JAPAN
DISCOVERY SUMMIT
EXPLORING DATA
INSPIRING INNOVATION

退官記念講演にかえて

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

成蹊大学　名誉教授
新村秀一

1. 発表の概要

- 大学卒業後、NECと大阪府立成人病センターの共同プロジェクトで、心電図の自動診断論理を判別分析で4年間研究した。
  - しかし、医師の開発した枝分かれ論理にかなわなかった。Why!?
  - 当初、自分の未熟さと考えたが、統計的判別分析が医学診断に使えないと結論した。
  - これが、新しい判別理論を開発しようという、研究の動機
- その後、多くの医学データの実証研究を、SASの計算サービスやVax/SAS（32製薬メーカ）の代理店をやりながら行った。
- 2016年末に判別分析の5つの問題を解決し、Springerから出版した。また30年以上成功していない癌の遺伝子解析に世界で初めて成功し、2017年6月にAmazonから出版した。

2. Fisherの線形判別関数(F-LDF)

2.1　判別する2群が、平均だけ異なるガウス分布である

（Fisherの仮説）：良い検定がない

- 分散比（相関比）最大化基準で導いたと紹介し、偏微分……。
  - 最尤推定を用いていない
- 英明なFisherは計算機環境の乏しい時代に、指数関数で表される正規分布の比の対数を取れば、簡単にF-LDFの判別超平面が線形式になることから導いたと考える。$\mathrm{Log}(f_1(x)/f_2(x)) = bx+c = 0.$
- 統計ソフトは、分散共分散から一意にF-LDFが導出でき便利である
  - 統計ソフト企業は大きくなり　〈…〉　数理計画法
  - 統計ユーザーが増えた発展した。
- Fisherらは、記述統計学を数学のように純粋理論に近づける目論見は見事に成功した

2.2 Fisherの叡智をなぜ学ばないのか？！

- 田辺國士（2011）. 応用数理の遊歩道(67)帰納という原罪. 21/4, 304-309.
  - 有意性検定において検証される仮説は「帰無仮説」と呼ばれます。この仮説の評価は統計量の観測データに基づいて行います。その手続きは、まずこの仮説が真であると仮定したときに観測されるべき統計量の分布を数学的に割り出し、この仮説の下で稀にしか起こらない事象群を定め、観測データがこの稀な事象群に入るなら、観測データから見てこの仮説は考慮に値しないものとして「棄却」します。**棄却の論理的な意味は「仮説が真であるがその下で非常に稀な事象が置きたか、あるいは仮説自体が真でないかのどちらかである」とFisher自身が述べています。**
  - Fisher R.A. Statistical Methods and Scientific Inferences. Hafner Publishing. New York, 1956.

2.2 Fisherの叡智をなぜ学ばないのか？！

- Fisherと同時代の研究者は、Fisherの仮説を満たさない場合、2次判別関数（QDF）を推奨した。
- Fisherはアイリスデータで実証検証した。
  - 正規分布で学習標本と検証標本で評価することは馬鹿げている
- 最尤推定法を提唱したが、F-LDFに用いていない。
  - 極大値しかもとめず、真の最大値は分からない
- Fisherは決して、正規分布の無条件信仰を普及したわけではない
  - しかし、正規分布を不磨の大典の如く信仰している研究者も多い
  - それが、40年以上癌の遺伝子解析が失敗した原因であ
  - 統計的判別関数が役に立たず、…。

jmp Statistical Discovery.™ From SAS.
Copyright © 2017 SAS Institute Inc. All rights reserved.

2.3 私的判別分析の流れ

- 多くの人は、次を判別分析を主流と考えている
  - 分散共分散による判別分析
    - F-LDF（QDF） → RD → LASSO
- 私的な正統派
  - F-LDF（QDF） → logistic回帰 -> SVM
- ロジスティック回帰は、Fisherが開発した最尤推定法で、データに依存した繰り返し計算する。
  - 統計手法で唯一使ってよい判別手法。しかし、F-LDFとQDFは歴史的な意味があるので、JMPはサポートを継続すべき。
- Vapnikは、数理計画法の2次計画法（QP）で定式化したが、統計やORでなく、パターン認識で普及を図った。Why？

2.4 検討した8LDFs

- 本研究では、F-LDF(QDF)とロジスティック回帰と、6種の数理計画法(MP)によるLDFsを取り上げ比較評価。
- MP-LDFsは、Stam (1997)が米国のOR学会誌に総括論文を出して1997年以前の研究を総括している。
  - なぜ統計ユーザーはMP判別を利用しないのか？
  - SVMは多くの実証研究をパターン認識の学会で行っている
- Log (p/(1-p)) = f(x) (1)

p: class1に属する確率; x: 独立変数.

  - 新村の地球モデルと同じ。
  - 変数xが連続的に大きくなると、正常から異常への確率が高くなる。
  - 判別境界に異常所見が多い。

SVM (Support Vector Machine)

- H-SVMはLSD (Linearly Separable Data) を明確に定式化したが、LSDでないデータに適用できないのでLSD判別の研究はない。
- ハードマージン最大化基準

H-SVM: $MIN=||b||^2/2;\quad y_i*({}^t x_i\, b + b_0)>=1;$ （3）

$y_i = 1 / -1$ for $x_i \in$ class1/class2; x_i : p-変数;

b: p-判別係数; b_0 : 定数項、自由変数.

- S-SVM: $MIN = ||b||^2/2 + c^* \Sigma e_i;$

$y_i*({}^t x_i\, b + b_0) >=1- e_i;$

c: penalty c; e_i : 非負決定変数.

  - SVM4 (C=10000), SVM1 (C=1). （4，5）
- Kernel SVM

新村（2010）. 最適線形判別関数(Optimal LDF)

- IP-OLDF: $MIN=\Sigma e_i;\quad y_i*({}^t x_i\, b+ 1)>=- e_i;$

e_i: 0/1 整数変数;

- RIP: $MIN=\Sigma e_i;\quad y_i*({}^t x_i\, b+ b_0)>=1-M^* e_i;$ (6)

e_i: 0/1 整数変数; M: 10000; b_0: 自由変数.

  - MNM基準をIPで定式化
  - MNMはNMより優れている
  - RIPが遺伝子解析を54日で解決した
- Revised LP-OLDF (7)
  - RIPのe_iを、非負の実数で定式化。
- S-SVMの目的関数の第2項と同じ
- Revised IPLP-OLDF：RIPとRevised LP-OLDFの混合モデル （8）

3. 判別分析の5つの問題

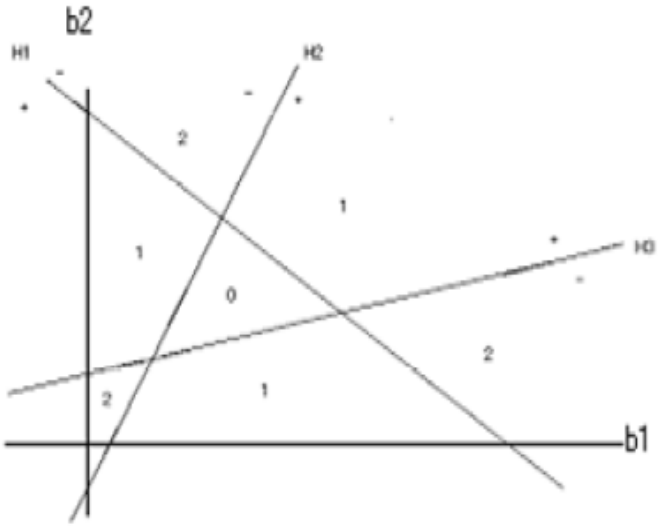
- 判別分析の5つの問題を、新しい判別理論で全て解決
  - （問題1）判別超平面上のケース($f(x_i)=0$)は、どちらの群に属するかは判定不能
  - （問題2）線形分離可能なデータ（LSD)を正しく判別できるのは、ハードマージン最大化SVM(H-SVM)とRIP
    - F-LDFの誤分類数（NM)は、非常に高いものもある。
  - （問題3）一般化逆行列は、一方の群の変数が一定値を取る場合、2次判別関数（QDF)は、全てを他群に誤判別。解決に3年かかった。
  - （問題4）判別分析は、推測統計学手法ではない。
  - （問題5）癌の遺伝子解析は30年以上成功していない。

3.1 問題1

- 判別超平面上のケースはどちらの群に属するかは判定不能
- 多くの統計家は$f(x_i) >= 0$であれば1群、 $f(x_i) < 0$であれば2群に判別と考える。
- しかし、「$y_i * f(x_i) > 0$であれば正しく判別され、 $y_i * f(x_i) < 0$であれば誤判別され、 $f(x_i) = 0$は分からない」
  - もし、大門2個の得点合計が50点以上を合格、49点以下を不合格とすると、$f = T1+T2-50 >= 0$で合格、 $f < 0$で不合格とする自明なLDFがあり、最小誤分類数（MNM)=0 で判別可能。
  - 等号を含むのは、説明変数で合否判定の規則が定義
  - しかし、F-LDFやQDFのNMは非常に高い。
  - 大学入試センター試験で、誤分類確率が20%を超える

判別分析の2つの新知見

問題1：$f(x_i)=0$のケースの扱い
問題5：MNM=0になる部分空間



- データ　(3症例　＊　2変数)
  - Class1:　x1=(-2, -3)
  - Class2:　x2=(-2,1), X3=(1, -3)
- IP-OLDF
  - MIN=Σ e_i;

    y_i ＊ (　-2b1 - 3b2+1　)>= - M*e_i;

    y_i ＊ (　-2b1 + 1b2+1　)>= - M* e_i;

    y_i ＊ (　1 b1- 3b2+1　)>= - M* e_i;
- NMと判別関数の関係と問題1の解決
  - 頂点や辺上にケースがあると、NMは増える
  - 内点に対応したLDFは、判別超平面上にケースがなく、問題1を避けることができる。
- RIPは最小NMすなわちMNMを持つ最適凸体の内点を唯一求める。
- スイス銀行紙幣データ：真札と偽札各100枚、6変数
  - (X4, X6)がBGS。(X1, X2, X3, X5)の判別係数が0で最適凸体の内点を求める。

NMとMNMの比較

- NMの問題点
  - 判別手法によって値が異なる
  - 判別超平面を動かすとNMは異なる(ROCで評価)
  - 事前確率:　等確率(デフォルト)と
    ケース数に比例(こちらが良い)
  - 問題1が生じると、判別超平面上のケース数だけ増える可能性
- MNM
  - 愚かな基準と言われてきた
  - データにUnique
  - 教師データのフルモデルが最小であり、教師データでモデル選択できないことを示す
  - MNM=0と線形分離可能(LSD)は同じ

3.2 問題2 （LSD判別の研究がない）

LSD判別は、新村しか行っていない

論文の査読者は、判別分析の目的はoverlapデータを判別

- スイス銀行データ：1000フラン紙幣の真札と偽札各100枚をFluryら(1988)が6個の計測値を計測し、判別分析の本を出版。
  - (X4, X6)でMNM=0になる。 (X4, X6)を癌の基本遺伝子 (BGS)
  - MNMの単調減少性：$MNM_k \geqq MNM_{(k+1)}$
  - もしMNM_2 =0なら、この2変数を含む16個のモデルのMNMは全て0で信号、残り47モデルのMNMは1以上で雑音。
- LSDはMatryoshka構造になるという表現に築かなかった

分散共分散に基づく判別関数の問題

LSD判別でNM（または誤分類確率）が高い

- 試験の合否判定
  - 大学入試センター試験の3年間の全科目の本問と予備問を解析し、誤分類確率が高く合否判定できなかった
  - 成蹊大学で、2010年から2015年まで、経済学部1年生を対象とした「統計入門」
    - 約120名
    - 10択100問の中間と期末試験を実施。
    - 大学入試試験の結果と同じ
- 大学入試センター試験の確認：誤分類確率が2割を超えるものもある。
- 問題3を数学ⅠとⅡで発見、統計入門で確認

3.3 問題3 （一般化逆行列の瑕疵）

X1とX3単独でMNM=0

すなわち、2個のBGS

p	Var.	LDF	QDF	MNM	λ=γ =0.8	0.1
1	Emission (X1)	2	0	0	2	0
2	Price (X2)	1	0	0	4	0
3	Capacity (X3)	1	29	0	3	0
4	CO_2 (X4)	1	29	0	4	0
5	Fuel (X5)	0	29	0	5	0
6	Sales (X6)	0	29	0	5	0

- QDFで判別するとX3が入ると、普通車全てが小型車に誤判別
  - 小型車の座席数が4で、普通車が5席以上のため
  - 当初、QDFが無条件にRDAに切り替わるので理由が分からなかった。
  - 現在は、RDAはユーザーが2つのパラメータを指定・・・面倒でデータごとに異なるチューニングの問題
- 一定値に小さな乱数を加えて、変動さすだけで簡単に解決
- 問題解決に3年かかったが、単に多変量によるアプローチが悪かった。

3.4 問題4 （新手法1で推測統計が可能）

- 判別分析は、誤分類確率や判別係数の標準誤差のない推測統計学と無縁な手法である。
- 「小標本のための100重交差検証法（新手法1）」を開発
  - JMPで標本を100回コピーし、疑似母集団（検証標本）を作成
  - 乱数で並べ替え、100組の教師標本を作る
  - 100組の教師標本を判別し、100組の教師標本の判別係数とNMを求め平均誤分類確率（M1）を計算。
  - これらのLDFを検証標本に適用し、NMから平均誤分類確率（M2）を求める。M2最小モデルをBestモデル
  - LDFとロジスティック回帰は、JMPのスクリプト。
- 全ての判別モデルで、M2最小のモデルをBestモデル。
- 多くのデータで、RIPが他の7種のLDFより良かった。

4. 問題5

- 2014年まで、新手法1で求めた判別係数の解釈ができなかった。
  - 合否判定データで得られた判別係数の定数項で判別係数を割ると、F-LDFを除いた7種の判別係数は自明な判別関数になった。
- 2015年10月25日に、富山市で開催された統計シンポジュームでこれを報告。4つの問題を解決し、研究が終了と早合点。
- 翌26日に、米国の6研究グループがNew England Journal of Medicineやサイエンスに論文を発表しデータを公開していることを知った（1999年から2004）。
- 10月28日に6種のMicroarrayデータをダウンロードし、12月20日にすべて解析し、「Matryoshuka Feature Selection Method（新手法2）」で驚く結果を得た。

4.1 概略

- ハーバード大学医学部教授のGolubら（1999）は、サイエンスに発表した論文で30年以上研究し、統計的に良い結果を得ていないと告白。
  - すなわち1970年から凡そ半世紀研究し結果が出ず、
  - 外野席では無理と考えていて、
  - おそらく研究の終焉報告。
- これは、分散共分散に基づく判別分析が、
  - LSDの判別も正しく行えず、
  - 多数ある部分空間のMNM=0になる、小さな遺伝子の組を見つけることができないため
  - F-LDF同様、ロジスティック回帰、SVMも定義域で一つの解しか求めない

4.2 経緯 （10月28日から12月20日）

- 10月28日 : Shippら(2002) を判別し、 7129個の遺伝子の32 個の判別係数が0でなく、残り7097 個が0 (Shinmura, 2015e).
  - 遺伝子空間全体がMNM=０で、変数選択しないで自然にn=77以下のMNM=0の部分空間が求まった。
  - これらが癌を特定する遺伝子である。
  - 癌の遺伝子解析でMNM=０になる空間と部分空間をMatryoshka と呼ぶ。
  - できるだけ早く世界に知らせるためResearch Gateにフリーペーパで投稿を決定。
    - これが一番早く、研究のタイム・スタンプを押す近道
    - 私が、医学誌に投稿する能力がない。

11月1日: (Shinmura, 2015f).

- 8種のLDFで、Alonら(1999) 、Shipp ら（2002）、Golubら(1999)を判別し、6種のMP - LDFでLSDを確認。
- F-LDFとロジスティック回帰はエラー。
  - 残りは32bitのExcelに展開できなかった。
- 3種のOLDFの判別係数はn以下で0でなく、遺伝子解析が可能。
- SVMの多くは0でないので、遺伝子解析に利用不可能。
- F-LDFやLassoはMNM=0であることも指摘できないであろう。

- 11月3~11月9日(Shinmura, 2015g-i):
  - 求めたSMを全体の遺伝子から省いて判別するとまた別のSMが得られた。これをrepeated feature selection method.
- 11月11日: (Shinmura, 2015j).
  - 2015年のDiscovery Summitの基調講演で、JMP ver.12を用いたMIcroarrayデータの判別例が紹介。
  - 1か月評価版を借りて3種を判別すると、AIon らはNM=0で残りはO.
  - 分析結果を技術担当に報告すると、暫くしてAIon らは5に変更された。
  - 20年ほど前に、「small n large p　問題」の典型である100件程度のデータで1万個程度の変数の分散共分散を推定し、この問題にアプローチする研究があり、いつの間にかなくなった。
  - JMPが、高次元データ用のF-LDFを提供してくれたことで、F-LDFが遺伝子研究に利用できないことが確認できた。
  - Rなどで出た成果では、確定的なことが言えない。
  - もしLASSOをリリースしてくれれば、LSDであることもいえないという結果が出ると期待される。

JAPAN DISCOVERY SUMMIT EXPLORING DATA INSPIRING INNOVATION

- 11月18日と22日：（Shinmura,2015k-i).
  - Matryoshka（Trap of）Feature Selection Methodを Golubで提案
    - 手法2を手作業で行うが、全てのSMをリストアップできない
- 12月4日：　(Shinmura, 2015m-o)
  - そこで、LINGO Program 3 of Method を完成し
  - 3種の全てのSMを作成.
- 12月6日～20日 (Shinmura, 2015p-s)
  - 64bit版のPC、OS、Excelを購入し、Singh（2002）, Tian（2003）、Chisretti（2004）　の全てのSMのリストを作成。

jmp Statistical Discovery.™ From SAS.
Copyright © 2017 SAS Institute Inc. All rights reserved.

4.3　問題５の解決

章	Dataset	2群と患者数	JMP	SM	RatioSV	PCA
2	Alon			130	[0, **0.9**]	**4.50%**
3	Alon	Normal (22) vs. tumour cancer (40)	**5 (8.0)**	**64**	[2, **29**]	**30.40%**
4	Singh	Normal (50) vs. tumour prostate (50)	**2 (1.6)**	**179**	[0.2, 12]	14.35%
5	Golub	All (47) vs. AML (25)	**8 (11.6)**	**69**	[0, 16]	34.88%
6	Tien	False (36) vs. True (137)	**3 (3.9)**	**159**	[0.6, 19]	24%
7	Chiaretti	B-cell (95) vs. T-cell (33)	**10 (9.8)**	**95**	[11, 39]	**51.46%**
8	Shipp	Follicul (19) vs. DLBCL (58)	**29 (16.8)**	**130**	[5, 31]	31.70%

- 米国の著名な６研究グループが、論文に使ったデータを検証のため公開
- 30年以上研究が行われ、外野席では癌の遺伝子解析は無理と考える風潮。
- それを僅か54日で対象データが、排他的なSMの和集合とノイズの遺伝子の部分空間である特殊構造。
- 彼らの統計的な結論と、私の結果を比較すれば一目瞭然。
- 40年もかかったのは難しい問題でなくアプローチが悪かったため。
  - 問題3の解決に3年かかった。問題5は、研究人生で一番簡単なテーマである。

4.4 Springerから出版

6種の普通のデータで、判別分析の新理論を検証。全て8個のLDFを新手法2で評価。JMPの分析結果の表と図であり、分かり易い

1章：判別分析の新理論の紹介

2章:アイリスデータとFisherの仮説（問題4）

3章：多重共線性のあるCPDデータ（問題4）

4章:学生データ（問題1、4）

5章:合否判定（問題2、4）

6章:スイス銀行紙幣データとBestモデル（問題2、4、5）

7章：日本車データ（問題3、4、5）

8章：新手法2と6種のMicroarrayデータ（問題5）

9章:LINGOでMP-LDFのプログラム紹介

5. 本年6月に癌の遺伝子診断をAmazon出版

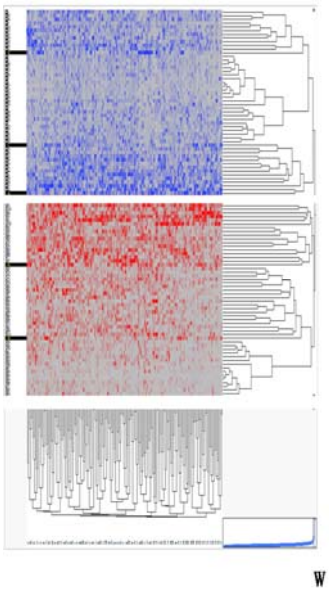
- From Cancer Gene Analysis to Cancer Diagnosis(1102円)、Unlimited会員は無料
- Alon関しては130個のBGSの排他的和集合で雑音空間がない
- 即ち、彼らが医学的に見つけた2000個の癌遺伝子は、130個の排他的なBGS（癌の基本遺伝子）の和集合である。
  - 49遺伝子の検査で約10万円。誤分類確率は高い。
  - 10万円*40= 400万円の検査が10万円でMNM=0
- 6種のデータの全てのSMは、小標本でありJMPで統計分析し、有用な情報が得られると期待したが、
  - ロジスティック回帰だけがNM=0
  - F-LDF、QDF、PCA、クラスター分析、一元配置の分散分析とt検定は、2群がLSDである兆候を示さなかった。
  - しかし、ある工夫で驚く結果が…

5.1 RIP判別スコアとRatioSV

- H-SVMとRIPは、BGSとSMをMNM=0で判別できるが、その評価尺度としてRatioSVを導入

  RatioSV＝SV間の距離2/RIPの判別スコアの範囲*100
- 6種のデータの最大値の範囲は[11.67%, 38.98%]
- 癌の悪性度指標と考えられるが専門医の検証が必要
- n件*k個の判別スコア（変数）とするデータで、PCAとクラスター分析は2群が完全に分離できる。
- RatioSV of PCAは、2群を明確に分離
- このデータの転置行列は、幾つかのRIP判別スコアが特異的であることを示した。

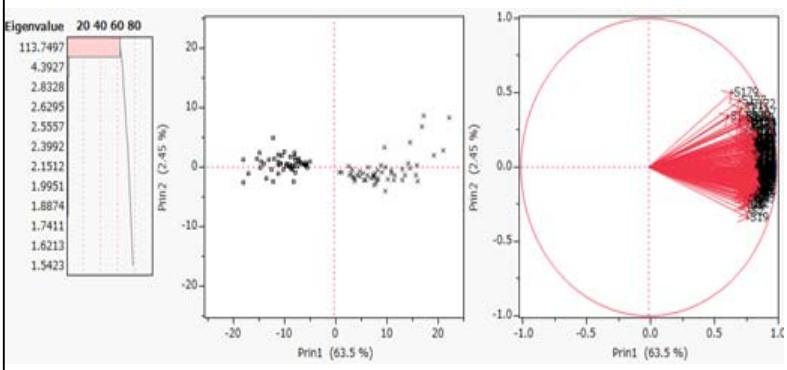
5.2　RIPの判別スコアデータの検討



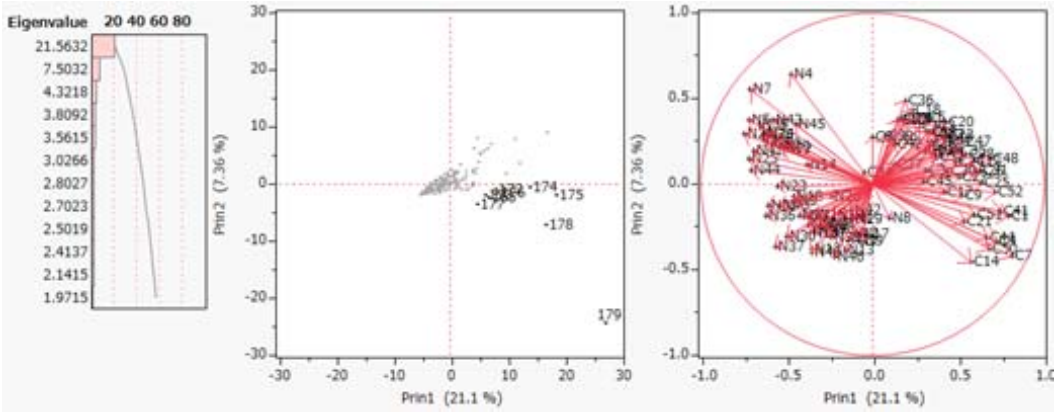
- Ward法で分析すると、例えばSinghらのデータでは、上の50正常症例、下の癌の52症例に完全に分かれる。
  - 専門家は遺伝子の樹状図から有意義な意味を導き出せると期待している。
- 他のデータでも同じ結果になる。
- 癌は遺伝子の病気であり、本来遺伝子で判別できて当たり前、という常識

5.3　PCAによるRatioSV

- 左の固有値は、第1主成分だけが異常に大きい
- スキャタープロットは、
  - 正常の50例が負のPrin1上に、癌患者の52例は原点から扇状に散布。
  - Prin1を悪性度指標を表すと考えれば、RatioSVは52%にもなる。
  - 非常に簡単な判別で、遺伝子研究以外では、論文にならない。
  - 研究者人生で、一番簡単なテーマ



5.4 転置行列の分析



- RIP判別スコアデータを転置して分析すると、個々の判別スコアで特異的なものが、6種類で分かる。
- これらの意味は、データを集めた専門家であれば、有用な意味が説明できるかもしれない。
  - 共同研究の呼びかけ。

6. JMPユーザー会の皆さんへの提案

6.1 人に馬鹿にされるくらいの高い目標を掲げる

- 大学卒業後、Fisherの仮説（正規分布、分散共分散）に基づく判別分析は、医学診断、各種格付け、試験の合否判定に不適と考え、MNM基準によるOLDFを提案した。
- JSCSに数編、応用統計と統計数理の企画号で各1編だけ掲載。それ以外の統計の学術誌は、全て棄却。
- 一方、医学系で棄却されたことがない。
- ORの月刊誌は多数論文があるが、論文誌はすべて棄却。
- 大学の紀要の活用： 半年ごとに出版される紀要は、研究記録の代わりに最適であり、出来立ての成果を最短で発表可能。

- 5年ほど前に大学の統計専攻の同僚から、研究をしていないと説教される。
- しかし、4年ほど前にResearchMap（日本の研究者DB）で成蹊大学の教員の順位が公表されると、理工学部の小島教授に次いで断トツの2位。
- 意味不明の査読結果の場合、気難しい統計やORの学術誌を避け、医学や理工学の応用誌に投稿。
- あるいは、SASやJMPのユーザー会で発表。
- Vapnikは、賢く統計やORを避け、パターン認識で普及を図った。
  - 統計は、多くのSVMによる判別分析の研究を失った。
- 私は本来統計の世界から除去されるはずだが、日本にSASやLINDOの紹介実績で生き残った？

JAPAN DISCOVERY SUMMIT EXPLORING DATA INSPIRING INNOVATION

6.2 世界への普及

- 2013年秋（翌年の定年を前）に、
  - 20年間自費で3000万円の研究費（大学より約1000万円）の圧縮と、世界へ普及のため、日本の費用の掛かる年4~5回の発表を停止し、英語で6頁以上のレフリー付きの論文3編以上を目指す。
- 2014年秋にResearch Gateを知る（無償でUPと利用）
  - 最初の1年は遅々とした実績だが、3年間でReed数6500、引用文献数1300の手ごたえを得た。
- 2015年10月に問題5というBigな応用問題を発見。
  - 世界中で40年以上成果がなく、無理と考える風潮があったが、自分の研究人生で一番優しい応用研究であった。
  - しかし、無職になって研究がピークになるとは！トホホ

JAPAN DISCOVERY SUMMIT EXPLORING DATA INSPIRING INNOVATION

6.3　なぜ30年も成功しなかったか？

- 既存の判別関数が、全くMicroarrayデータの特殊構造に適していなかった。
  - そのため、t検定やクラスター分析といった寝ぼけた統計手法に頼る。
  - AIに組み込まれたクラスター分析に、近年医学者は期待。
- 既存の判別関数は、MNM=0になる全遺伝子の中から、MNM=0になる部分空間であるSMの組を列挙できない。
- MNM=0すなわちLSD判別は、筆者しか行っていない。

F-LDFの問題点

- 分散（相関）比最大化基準に基づいているが、2群が正規分布であれば、指数関数の特徴で計算機環境の乏しい時代に、Fisherは簡単にF-LDFを得られることに気づいた。
  - 最尤推定法も用いていない。
  - Irisデータによる検証
  - 同時代、Fisherの仮説を満たさない場合、QDFが提案
  - 分散共分散行列に基づく判別関数は、LSD判別でも誤分類確率は非常に高い場合（試験の合否判定）。
- LSDでMNM=0にならない手法は、遺伝子解析には不適節
- また、部分空間の解を求めず、全遺伝子で唯一の極値解の推定値を求めるだけ。

SVMの問題

- SVの距離の最大化基準による最大値を選ぶ。
- 2次計画法(QP)を用いているので、全遺伝子空間で最大値を求め、部分空間の最大値を求めることはできない。
- 全てのモデルを検討すれば、部分空間の最適解を求めることができるがNP-Hardになる。
- SASのNLINに格子探索がある。初期値を変えることで局値解を場当たり的に探索するが、本来は数理計画法のNLPで大域的探索が必要。

なぜ改定LP-OLDFも部分空間の最適解を求めることができるのか？

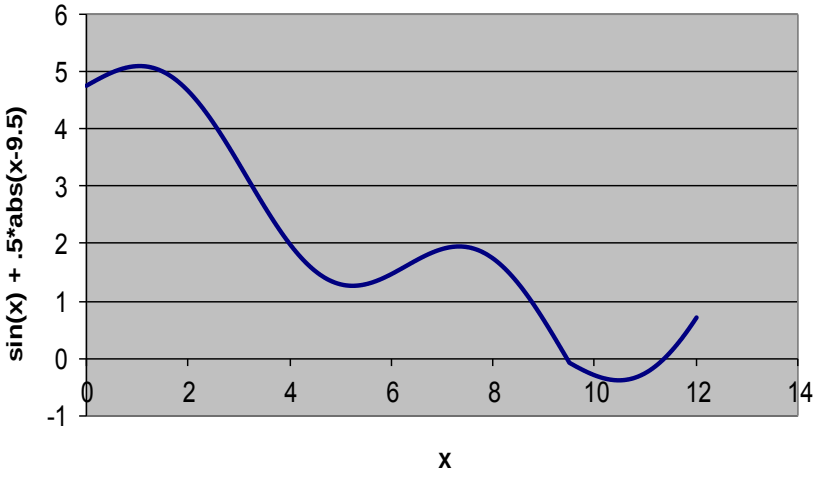
- 最小値/最大値と極小値/極大値の違い
  - 微分や偏微分は・・・
  - NLPの知識が必要
- $MIN = \Sigma e_i; \quad y_i \times ({}^t x_i b + b_0) >= 1 - M \times e_i; \quad (2)$

  b_0: 自由変数. e_i:非負の実数

  $MIN = \|b\|^2/2 + c \times \Sigma e_i; \quad y_i \times ({}^t x_i b + b_0) >= 1 - M \times e_i \quad (3)$
- 改定LP-OLDFは、誤分類される症例のSVからの距離の最小化。
  - MNM=0であれば、距離の和は必ず0という最小値になるので、MNM基準のRIPと同じ働きがある。
  - S-SVMの第2項目だけを用い、QPを用いていない。
  - QPは、全遺伝子空間で最小値を求める。

JAPAN DISCOVERY SUMMIT EXPLORING DATA INSPIRING INNOVATION

局所解と大域的最適解



非凸関数

(SIN(x)+0.5*ABS(x-9.5))

- 初期値の違いで$x = 0$, $x = 5.235987$, $x = 10.47197$の何れかの解を求める.
- Globalオプションを指定すれば, 解$x = 10.471$が大域的最適解

6.4 提案

- 長らく編集委員をやってきた学術誌で、「MNM基準は、統計のイロハを知らない愚かな提案で、正規分布を仮定したF-LDFが検証標本でも結果も良いに決まっている」というリジェクトコメントで退会し、日本ME学会に投稿しすぐに採択された。
- 時系列的に各種問題を順次解決し、論文にする必要があり、査読者とのやり取りに不要な時間を割くのを避ける
- 最近の国際会議で、査読付きで6頁から8頁の論文を受け付けるものや、Freeジャーナルが多く、編集長が最終判断。
- それらを全てRGにUPし、Impact factorでなく、引用文献数で勝負すべき。
  - 単に成果物のPDFを無償でUPするだけ

6.5　共同研究の提案

- ボブディランでもノーベル文学賞をもらった。
- もし統計でノーベル医学賞が認められるのなら、この研究は最有力候補と考える。
  - 一緒に夢を見ませんか？
- 共同研究探しています
  - 製薬メーカー
  - 大学医学部や病院
  - SAS、JMPやIBMといった統計ソフト
  - Microarrayチップの開発メーカー
  - 情報起業の新規事業（癌の遺伝子解析）

JAPAN DISCOVERY SUMMIT EXPLORING DATA INSPIRING INNOVATION

6.5　JMPへの期待

- 統計家とJMPの立場は異なる。
  - 統計家は10年以上前、Rなどを用いて横長データのF-LDFを開発していたが、研究の結末が不明。
  - JMPが2015に横長データのF-LDFを提供したことで、遺伝子解析に不適なことが証明できた。
  - LASSOも提供してくれれば、遺伝子解析に不適なことがすぐに証明でき、自分の研究時間を節約。
- 癌の遺伝子解析という大きなテーマに対応するために、MP判別を開発する元気な企業は手を挙げてください。

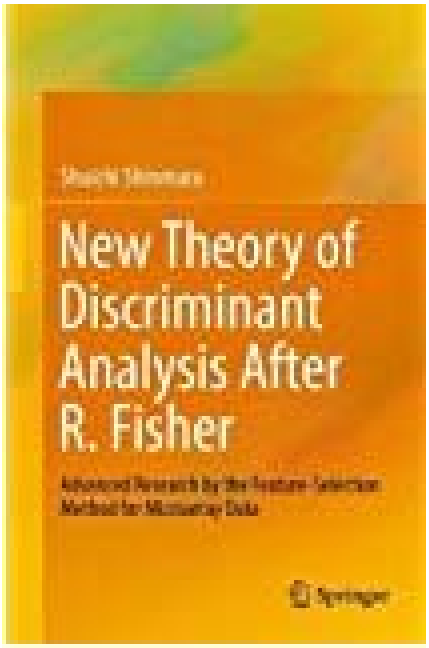
JAPAN DISCOVERY SUMMIT EXPLORING DATA INSPIRING INNOVATION

jmp Statistical Discovery.™ From SAS.
Copyright © 2017 SAS Institute Inc. All rights reserved.

6.6 問題点

- 今回の結論は、分析に用いた6種のデータの範囲。ただし、他のデータもMNM=0と期待。
- クラスター分析で、非常に近い距離の遺伝子のペアが数組観測されるが、これらが互いに互換であることの影響。
- 私の能力を超えた、癌のしぶとい狡猾な戦略に打っちゃりを食らうこと。
- 変数名が遺伝子名を他の変数名に置き換えている。

New Theory of Discriminant Analysis after R. Fisher



1章:新理論の紹介

2章 : Irisデータ

3章:CPDデータ

4章:学生データ

5章 : 合否判定データ

6章 : スイス銀行紙幣データ

7章:日本車データ

8章:6種のMicroarrayデータ

From Cancer Gene Analysis to Cancer Gene Diagnosis

Amazon Kindl: $9.99 447 pages

1 New Theory of Discriminant Analysis and Cancer Gene Analysis

2 Cancer Gene Analysis of 130 BGSs of Alon et al. Microarray Dataset

3 Cancer Gene Analysis of 64 SMs of Alon et al. Microarray Dataset

4 Cancer Gene Analysis of Singh et al. Microarray Dataset

5 Cancer Gene Analysis of 69 Small Matryoshkas (SMs) of Golub et al. Microarray Dataset

6 Examination of 159 Small Matryoshka (SM) of Tian et al. Microarray Dataset

7 Examination of 95 Small Matryoshka (SMs) of Chiaretti et al. Microarray Dataset

8 Examination of 130 Small Matryoshka (SMs) of Shipp et al. Microarray Dataset

9. Validation of Matryoshka Feature Selection Method by LINGO Program 1 using Common Data

Amazon：ビジネス、研究と教育のための問題解決学シリーズ

- 全て337円、Unlimited会員は無料

1：DEAによる可視的評価法

電力10社の3.11の前年と翌年、鉄道18社、日本の空港、東京都24区の区立図書館

2、3、4：シカゴ大学L.Schrage教授の叡智に学ぶ問題解決学

- 147冊の文献調査、それをLINGOの雛型モデルを作成

5：Excelによるビジネスと教育のための問題解決学

「Amazon」で「新村秀一」で検索。

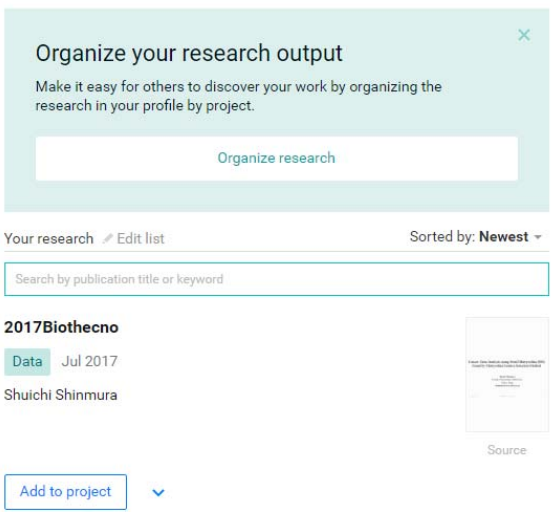
最後の一言

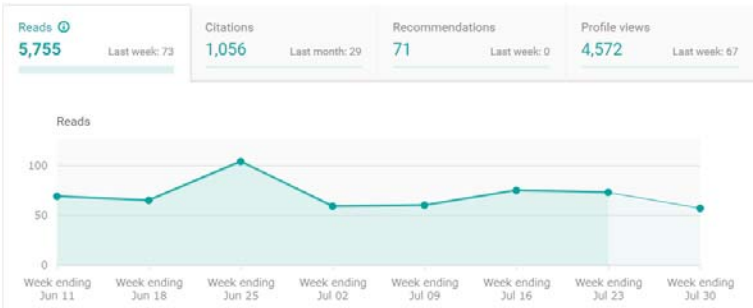
- 理想や目標は、人に馬鹿にされ揶揄されるくらいのものを掲げて、努力すれば、道は意外と開かれる
- しかし、もう5年ほど早く解決できればよかった。

- さて、これからどうしよう？
- 3冊目の決定打で、世界中の癌遺伝子診断で認められるか？
- 皆さんも、頑張ろう。

JAPAN DISCOVERY SUMMIT
EXPLORING DATA INSPIRING INNOVATION

ResearchGate





文献

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. Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data. Sequential Anal. (Editor's special invited paper) 30, 356-399.
3. Cox, D.R. (1958) “The regression analysis of binary sequences (with discussion).” J Roy Stat Soc B 20: 215-242.
4. 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
5. Edgar A (1945) The irises of the Gaspe Peninsula. Bulletin of the American Iris Society vol. 59: 2-5
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. Fisher, R. A. (1956). Statistical methods and statistical inference. Hafner Publishing Co.
9. Flury B, 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. (Shinmura, S. translated Japanese version)
15. Shinmura, S. (2014). “End of Discriminant Functions based on Variance-Covariance Matrices.” *ICORES*, 5-14.
16. Shinmura, S. (2015a). “The 95% confidence intervals of error rates and discriminant coefficients.” Statistics, Optimization and Information Computing, vol. 3, 66-78.
17. Shinmura, S. (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.
18. ISSN: 1865-0929, ISBN: 978-3-319-17508-9, DOI: 10.1007/978-3-319-17509-6.
19. Shinmura, S. (2015c). “A Trivial Linear Discriminant Function.” Statistics, Optimization, and Information Computing, Vol.3, December 2015, 322-335. DOI: 10.19139/soic.20151202.
20. Shinmura, S. (2015d). “The Discrimination of the microarray data (Ver. 1).” Research Gate (1), Oct. 28, 2015, 1-4.

jmp Statistical Discovery.™ From SAS.
Copyright © 2017 SAS Institute Inc. All rights reserved.

21. Shinmura, S. (2016a).“Matroska Feature Selection Method for Microarray Data.” Biotechno 2016, 1-6.
22. Shinmura, S. (2016b) “Discriminant Analysis of the Linear Separable Data -Japanese automobiles-.” Journal of Statistical Science and Application, vol4, No.07-08, 165-178. DOI: 10.17265/2328-224X/2016.0708.001.
23. Shinmura, S. (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.
24. Shinmura, S. (2016d). “The K-fold Cross Validation for Small Sample Method.” Data Analytic 2016, 1-6.
25. Shinmura, S. (2016f). The New Theory of Discriminant Analysis after R Fisher, Springer. DOI: 10.1007/978-981-10-2164-0
26. Shinmura, S. (2016g).From Cancer Gene Analysis to Cancer Gene Diagnosis, Amazon Kindle version.
27. 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
28. Simon N, Friedman J, Hastie T, Tibshirani R (2013). “A sparse-group lasso.” J. Comput. Graph. Statist, 22:231-245
29. Singh et al.(2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell: March 2002, 1/2: 203-209
30. 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

31. Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag.
32. 青嶋誠 (2016). 高次元の統計学. 21-2,5-15.
33. 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.
34. 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.
35. 新村秀一 (2004).『JMP活用 統計学とっておき勉強法』. 講談社.
36. 新村秀一(2007).『ExcelとLINGOで学ぶ数理計画法』. 日科技連出版.
37. 新村秀一 (2010).『最適線形判別関数』. 日科技連出版.
38. 新村秀一(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.

jmp Statistical Discovery.™ From SAS.
Copyright © 2017 SAS Institute Inc. All rights reserved.

39. 新村秀一(2011b). 『数理計画法による問題解決法』. 日科技連出版.

40. 新村秀一 (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.

41. 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.

42. 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.

43. 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、

44. 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.