化逆行列が正しく機能しないためである。この解決法は, 単に一定値をとる変数に乱数を加えるだけで解決できる。このことを論文に記述ところ,「JMPのバグを記述することは不適切」というレフリーコメントをもらった。バグは問題点が明確になればすぐに対応できる。しかし, 当初JMPの担当者はRDAに関していろいろな修正が行ったようだが, 数か月後に2つのパラメータを［0, 1］の範囲で自分で選んでほしいという回答が返ってきた。S-SVMでもそうであるが, ユーザーがチューニング・パラメータを選ぶような統計手法は好ましくない。一方, QDFに関してはいまだ

に解決されていないので，一般化逆行列に問題があると考えている。JMPがこのような場合を想定した製品検査を行わず，筆者が多変量的な検証にこだわって1変数の層別箱ひげ図で各項目の検討を行わなかったために，解決に2012年まで3年もかかった。即ち，癌の遺伝子解析も1970年以来解決できないのは，単にアプローチが適していなかっただけで，決して解決できない難問ではないことが分かる。以下がSpringer（Shinmura,2016）の概略である。

1章：判別分析の新理論の概説。

2章：IrisデータだけがFisherの仮説を満たし，F-LDFのNMがMNMに収束することと，相関係数の解釈の注意点等を紹介。Fisherは評価のために正規乱数を発生させて学習と検証標本を作製していないで実データを用いていることが重要だ。

3章：3個の共線性がある児頭骨盤不均衡（CPD）データで，誤分類数が増加法と減少法で著しく傾向が異なるため，共線性の解消法と変数選択法等を紹介した。さらにロジスティック回帰も問題1があることを示した。

4章：40人の学生の合否判定を5変数で判別し，判別超平面上に10人の学生が来るために各LDFの問題点と問題1を説明した。またデータを変換し，簡単にLSDの作成法を紹介し，LSD判別分析の重要な点を紹介（問題2）。

5章：18種類の合否判定を得点を説明変数として判別。例えば，大問2個の得点で50点以上を合格とする場合，$f = T1 + T2 - 50$ という自明なLDFで，$f \geqq 0$であれば合格，$f < 0$であれば不合格と正しく判別できる[4]。しかしF-LDFで判別するとNMから求まる誤判別率が20％を超える例も出てくる。正常（地球）と異常（山）を判別する医学診断は，計測値が連続的に大きく（小さく）なることで正常から異常に推移し，異常症例の典型例は異常群の平均でなく山の頂点である。そして，判別超平面である水平線の近傍に多くの症例がくる。このようなデータ構造を持つ医学診断，合否判定データ，各種格付はMNM=0であるのに，分散共分散に基づく判別関数の誤分類確率は異常に高いことを示した（問題2）。即ち過去に誤分類確率が20％を超えて研究を停止したものでも，MNM = 0である可能性は否定できないので重要な研究は再評価する必要がある。

6章：スイス銀行1000フラン紙幣の真札と偽札各100枚を6個の計測値で判別。（X4, X6）の2変数でMNM = 0になる。MNMは単調減少（$MNM_k \geqq MNM_{(k+1)}$）するので，（X4, X6）を含む16個のモデル（信号）はMNM=0に，残りの47個のモデル（雑音）はMNM＞=1になる。この事実は，6変数の全てのモデルを検討することで見つけた。しかし，新手法2を実現したLINGOのProgram3で判別するとMicroarrayデータと同じことが分かり，新手法2の内容を本データで紹介した。遺伝子診断では，6変数の空間に2変数から5変

[4] 等号を含むことができるのは，判別規則が説明変数で記述できるからである。

数のMNM=0になる15個の部分空間が含まれるのでMatryoshka構造と呼び，最小次元の（X4, X6）を癌の基本遺伝子（BGS）と呼んでいる。LSD判別では47個の雑音の解析は不要と考える。ただし，IPの分枝限定法は全てのモデルを探索することと同じことを行うため計算時間がかかる。そして，最初にMNM=0という最適解を見つけると計算を終了する。このためBGSを直接見つけるように制御できないので，Program3が見つけたものをSMと呼ぶことにした。

7章：小型車15車種と普通車29車種を排気量（X1）や座席数（X3）を含む6変数で判別した[5]。小型車のX1とX3は，普通車より小さく，この2変数は2個のBGSになる。Program3でSMの一つとしてX3が求まるが，1変数なのでBGSであることが分かる。小型車の座席数は4席で，普通車は5席以上のため，X3がモデルに入ると普通車の29車種が小型車に全て誤判別される一般化逆行列の問題3がある。この解決に3年かかったが，小型車の座席の一定値4に小さな乱数を加えるだけで解決できる。

8章：米国の6研究グループがMicroarrayデータを集め論文を書いていてJefferyらのHP[6]から入手できる。2015年10月28日から8種のLDFで判別した。3種のOLDFはMNM=0で，n個以下の遺伝子の判別係数が0でなく残り全て0になった。即ち自然にn個以下の遺伝子で癌遺伝子が特定できる。これをSM1と呼ぶ。このSM1を全遺伝子から省いて再度判別すると別のSM2が求まる。そしてデータは複数個のSMの排他的和集合の信号と，高次元のMNM≧1の雑音に分かれる事が分かった。1970年以降癌の遺伝子解析が行われ，有用な結果が得られなかった（問題5)。これは，統計手法で雑音を含んだデータの有効な分析ができない事と，SMの排他的和集合という特殊な構造が理論的に発見できないためと考えられる。新手法2で表1の結果を得た。章列は，2017年6月に出版したAmazonの章である（Shinmura,2017)。新手法2で3章から8章の結果を得た。即ち，Alonら(1999)は64個，Singhら(2002)は179個の排他的なSMがある。JMP列は，2015年10月に日本のJMPユーザー会で特異値分解を用いた高次元データが判別できるF-LDFが発表され，それを一時借用してNMを求めた。誤分類数と括弧の数字の誤分類確率は1.6％から16.8％で線形分離できない。多くの研究で，データはLSDである認識がなかったことが問題である。SM列は求まったSMの数である。2章とRatio列以降は，2017年にAmazonから出版した内容を紙面の節約のため併記する。

---

[5] このデータは，大学院の岡野さんが統計レポートに用いたデータである。彼女から研究に用いることの了解を得てある。

[6] http://www.bioinf.ucd.ie/people/ian/

表1 6個のデータの新手法2の結果（SM列まで）と遺伝子診断

章	データ	2群と患者数	JMP	SM	Max Ratio	Min Ratio	＞=5%	PCA
2	Alon et al.	Normal（22）vs. tumour cancer（40）		BGS130	0.90%	0.00%	0	4.50%
3	Alon et al.	Normal（22）vs. tumour cancer（40）	5（8.0）	64	26.76%	2.35%	63	30.40%
4	Singh et al.	Normal（50）vs. tumour prostate（50）	2（1.6）	179	11.67%	0.28%	38	14.35%
5	Golub et al.	All（47）vs. AML（25）	8（11.6）	69	15.69%	0.00%	13	34.88%
6	Tien et al.（2003）	False（36）vs. True（137）	3（3.9）	159	19.13%	0.63%	27	24%
7	Chiaretti et al.	B-cell（95）vs. T-cell（33）	10（9.8）	95	38.98%	10.73%	95	51.46%
8	Shipp et al.	Follicular lymphoma（19）vs. DLBCL（58）	29（16.8）	130	30.67%	4.99%	129	31.70%

9章：研究データは小標本のことが多い。これを100回コピーし疑似母集団を作成する。乱数を与えて昇順で並べ替え100組の学習標本を作る。疑似母集団を検証標本にして100重交差検証法を行う。LOO法のように検証標本が一定しない方法は問題である。新手法1で6種のデータの学習標本と検証標本を作成し判別係数と誤分類確率の95％信頼区間を求めて問題4を解決した。さらに検証標本の平均誤分類確率最小（M2）のモデルをBestモデルとして選ぶ。8種のLDFのBestモデルを比較し，M2の一番小さいモデルを最終的に選べば簡単にモデル選択ができる。そして，改定IP-OLDFが一番良いことが分かった。

4. 癌の遺伝子解析が困難な3つの言い訳

Golubeら[7]が指摘する通り「癌の遺伝子解析」は1970年ごろから行われてきたようだ。医学研究者はMicroarrayデータの発現量から真剣に従来の形態学などのアプローチでない方法を模索し既存の統計手法や新しい方法を開発し検討している。一方，統計研究家はBigデータ解析が次世代の研究テーマに格好であり，その中で質の高いMicroarrayデータが容易に利用できる世になり，多くの研究者が多くの研究を発表しているが，私が僅か54日で解決した結果に比べて見劣っているといわざるを得ない。その一番大きな原因は，統計的判別関数や一般的な手法が全く「癌の遺伝子解析に無力である」と断言できる。現在まで行ってきた研究をまとめると次のようになる。

[7] Although cancer classification has improved over the past 30 years, there has been no general approach for identifying cancer classes（class discovery）or for assigning tumors to known classes（class prediction）.

4.1 3つの言い訳

⑴ Bigデータの位置づけ

通常統計が対象とするのは，小標本の「Small n Small p」データである。このため得られた統計量の有用性を高めるため，推測統計学の体系が整備された。今日Bigデータが注目されているが，「Large n Small p」データと「Small n Large p」データと「Large n Large p」データに分けて考える必要がある。「Large n Small p」データは，現状のPCと統計ソフトで十分対応でき問題が少ないので検討対象から省く。ほどんとの検定が棄却され推測統計の役割が減少する。また，データを100分割して検証すれば新手法1が必要なくなる。「Large n Large p」データは，各種センサーや自動車に搭載のセンサーや携帯情報など，これまでの統計データと異質なデータを用いて工学的に成果を上げている。このため「Small n Large p」データが一番統計に適していて，それの代表が癌と正常あるいは異なった2種類の癌症例を100件ほどを対象に数千から数万個の遺伝子の発現量をMicroarrayで検査した発現量が研究に用いられている。そこで癌の遺伝子を特定する「Feature Selection Method」すなわちモデル選択で多くの研究が行われている。不思議なことに何を成功の評価基準にしているのかわからない。それで，論文の中には以下のように「癌の遺伝子解析が困難な3つの言い訳」がバズワードのように指摘されている。

⑵ 「Small n Large p」データの高次元解析の困難性

このデータに関して何が困難か整理されていないようである。1変数や2変数の分析であれば，計算結果が多くて整理が大変なだけである。問題は多変量であり，例えば分散共分散行列を求めること自体が難しい。2000年以前に国内や国際会議での発表を目にしたがいつの間にか消えていった。JMPが2015年のDiscover SummitでSASの創立者の一人のSall博士（2004）が基調講演で特異値分解を利用して，高次元のMicroarrayデータをF-LDFで判別できることが可能になったと報告した。筆者が質問し，現在Microarrayデータの分析中で，借用することにした。表1で示すように誤分類数は高いので，製品発表と同時に役に立たないことが分かった。よしんばNM=0であったとしても，例えば1万変数の中から後で紹介する少数のk個の遺伝子の組であるSMを見つけるには変数選択か全ての組み合わせモデルの検討が必要になる。しかし，MP-LDFではn≫pよりもn≪pのデータの判別の方が容易である。しかもMNM=0のデータでは，3つのOLDFも3つのSVMも10秒もかからずMNM=0あるいはNM=0であることが分かる。研究論文の中ではSVMの利用を行ったという記述があるが，明確な結果の記述はない。またMicroarrayデータを判別し，NM=0であったという事例もないのが不思議である。n≪pのデータでは，過推定すると誤解し検討しなかった可能性がある。プリチャード真理と江口真須透（2009）は，「関連遺伝子セットの多重解の存在」でMicroarrayデータから詳細な説明がないが，2つの遺伝子の組を用いてSVMで判別し複数の

NM＝0になる多重解があることを報告している。明確に記述してないが，どちらの解が正しいのか決めかねるような印象である。これはLSD判別の成果を知らないので，無数の多重解があることを受け入れられなかったためと考える。

⑶ NP-Hard

⑵と関係するが，pが大きいと“Feature Selection”すなわち統計でいう変数あるいはモデル選択に時間がかかることを言ったものと考えられる。しかし高次元データで逐次変数選択法や全てのモデルの組み合わせを考えることは現実的ではない。そこでLASSOなどで回帰係数や判別係数を0にすることが考えられたのかと考える。例えできたとしても，図1で示したように最適凸体の内点を求めることはできない。それ以前に，LSDを正しく判別できないことが問題である。

⑷ 信号をノイズから分離する困難さ

高次元空間で役に立つ信号はノイズに埋もれていて，その分離が困難であるといわれている。そのために工学的なフィルタリング手法が種々提案されている。しかし信号の定義は研究の初期には議論されたかもしれないが，調べた限りでははっきり示されていない。筆者の研究ではMicroarrayデータ自体がMNM＝0であり，癌と正常群が高次元空間で完全に分かれている。その上で，それが小さなMNM＝0になる部分空間に分割され，残りはMNMが1以上の雑音と考えている。これが，筆者の信号と雑音の定義である。

4.2 なぜ1970年以降「癌の遺伝子解析」はなぜ成功しなかったか

なぜ1970年以降「癌の遺伝子解析」は成功しなかったかの一番大きな理由は，分散共分散行列に基ずく判別分析が次の理由で全く役に立たないからであることが私の研究で分かった。

⑴ 相関比最大化基準による判別関数は，LSDの判別でNMは非常に高いことが多い。これは，小標本に分割したSMでも誤分類数が0にならないことが多いことでも裏づけられる。結局，癌の遺伝子解析には役に立たない。Wardクラスター分析やPCAでもSMで線形分離可能な兆候は見つけれなかった。2群の平均値の差の検定で，癌のクラスの平均は正常の平均値より高いという前提で，癌の遺伝子を探す論文もあるが，間違いである。SMのt値は，正からほぼ0に近いもの，負のものがある。以上から，癌の遺伝子はMatryoshka構造になるという特殊構造があり，通常の統計手法はほどんと役に立たない。

⑵ H-SVMとSVM4とSVM1は6種類のMicroarrayデータでNMは0になる。しかし判別係数のほどんとは0でないので，改定IP-OLDFのように判別係数の多くが0になり，自然に癌の遺伝子を特定したようなことができない。高次元データで変数選択を行うか全ての判別モデルの組み合わせを検討するのはNP-hardであるので，医学的な見地から遺伝子を特定

しない限り難しい。これは，SVMがQPで定式化されているため，統計的判別関数と同じく高次元空間の全定義域で一つの最適解しか求めることができないためである。

⑶ 改定IP-OLDFが高次元の遺伝子空間を，幾つかのSMに分割できたのは，IPのアルゴリズムの分枝限定法が全てのモデルの組み合わせを行ったのと同じ効果があるためであろう。このため，Matryoshka構造にある多数のSMの中から最初に見つけたものを最適解として出力する。そしてそれを省いて判別すると別のSMが求まる。

⑷ 改定LP-OLDFでも別のSMの組を見つけることができることが分かった。この特徴はまだMPの世界で一般に確認されていないが，LPも部分空間の解を見つけることができる。ただし，式⑹の改定LP-OLDFは式⑻のS-SVMの目的関数からQPが必要になる第1の目的関数を省いただけの違いである事実が重要だ。

5. From Cancer Gene Analysis to Cancer Gene Diagnosis (Shinmura, 2017)

Springerから出版した当初，SMは統計的には小標本であり，ロジスティック回帰などの統計的判別分析に加えて，一元配置の分散分析とt検定，クラスター分析，相関分析と主成分分析（PCA）等の標準的な統計手法で分析すれば，有用な情報が得られると考えていた。表1の2章では，AlonからLINGO Program4と手作業で130個のBGSを求めた結果を示す。3章以降で6種類全てのSMを，一元配置の分散分析とt検定，Wardクラスター，PCA，ロジスティック回帰，F-LDFとQDFで分析した。しかしロジスティック回帰だけが全てのBGSとSMがNM=0であったが，他の手法はLSDである兆候が得られなかった。そこで考えたのが，改定IP-OLDFは各BGSとSMを正常をSV=－1以下に，癌をSV=1以上で判別している。これらの判別の良さを示す統計量としてRatioSVを導入した。さらに，判別スコアを1個の変数と考えた。例えばAlonらでは130個のBGSが見つかるので，62例＊130変数のデータとして分析することにした。その結果，クラスター分析とPCAは完全に2群に分かれ，癌の遺伝子診断の可能性が開けたので，ここではそれを説明する。

5.1 BGSとSMの分析

ロジスティック回帰のNMは全て0[8]であり，新手法2のLINGOプログラム3は正しくMNM=0のSMを求めたことが分かる。しかし，Alon他の130個のBGSは，QDFは58個だけがNM=0で，F-LDFは130個全て0でない。64個のSMの判別ではQDFは全てNM=0で，F-LDFは13個だけが0になる。他の手法は結果は得られるが，線形分離可能な兆候は認めら

[8] ロジスティック回帰でLSDを判別すると，繰り返し計算が不安定になり回帰係数の標準誤差は異常に大きくなる。このようなモデルは，本来であれば利用できない。しかし，MNM = 0であり判別境界を移動してNM = 0になるものがある場合，LSDと判定している。

れない。t検定で癌遺伝子を探している研究もあるが, t値は－10程度から10程度にばらつき, 2群の平均に差がないt値が0に近いものも含まれる。すなわちt検定は，癌の遺伝子の特定の役には立たないと考えられる。1個の遺伝子で癌は特定できなく種々の遺伝子の組み合わせで癌が特定できるようだ。得られたBGSの役割は今後の課題である。すなわちロジスティック回帰以外の手法で，LSDの兆候は発見できない。

5.2 改定IP-OLDFの判別スコアの分析

6種のMPによるLDFは，SVで－1以下に正常患者，1以上に癌患者を判別した。しかし判別結果の評価基準がなかった。SVの距離の2が判別スコアの範囲の長さの何％であるかを表すRatioSV（=2/ 判別スコアの範囲*100）を考えた。これが表1のMax RatioとMin Ratio列である。Alon他の130個のBGSは改定IP-OLDFの中で最大は0.90％で最小は0.00％（0.005未満）である。これでは新規の患者の検証標本でよい結果が出るとは考えられない。しかしSMでは最大が26.76％で，最小が2.35％である。この値が5％以上のものが63個あり，1個だけが2.35％である。この閾値は今後の課題であるが，仮に5％以上とすれば，63個の改定IP-OLDFが癌診断に利用できる。すなわち，新手法1で考えた100重交差検証表の検討が不要と考えている。Chiaretti他（2004）の場合，約40％の大きな窓が開き残りの60％に2群が散らばっている。またWard法とPCAの結果は綺麗に2群に分かれる。

図2はAlonらの64個の判別スコアデータのPCAの結果である。左の固有値を見ると第1主成分の固有値が39.4とSpike状に大きい。これは2群が大きく離れ，2群の群間分散が大きいためと考える。右の因子負荷量は，4象限から1象限に布置している。真ん中のスコアプロットは，正常群がPrin1の負にほぼ直線上に布置し，癌は原点から正の方向に扇のように布置しているが，恐らく2群の平均は主成分上と考えられる。右の癌患者は原点近くの癌は比較的軽度の癌患者であり，正常の原点に近い患者は癌になりやすいか否かは，病理学者が患者カルテを調べれば分ると考えている。即ち，Prin1の主成分軸は63個の改定IP-OLDFの判別スコアと同じく癌の悪性度の指標に使えることが期待される。もう一つの応用としては，癌患者が治療で正常領域に誤判別されれば5年も経過観察なく治癒したと確定できると素人的に考える。主成分軸上のRatioSVは30.4％で26.76％よりも約4％も大きいのは64個の合成のためである。このため，症例を持っている米国の6研究グループに，RGを通して共同研究を呼びかけたが，医師の多くは会員でないこともあり返事はないので医学的な検証は困難であり大きな壁に阻まれている。一番大きな理由は，筆者自身がこれから著名な医学誌に論文を投稿する力量がないことである。

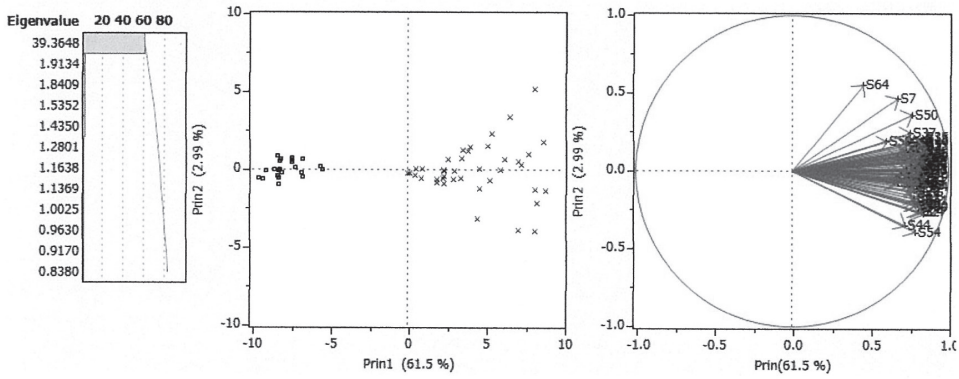

図2 PCAの結果

Aoshima & Yata（2017）は，高次元遺伝子空間で2群が離れた球上に布置し，このためPCAの第1主成分の固有値はスパイク状に第2主成分以下に比べて大きいことを指摘している。もし図2のスコアプロットがその陰であれば，図2の内容と合致している。しかし，なぜF-LDFのNMが0でないのか大きな疑問が出てくる。よほど2群の分散共分散の構造が，Fisherの仮説とは異質であるためであろうが，その具体像は分からない。

6. まとめ

MNM基準による改定IP-OLDF（と改定LP-OLDF）で1970年以降解決できなかった癌の遺伝子解析が僅か54日で解決できた。そして6種類のMicroarrayデータ全てでSMとAlonらのデータでBGSを見つけている。BGSに含まれる遺伝子を1個除くとMNMは0でなくなるので，RatioSVは著しく小さい。これまでの研究で，癌遺伝子を明確に定義した論文を発見できなかった。ここ10年の研究では，t検定やクラスター分析が多く用いられている。多分統計的判別関数は，2000年以前に役に立たないと医学研究者から見放されたのであろう。恐らくt検定で，t値の大きなものあるいは「近傍分析や荷重投票法」と呼ばれるこの分野で開発された手法で癌遺伝子として狙い撃ちして探していると推測できる。しかし筆者のSMとBGSの分析結果では，t値が正のものから，ほぼ0に近いもの，負のものが含まれている点である。負のものは，癌の抑制遺伝子と考えられるが，ほぼ0のものは他の遺伝子との交互作用を持つとしか考えられない。このような遺伝子は，t検定で特定できないし，従来の医学における癌の遺伝子の発見の範囲外にあると考えている。実証研究を行っている統計家の立場で言えば，クラスター分析は多くの手法があり，オプションの設定の違いで，数多くの異なったクラスターの結果が得られる。医学研究の場合は，研究者が医学的知見に基づきある程度正しい結論を得ていて，それをうまく説明するためにクラスター分析の結果を用いる

と考えるべきであろう。医学にAIを用いた研究が注目されているが，その中核はクラスター分析であり煩わしいクラスター分析の解読時間がセーブできる利点はある。

三宅，新村（1980）は，ヒューリスティックな方法によるMNM基準に関する筆者の最初の論文である。それ以前3年間ある編集委員を努めた統計関連の論文誌に投稿し続けたが，レフリーコメントで「内部標本の誤分類数を最小化する基準は，統計のイロハも知らない愚かな議論であり，外部標本の検証に耐えられない。それに対して正規分布を仮定して導かれたF-LDFは外部標本での検証でも良いことは常識である」として数度の修正にもかかわらず却下された。それが判別分析の一番の応用分野である「医用電子と生体工学」に簡単に受理された。また2010年ごろに投稿した論文は，「数理計画法を使った判別分析は，我々統計の文化となじまない」と棄却されている。しかし，SVMは統計家の多くがQPを使っているのに素直に受け入れている。これはVapnikが，パターン認識などの分野で発表し，多くの実証研究で広く認められたため，統計やORにも受け入れられるようになったと考える。このためSVMの研究発表は，統計やORより工学分野での発表がはるかに多く，重要な研究テーマを統計やOR研究は他の研究分野にSVMの研究者をとられたと考えるべきである。

結局，統計的判別関数が癌の遺伝子解析に役に立たず，t検定やクラスター分析，そして工学的なフィルタリング技術の研究が行われてきた。しかしMNM基準による改定IP-OLDFは自然にSMを見つけ，新手法2で高次元の遺伝子空間は複数の信号であるSMあるいはBGSの排他的和集合と線形分離可能でない雑音空間に簡単に分けることができる。多分LASSOはMNM基準を用いていないので，SMを正しく求めることができないと推測している。データはすでに公開されているので，早く実証研究で筆者の意見が正しいか否かを明白にすべきであろう。

以上から，Fisherの仮説に基づいて分散共分散行列を用いた相関比最大化基準による判別関数は，医学診断や各種格付けのような「地球モデル」で表されるデータに適していないことを6種類の実データで実証研究し新しい判別理論を提案した。さらに応用研究として1970以来行われてきて応用研究として取り上げた「癌の遺伝子解析」では「Matryoshka Feature Selection Method（新手法2）」では簡単に「高次元の遺伝子空間で癌と正常あるいは異なった2群の癌は完全に分かれていて，さらに癌を特定する64個から179個の小さな遺伝子の部分空間に分割された。これらは遺伝子解析が困難な3つの言い訳が単なる言い訳であり，アプローチが間違っていることを示す」。

最後の大きな問題は，青島と矢田が高次元のMicroarrayデータで2群は離れた2個の球体に張り付いていると主張し，筆者は図1のように2群はSVの大きな窓で完全に分離している。「なぜ，それがF-LDFでNM＝0で判別できないのか？」という大きな疑問である。恐らく2群の分散共分散の違いを示せば良いだろうが，筆者の目標ではない。問題ははっきりしていて，

データもあるので，まだ研究者人生に余裕のある人がチャレンジすることを薦めたい。

(成蹊大学名誉教授)

REFERENCES

1. Alon, U. et al. (1999). "Patterns of Gene Expression Revealed by Clustering Analysis of cancer and Normal Colon Tissues Probed by Oligonucleotide Arrays." *Proc. Natl. Acad. Sci*. USA, 96, 6745-6750.
2. Anderson, A. (1945). "The irises of the Gaspe Peninsula." *Bulletin of the American Iris Society* vol. 59: 2-5
3. Aoshima, M. and Yata, K. (2017). "Two-sample tests for high-dimension, strongly spiked eigenvalue models." *Statistica Sinica*, in press (arXiv: 1602. 02491).
4. Cox, D. R. (1958) "The regression analysis of binary sequences (with discussion)." *J Roy Stat Soc* B 20: 215-242.
5. Chiaretti, 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: 2771-2778
6. Firth, D. (1993). "Bias reduction of maximum likelihood estimates." *Biometrika*, vol. 80: 27-39
7. Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic problems." *Annals of Eugenics*, 7, 179-188.
8. ———— (1956). *Statistical methods and statistical inference*. Hafner Publishing Co.
9. Flury, B. and Rieduyl, H. (1988). *Multivariate Statistics: A Practical Approach*. Cambridge University Press
10. Friedman, JH. (1989). "Regularized Discriminant Analysis." *Journal of the American Statistical Association*, 84/405: 165-175
11. 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.
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. Sall, J. P., Creighton, L., Lehman, A. (2004). *JMP Start Statistics*, Third Edition. SAS Institute Inc. (Shinmura, S. edited Japanese version)
14. Schrage, L. (2006). *Optimization Modeling with LINGO*. LINDO Systems Inc.
15. Shinmura, S. (2016). *The New Theory of Discriminant Analysis after R Fisher*, Springer. DOI:

10.1007/978-981-10-2164-0

16. ――――（2017）. *From Cancer Gene Analysis to Cancer Gene Diagnosis*. Amazon Kindle Version.
17. Shipp MA, et al.（2002）. “Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.” *Nature Medicine* 8/1: 68-74. Doi: 10.1038/nm0102-68
18. Simon N, Friedman J, Hastie T, Tibshirani R（2013）. “A sparse-group lasso.” *J. Comput. Graph. Statist*, 22:231-245
19. Singh et al.（2002）. “Gene expression correlates of clinical prostate cancer behavior.” *Cancer Cell*: March 2002, 1/2: 203-209
20. Stam, A.（1997）. “Non-traditional approaches to statistical classification: Some perspectives on Lp-norm methods.” *Annals of Operations Research*, 74: 1-36
21. Taguchi G, Jugular R（2002）*The Mahalanobis-Taguchi Strategy - A Pattern Technology System*. John Wiley & Sons.
22. Tian 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
23. Vapnik, V.（1995）. *The Nature of Statistical Learning Theory*. Springer-Verlag.
24. 石井　晶（2015）. First principal component and its applications to tests of means and covariance matrices for high-dimensional data. 多様な分野における統計学の新展開. 1-10.
25. ――――（2017）. Asymptotic properties of classification procedure based on eigenstructures in high - dimensional context. 多様な分野における統計学の総合的研究. 1-7.
26. 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.
27. ――――（2004）.『JMP活用　統計学とっておき勉強法』. 講談社.
28. ――――（2007）.『ExcelとLINGOで学ぶ数理計画法』. 日科技連出版.
29. ――――（2010）.『最適線形判別関数』. 日科技連出版.
30. ――――（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.
31. ――――（2011b）.『数理計画法による問題解決法』. 日科技連出版.
32. ――――（2015）. 判別分析の誤分類確率と判別係数の95％信頼期間. 多様な分野における統計学の新展開（富山県民会館）. 1-10.
33. ――――（2017a）. 横長データの代表であるMicroarrayデータによる癌の遺伝子診断―退官記念講演に変えて―. Discovery Summit　2017（11月17日）. 発表資料はPPの48ス

ライドで，ResearchGateからダウンロード可。

34. ――――（2017b）．Cancer Gene Analysis by Singh et al. Microarray Data. 多様な分野における統計科学の総合的研究（新潟大学11月19日）．1-10.
35. ――――（2017c）．なぜ癌の遺伝子解析は30年以上成功しなかったのか？大規模複雑データの理論と方法論，及び，関連分野への応用（筑波大学11月19日）．1-10.
36. ――――（2018）．Cancer Gene Analysis using Small Matryoshka（SM）found by Matryoshka Feature Selection Method。生命・自然科学における複雑現象解明のための統計的アプローチ（滋賀大学2月16日）．1-10.
37. 竹内　啓（20011）．書評:小西定則「多変量解析入門―線形から非線形へ―」, 新村秀一「最適線形判別関数」．統計，71-74.
38. 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.
39. プリチャード真理，江口真須透（2009）．関連遺伝子セットの多重解の存在．日本統計学会誌,
40. 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.
41. ライナス･シュラージ（新村訳）（1992）.『実践数理計画法―LINDOを用いて―』, 朝倉書店。
42. ――――（新村訳）（2017）．LIINGOを用いて種々の最適化問題を実際に解決しよう。Amazon Kindle版。