高次元遺伝子データ解析理論 (Theory2) の完成報告 3

01202720　成蹊大学名誉教授　新村　秀一　　SHINMURA Shuichi

1. はじめに

RIP で 2 群判別すれば、データに一意に MNM が決まる。判別分析は、データに重要な 2 群の分類情報を付加し有効な情報を得る手法だ。誤判別例を省き Test 標本とすれば、「省かれたデータは LSD になり Fact3 という 4 つのデータ構造」を持ち、明確な分析結果で誤判別例が検討できる。今回児頭骨盤不均衡の乳児の出産を自然分娩(180 例)にするか帝王切開(60 例)にするかを事前に決めた CPD240 を再検討した。

2. 19 変数の CPD240 データ

出産に際して妊婦の骨盤と児頭が不均衡のため CPD 症候を示す場合、分娩法を出産前に決める必要がある。日本医科大学産婦人科の鈴村教授は、X 線写真から児頭の形状を厚紙に切り取り骨盤と比較し、この判断を簡易に行う「鈴村氏法」を提案した。筆者が統計分析で検討した。医学論文以外に、詳しい分析は、Springer [3]で紹介した。表 1 に示す 17 個の計測値と鈴村氏法の判定基準の X9=X7-X8 (VIF=21)と X12=X13-X14(VIF=1484)がある。この 2 変数のため多重共線が生じ、どの変数を省いて解決するかに悪戦苦闘した。多重共線性の良い対応法がなかったが、変数減少法と PCA を組合わせて解決した。

Table 1. The 19 variables and VIFs.

Var.	Description	VIF
X1	age of the pregnant woman	1.2
X2	the number of times delivery	1.3
X3	the number of the sacrum	1.1
X4	the utero-posterior distance at the pelvic inlet	24.6
X5	the utero-posterior distance at the wide pelvis	8.7
X6	the utero-posterior distance at the narrow pelvis	3.1
X7	the shortest Anteroposterior distance	57.0
X8	Biparietal fetal diameter	5.3
X10	the utero-posterior distance at the pelvic inlet	3.7
X11	biparietal diameter at the pelvic inlet	1.7
X13	the area at the pelvic inlet	1466
X14	the area of the fetal head	638
X15	the area at the bottom length of the uterus	1.4
X16	abdominal circumference	1.7
X17	external conjugate	1.6
X18	intertrochanteric diameter	1.6
X19	lateral conjugate	1.4

3. LINGO の 4 つの Program による分析結果

RIP で 2 例の誤判別例を Test 標本とし省いたデータは LSD になる。そして**高次元の遺伝子データという横長データ（n<p）**で見つけた LSD の 4 つの汎用的なデータ構造（Fact3）が利用できる。従来の研究では、新しい理論や手法の提案だけが重要視された。今回「**分析データの構造**」の新知見の応用である。これで判別理論が真に役に立つと考える。

Step1: 最初に **Program1** (RIP や H-SVM 等 5 種の LDF)の RIP で、正常例(N30)と帝王切開(C24)の誤判別例が分かった。これらを省いた **CPD238(238*19)**を作成する。ただし、X13 の係数も自然に 0 になった。これが Microarray を判別するだけで自然に n 個以下の変数が選択できる **Fact1** である。n 個以下の遺伝子を含む SM1 を自然に選ぶ FS で、SM1 省いた残りの遺伝子から順次 SM2 以降の SM に分割した。**Program3** は、この SM 分割を連続的に行い、最後に LSD でない(MNM＞0) ほぼ一組の Type2 の SM で終了する。他の研究が悪戦苦闘し「わずか一組の LSD でない遺伝子を選択し、多くの分類手法で ER が最小のデータと手法の組み合わせ」を発表している。出発点から間違っている理工学研究の大不祥事である。

Step2: Program3 と 4 で **CPD238** を SM と BGS に分割すれば、簡単に **19 変数の SM1** と、**14 変数の BGS1** と LSD でない残りの 5 変数(X1,X3,X4,X6,X13)に分割できた。BGS2 のロジスティック回帰の NM は 38 例(31/7)である。このため、**Program4** で BGS2 を求めるのに時間がかかり打切り MNM は不明である。

Step3:判別分析には 4 つの深刻な問題がある。**問題 4** は判別理論は推測統計学でないので、CPD238 を 10 回コピーし**擬似的な母集団である検証標本**とする。それを乱数で並べ替えて 10 分割し 10 組の学習標本とする。これは検証標本からの標本であり、統計学の「母集団と標本の関係」を満たす。一般に使われている **k 重 CV** は、この基本を満たさない間違った検証法だ。Theory2 では 10 組の検証標本の平均 ER (**M2**) が 0 の BGS を選ぶ（**Validation1**）。SM1 と X13 を省いた SM1 と BGS1 と BGS2 を **Program2** で検証すると、M2 は 2.8%、2.9%、**2.5%**と 11.8%であった。これは変数の多い 2 つの SM1 より 15 変数の BGS1 の方が、頑強であることを示す。一方、BGS2 は明らかに悪い。これは統計モデルの**ケチの原理**も成立し、**Theory2 が学際的な理論**である事が分かる。多くの医学研究は DNA の発現データで結果が出ないので、現在 RNA に注目している。しかし筆者は DNA の 169

の Microarray で良い結果を出したので、RNA の品質が良ければもっと良い結果が得られるだけだ。一方、DNA で成功しなかった研究者が RNA でも成功しない。これが学際的な Theory2 の結論である。

4. 5 つの検証法

「多次元空間では、乱数で作った 2 群でも LSD になる」という都市伝説がある。しかし論文や実際のデータは調べたが分からない。これは**普通のデータ(n>=p)と横長データ(n<p)の違い**を知らない無知からきたと考える。高次元遺伝子解析の「3 つの困難」はこの違いを知らないことから生れ、多くの研究が失敗した原因だ。なぜ高校数学で習う「**連立方程式の解を求める条件**」を横長データの研究に生かせなかったのか不思議だ。p 変数から n 変数を選ばなければ、連立方程式の解は求まらない。また LDF や重回帰の DF は p でなく n になることを知らないのは、MP や統計ソフトで実証経験に乏しいからである。

すなわち判別理論は、2 群が平均だけ異なる同じ正規分布という Fisher の仮説でなく、「**組合わせ理論**」で考えるべきだ。試験の合否判定と同じく高次元遺伝子データは簡単な判別である。しかし p が大きいと ${}_pC_n$ の多くの組み合わせが生じる。多くの LSD の SM や BGS に分割できても、多くの組み合わせの偶然で LSD になった偽の多変量の癌遺伝子を省く必要がある。そこで **Validation1** は M2＝0 になる BGS を選ぶ。しかし CPD では 2.5%という結果になった。医学の計測機器は品質が良いことが多いので、他の分野ではもっと悪い選択基準になると考える。分野毎にこの基準は変わる。**Validation2** は RatioSV（2＊100／判別スコアの範囲）で、2 群の線形分離度を調べるがあまり役に立たないようだ。

図 1 は BGS1 の PCA の散布図（Validation3）である。Prin1 で青の正常と赤の帝王切開が分かれていない。発現データでは 2 群が分かれるデータが多かった。SM や BGS は全て LSD であり、散布図の累積寄与率は 91.8%と大きいが、残りの 12 主成分の僅か 8.2%のばらつきが加わって LSD になるようだ。このような LSD を省くことにした。

「癌は不均質な病気」と言われている。Fact3 はこれに反して、癌と正常が均質に分かれるので不思議であった。正常細胞が癌細胞になると**エピジェネチックな変化**が起きて、発現データの質と量が変わるので素人ながらこれが原因と考えていた。すなわち**発現量は人類が出会った最高品質の計測値**だ。一方 2 次元の大きなバラツキで誤判別される症例を取り除くことで均質になるので、散布図の誤判別例が不均質を表すと考えた（**Validation3**）。NIH の Liver3 は肝癌（181）と正常(176)と、正常と癌の 2 群で、ほぼ同数で、症例数が一番大きな最良のデータである。しかし誤分類例が非常に多いので、これを省いて各 30 例の中規模なデータに作り代えると 3 検証を満たし、多くの重要な BGS が見つかった。これは発現量に限って、大標本より中規模の標本の方が良いことを示す。研究費が潤沢だと、多くの症例数を集める。そしてその中に医学研究に興味のある研究対象も多く含まれる。しかし、最初は癌とコントロール群の正常は、典型症例に限った方が良いと分かった。

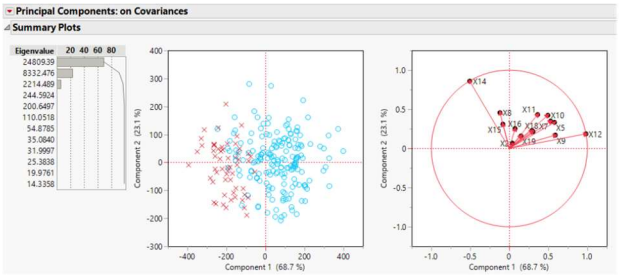

図 1　BGS1 の分散共分散行列による散布図

癌研究では、生物学的知識に基づいて顕微鏡で 400 個以上の発癌遺伝子を見つけた。**Validation4 は、**多くの M2=0 の重要な BGS に含まれる発癌遺伝子で医学的な多変量の発癌遺伝子の特徴が分かる。**Validation5** は本庶先生が普及に尽力された **Genome Cohort** による検証と生存時間解析である。癌研究の問題は、データ解析の利用を考えず、クラスター分析で症例と遺伝子の関係を調べて生存時間解析を行う研究が多い。5 つの検証を、順を追って行うべきである。

5. おわりに

Theory2 で得た成果は、次の適用が可能である。

1） DNA や RNA を用いた他の医学診断。あるいは動物などの遺伝子診断。無償で品質の良いデータを探せば、向学心のある高校生でも成功する。
2） しかし普通の LSD でないデータに適用した本研究は、医学診断に限らず多くの人の福音になる。

学会や大学や企業が ZOOM 会議を準備すれば、筆者は 2 時間程度で直ぐに利用できる講習会の準備をする事は可能だ。日本から最先端の遺伝子データ解析の研究が続出することを期待したい。

2015 年まで OR や統計で年 4 回の発表と研究費捻出に日本語の出版に拘ったため、Google Scholar の引用数は皆無だった。それが 15 の発現データの報告書を Research Gate にあげると一気に 299 引用され現在 1135 に増えたが、RG では Theory1 だけ引用され Theory2 の引用が少ないことが分る。