連載講座

ORと統計の融合

数理計画法を用いた最適線形判別関数(3)
—多重共線性のある医学データの分析—

新村　秀一

1. 対象データ

ここでは，新生児が児頭骨盤不均衡（Cephalo-Pelvic-Disproportion，略してCPD）か否かを判定する日本医科大学鈴村名誉教授が開発した鈴村氏法の妥当性を調べる目的で集められた医学データを用いる．鈴村氏法とは，CPDの疑いのある新生児に対して，妊婦のレントゲン写真から児頭の輪郭を紙に写し取り，それをレントゲン写真上で動かし，児頭と骨盤の関係を検討する．そして，自然分娩（180例）や帝王切開（60例）などのいずれの処置方法を選択するかを決める方法である．

表1のデータは，これを統計的に実証するために日本医科大学第1病院の鈴村・松木・武井らによって集められた．松木（1978）は，本データを用いて後で紹介するDOC1とDOC2と呼ばれる2つの判別モデルを導いているが，このモデルは鈴村氏法の妥当性を示す変数を含んでいない．

表1　説明変数

X1	：年齢
X2	：経産回数
X3	：仙骨の数
X4	：入口部前後径
X5	：かつ部前後径
X6	：狭部前後径
X7	：最短前後径
X8	：児大横径
X9	：X7－X8
X10	：入口部前後経
X11	：入口部横径
X12	：X13－X14
X13	：入口部面積
X14	：児頭面積
X15	：子宮底
X16	：腹位
X17	：外結合線
X18	：大転子間径
X19	：側結合線

本データを用いた理由は，三宅・新村（1979）で，探索的な最適判別関数の評価に用いたが，計算時間の制約で6変数までしかアプローチできなかったことである．今回は，整数計画法による新しいアルゴリズムと進歩した情報処理環境による20年後の再挑戦である．

また，アイリスデータと異なり，変数の数が19変数と多く，多重共線性を含み，データ件数も240件（すなわち，240整数変数）と少なくないからである．すなわち，現実のデータで問題になる，多重共線性を考慮したモデル選択を考えている．

X9とX12は，それぞれ児頭と骨盤に関する2つの計測値の差であり，鈴村氏法の妥当性を示すと期待される変数である．しかし，このため説明変数間に多重共線性が含まれることになる．また妊婦であるため，X1（年齢）やX2（経産回数）とX3（仙骨の数）は，正の打ち切りのある整数値である．このため，従来の判別手法の理論的前提である多次元正規分布という条件を満たしているとは思われない．

2. 解析結果

2.1 基本的な分析

(1) 分析の枠組み

表2は，SASのRSQUAREプロセジャーに，今回提案する手法の結果を付け加えたものである．

SASのRSQUAREプロセジャーは，説明変数の全ての組み合わせの回帰モデルを評価している．すなわち，19個の説明変数から，$(2^{19}-1)$個すなわち約50万個の回帰モデルを検討したことになる．そして，デフォルトでは，説明変数の総数と同じ数の上位19個のモデルを説明変数が1個から18個の場合に出力す

しんむら　しゅういち
成蹊大学 経済学部
〒180-8633 武蔵野市吉祥寺北町3-3-1

表2　CPDデータの総括表

P	Rank	Type	IP	LP	FP	QP	F5	Q5	R-square	C(p)	AIC	説明変数
1	1	FBfb	19	21	23	22	26	27	0.521	22.3	-568.4	X12
2	2	FBfb	13	16	17	20	32	27	0.559	4.2	-585.9	X9 X12
3	3	FBfb	12	19	19	22	22	20	0.565	3.0	-587.2	X9 X12 X18
4	1	Ffb	10	19	17	18	22	21	0.572	1.3	-588.9	X9 X12 X15 X18
4	3	B	11	22	20	20	23	23	0.568	3.2	-587.0	X9 X12 X13 X18
5	1	Ff	10	22	17	18	24	17	0.575	1.6	-588.7	X9 X12 X15 X17 X18
5	2	b	8	16	17	16	24	24	0.574	2.0	-588.3	X2 X9 X12 X15 X18
5	3	B	11	21	19	31	22	39	0.573	2.5	-587.7	X9 X12-X14 X18
6	1	B	9	20	16	32	22	36	0.578	1.9	-588.5	X9 X12-X15 X18
6	2	b	7	14	17	15	22	24	0.577	2.3	-588.1	X1 X2 X9 X12 X15 X18
6	3	Ff	8	18	16	16	24	25	0.577	2.5	-587.9	X2 X9 X12 X15 X17 X18
6	*	DOC1	13	18	20	20	22	23	0.565	8.3	-587.3	X5 X9 X13 X14 X17 X18
6	*	DOC2	11	24	19	22	22	20	0.564	8.4	-587.2	X7 X9 X13 X14 X17 X18
7	1	B	9	20	15	30	21	35	0.582	1.8	-588.7	X9 X12-X15 X17 X18
7	2	Ffb	7	16	16	15	23	17	0.580	2.7	-587.8	X1 X2 X9 X12 X15 X17 X18
8	1	F	6	18	14	9	19	16	0.584	3.0	-587.6	X1 X2 X7 X9 X12 X15 X17 X18
8	2	B	8	17	17	27	21	35	0.584	3.0	-587.5	X1 X9 X12-X15 X17 X18
8	5	fb	6	18	14	9	19	16	0.583	3.2	-587.4	X1 X2 X8 X9 X12 X15 X17 X18
9	1	B	6	16	16	23	19	29	0.586	3.6	-587.1	X1 X2 X9 X12-X15 X17 X18
9	2	F	4	10	15	9	20	16	0.586	3.6	-587.1	X1 X2 X5 X7 X9 X12 X15 X17 X18
9	3	fb	4	10	14	9	20	15	0.585	4.0	-586.6	X1 X2 X5 X8 X9 X12 X15 X17 X18
10	1	B	6	17	15	24	18	26	0.588	4.5	-586.2	X1 X2 X7 X9 X12-X15 X17 X18
10	6	F	4	10	14	8	20	15	0.587	5.1	-585.7	X1 X2 X5 X7 X9 X12 X15 X17-X19
10	*	fb	3	10	14	10	20	15	0.586	5.7	-585.2	X1 X2 X5 X8 X9 X12 X15 X17-X19
11	1	B	4	10	13	22	20	28	0.590	5.4	-585.5	X1 X2 X5 X7 X9 X12-X15 X17 X18
11	*	F	4	8	13	9	19	13	0.582	12.0	-586.9	X1 X2 X5 X7 X9 X12 X13 X15 X17-X19
11	*	fb	3	8	13	11	20	13	0.587	7.4	-583.5	X1 X2 X5 X8 X9 X12 X13 X15 X17-X19
12	1	FB	4	9	13	21	19	26	0.591	6.9	-584.0	X1 X2 X5 X7 X9 X12-X15 X17-X19
12	*	fb	3	9	13	11	22	15	0.588	9.1	-581.9	X1 X2 X5 X8 X9 X12 X13 X15-X19
13	1	FB	3	9	13	17	20	23	0.592	8.6	-582.3	X1 X2 X4 X5 X7 X9 X12-X15 X17-X19
13	*	fb	3	7	15	9	23	13	0.588	11.0	-580.0	X1 X2 X5 X8 X9 X11-X13 X15-X19
14	1	FB	3	8	15	16	20	22	0.592	10.4	-580.6	X1 X2 X4 X5 X7 X9 X11-X15 X17-X19
14	*	fb	2	6	15	10	23	10	0.588	13.0	-578.0	X1-X3 X5 X8 X9 X11-X13 X15-X19
15	1	FB	3	6	15	17	20	20	0.592	12.2	-578.7	X1 X2 X4 X5 X7 X9 X11-X19
15	*	fb	2	5	15	7	23	11	0.588	15.0	-576.0	X1-X3 X5 X8 -X13 X15-X19
16	1	FB	3	6	16	21	21	23	0.593	14.1	-576.8	X1 X2 X4 X5 X7-X9 X11-X19
16	*	fb	2	4	15	7	23	10	0.588	17.0	-574.0	X1-X3 X5 X6 X8-X13 X15-X19
17	1	FB	2	5	16	19	22	21	0.593	16.0	-574.9	X1 X2 X4 X5 X7-X19
18	1	FB	2	5	15	17	22	21	0.593	18.0	-573.0	X1 X2 X4-X19
19	1	FB	2	3	15	16	22	20	0.593	20.0	-571.0	X1-X19

る．19個の場合は，1個のモデルの統計量を出力する．すなわち，(18×19+1)個のモデルを出力した．

最初の数字(p)は，回帰分析に用いた説明変数の個数を表わしている．p=1から18のそれぞれで，決定係数の大きい上位19個のモデルを出力したが，その中から重要なもののみを表示してある．

最右端の変数は，選ばれた説明変数を表わす．2番目（Rank）は，説明変数が同じ個数のモデルの中で，該当するモデルの決定係数が何番目に良いかを示す順位である．ただし20番目以降はRSQUAREで出力しなかったので，STEPWISEプロセジャーで別途計算し，順位は不明なので“*”で示してある．

(2) 基本系列

逐次変数選択法では，FinとFoutの打ち切り基準を用いて，変数選択を行っている．しかし，このような停止則は，収束に問題が起きず計算時間に影響のない逐次選択法では，実際の応用局面では意味がないと考える．

そこで，停止則を考えないで1変数からフルモデルまでのモデル系列を検討すべきと考えている．その意味で，停止則を用いないで得られた変数増加法と変数減少法のモデル系列を，上昇基本系列と下降基本系列と呼ぶことにする．

3番目（Type）の記号は，Fが19変数の上昇基本系列（19 F），Bは下降基本系列（19 B）を表わしている．fは，19変数から多重共線性のある（X 4，X 7，X 14）の3変数を省いた16変数の上昇基本系列（16 f）である．多重共線性の解消方法については，新村・三宅（1983）と新村（1996）を参照してほしい．bは，多重共線性のない16変数の下降基本系列（16 b）である．

DOC 1とDOC 2は，本データを集めた松木(1978）が選んだモデルである．

すなわち，19変数の変数増加法では，最初の1変数モデル（X 12）が選ばれ，次に2変数モデル（X 9，X 12），そして3変数モデル（X 9，X 12，X 18）というようにモデルが逐次選ばれ，最終的に19変数のモデル（X 1-X 19）になる．(-)は，X 1からX 19までの19個の変数を表す省略記号である．

19変数の変数減少法は，（X 1-X 19）から始まり，p＝18の行に示すように変数X 3がモデルから掃き出され（X 1，X 2，X 4-X 19）が選ばれる．p＝17ではX 6が掃き出され（X 1，X 2，X 4，X 5，X 7-X 19）が選ばれている．そして，最終的に1変数モデル（X 12）になる．

多重共線性を省いた16変数の変数増加法では，pが1から7まではFと一致しているので，19変数と同じモデルが選ばれていることが分かる．16変数の変数減少法では，pが16変数から4変数までは，bはBと一致せず，3変数ではじめて同じモデルを選んでいる．

(3) 誤分類数

4列目から9列目の数字は，各判別関数の誤分類数である．

IPとLPは，シカゴ大学のL. Schrage教授が社長を務めるLINDO社の提供する数理計画法ソフトWhat'sBest!を用いた．

FPとQPは，データ件数に比例した事前確率0.75と0.25を用いたFisherの線形判別関数と2次判別関数，そしてF 5とQ 5は事前確率が0.5である場合の判別分析の誤分類数である．SASのDISCRIMプロセジャーで計算した．

表中のモデルの全てが，χ^2検定で棄却された．しかし，必ずしも2次判別関数の成績（誤分類数）が良いわけではない．すなわち，2次判別関数の現実への適用が疑問視される．

(4) モデルの検討

その後は，モデル選択に用いられる，決定係数，Mallow'sのCp統計量，赤池のAIC基準である．

赤池のAIC基準は，4変数のモデル（X 9，X 12，X 15，X 18）が良いことを示唆している．このモデルは，タイプがFfbであるので，19変数の変数増加法と16変数の変数増加法と減少法の両方で選ばれていることが分かる．このモデルには，鈴村氏法の良さを示すために付け加えられたX 9とX 12が含まれている．

これに対して，松木が選んだ6変数のモデルには，X 9は含まれるがX 12は含まれていない．多重共線性に対する配慮や，基本系列という考え方がなかったためであろう．

2.2 基本統計量と平均値の差の検定

(1) 基本統計量

表3は，表2のIP列からQ 5列までの6個の誤分類数の基本統計量である．データ件数は，基本系列で

表3 誤分類数の基本統計量

	平均値	中央値	最小値	最大値	25%点	75%点	範囲	四分位範囲
IP	6.250	5	2	19	3.0	9.0	17	6
LP	13.000	12	3	24	8.0	18.0	19	10
FP	15.775	15	13	23	14.0	17.0	10	3
QP	16.875	17	7	32	10.0	21.5	25	11.5
F5	21.650	22	18	32	20.0	23.0	14	3
Q5	21.325	21	10	39	15.5	25.5	29	10
LPIP	6.750	6	1	13	4.0	10.0	12	6
FPIP	9.525	9	4	14	8.0	11.5	10	3.5
QPIP	10.625	8	3	23	6.5	16.0	20	9.5
F5IP	15.400	16	7	21	13.0	17.5	14	4.5
Q5IP	15.075	12	7	28	10.0	19.5	21	9.5

表4 t-検定

	IP	LP	FP	QP	F5	Q5
IP	0.000	-12.660	-24.490	-11.797	-27.659	-15.442
LP	12.660	0.000	-3.475	-3.967	-8.920	-7.940
FP	24.490	3.475	0.000	-1.151	-16.075	-5.345
QP	11.797	3.967	1.151	0.000	-4.349	-10.149
F5	27.659	8.920	16.075	4.349	0.000	0.283
Q5	15.442	7.940	5.345	10.149	-0.283	0.000

選ばれた38件とDOC 1とDOC 2の計40件である．Q5の後のLPIPは，LPからIPを引いた差を表す．多くの分析結果があれば，それらを用いてさらに統計分析し，分析結果を統計的に評価することができる．

平均値と中央値を比較すると，IP<LP<FP<QP<Q5<F5の順に大きくなる．Q5とF5の成績が悪いことから，このデータでは，事前確率はデータ件数に比例させた方がよいことが分かる．これ以降，F5とQ5は参考程度に考えることにする．

最大値は，IP<FP<LP<QP＝F5<Q5の順に大きくなる．LP-OLDFの中には，誤分類数が線形判別関数より悪くなるものもあるようだ．やはりQPは，他の3手法より最大値が大きいことが分かる．

範囲は，FP<F5<IP<LP<QP<Q5の順に大きくなる．四分位範囲は，FP＝F5<IP<LP<Q5<QPの順に大きくなる．線形判別関数の誤分類数のバラツキが一番小さく，それを除けばIP-OLDF，LP-OLDF，2次判別関数の順は分布の代表値である平均値や中央値と変わらないことが分かる．

(2) 平均値の差の検定

表4は，2変数の平均値に差があるか否かのt検定を表すt値を行列表記したものである．網掛けしてあるFPとQP，およびF5とQ5以外は，5%で棄却される．すなわち，FP（F5）の平均値はQP（Q5）より小さいが，統計的には差が認められない．

結局，平均値の差の検定からも，IPが一番判別成績が良いことが分かった．

3. 上昇・下降基本系列での評価

(1) 上昇基本系列

図1は，横軸がモデルに含まれる説明変数の数pを，縦軸は誤分類数を示す．表2のFで表される19個の上昇基本系列のIPとLPとFPとQPの誤分類数をまとめたものである．F5とQ5は，見にくくなるので省いた．この図から，アイリスデータと異なり，一般的に次のことが分かる．

- IP-OLDFは，1変数から19変数と増えるに従い，誤分類数は単調に減少している．しかも，他の3つに比べ成績は明らかに良い．
- この単調減少性は，一般的に次のように説明できる．p変数で誤分類数を最小とする判別関数が得られたとする．このモデルに，任意の残りの変数を加えたモデルで，追加した変数の判別係数を0にすれば，(p+1)変数モデルにおいてp変数の誤分類数は最低保証される．すなわち，単調減少になる．
- この単調減少性は，誤分類数で表され，決定係数の単調増加性よりも分かりやすい．

この他，このデータに依存した次の傾向が読み取れる．

- LP-OLDFは，1変数から8変数までで，誤分類数の変動が激しい．12変数以上で単調に減少している．
- 線形判別関数は，変数が増えても単調に減少しないで，6変数以上では誤分類数は16個から13個の間で変動している．判別分析における問題点は，このように誤分類数が決定係数のように単調減少でないことである．このため，モデルを逐次変数選択法で決定したり，特定のモデルでFisherの線形判別関数と2次判別関数を誤分類数で比較しても確定的なことは言えないことが分かる．
- 2次判別関数は，1変数から8変数までで減少し，8変数から11変数まではほぼ同じ値をとり，11変数以上になると急に誤分類数は増えている．表2を見れば分かる通り，11変数モデルには（X 12，X 13）が含まれているが，12変数モデルになってX 14が入って多重共線性の影響が出たためと思われる．この2次判別関数の振る舞いは，これまで指摘されていない知見である．

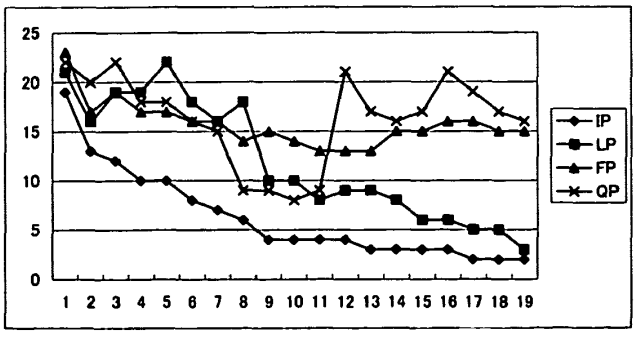

図1 上昇基本系列上の誤分類数

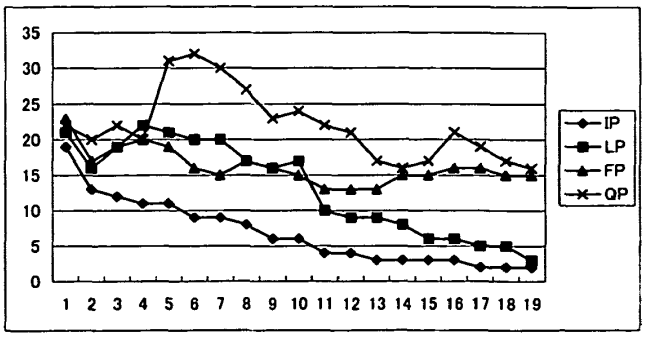

図2 下降基本系列上の誤分類数

(2) 下降基本系列

図2は，下降基本系列の誤分類数をまとめたものである．モデルは，19変数から1変数と降順で選ばれるが，説明は上昇基本系列と同じ昇順で行うことにする．上昇基本系列と同じく，次の結果が得られた．

・IP-OLDFは，1変数から19変数と増えるに従い，誤分類数は単調に減少している．しかも，他の3つに比べて成績は明らかに良い．
・LP-OLDFは，1変数から9変数まで変動しながら減少している．10変数以上で単調に減少している．
・線形判別関数は，変数が増えても単調に減少しないで，6変数以上では誤分類数は17個から13個の間で変動している．
・2次判別関数は，1変数から6変数まで増加し，6変数以上で変動しながら減少している．表2の19変数の下降基本系列を調べると分かる通り，5変数モデルまでは多重共線性が含まれており，4変数モデルでは多重共線性が解消されているため，減少していくのであろう．

すなわち，2次判別関数では図1と図2の増加と減少のパターンが逆であるが，いずれも問題であることが分かる．他の3手法では上昇基本系列でも下降基本系列でも同じことがいえる．

(3) 多重共線性の解消

図3は，多重共線性を解消した16変数の上昇基本系列である．2次判別関数が，LP-OLDFと似たような傾向になった．他の3手法は，19変数と同じ傾向である．

図4は，16変数の下降基本系列である．やはり，2次判別関数が，LP-OLDFと似たような傾向になった．他の3手法は，19変数と同じ傾向である．

すなわち，IP-OLDFと線形判別関数とLP-OLDFは，多重共線性に影響されず，次のような特徴がある．

・IP-OLDFは，単調減少する．
・線形判別関数は，単調減少しない．
・LP-OLDFは，変動しながら，pが大きくなってから減少していく．

2次判別関数は，多重共線性の影響を一番強く受け，多重共線性がなくなってはじめてLP-OLDFのように変動しながら減少していく．

4. 多重共線性の評価

(1) 基本系列での比較

図5は，表2のIP-OLDFの誤分類数を各基本系列毎にまとめたものである．19Bで表される19変数の下降基本系列の成績が一番悪く，16bで表される16変数の下降基本系列の成績が良いことが分かる．残りの2つは，重なる部分もあるがその中間にある．

経験的には，「逐次変数選択法では，変数増加法よりも変数減少法の成績が一般的に良いことが多い．」と言われている．しかし19変数のデータのように変数の数が比較的多く，多重共線性のある場合には，変数減少法では多重共線性のある変数を最初からモデルに含むので，それが掃き出されるまで（誤分類数で表される）成績が悪く，変数増加法ではそれをできるだけ最後まで含まないようにさける傾向があるので成績が良いのではなかろうか．それに対して，多重共線関係のない16変数では，従来の知見と同じく変数減少法の成績が良くなったと考えられる．

このような指摘は，従来のモデル選択に用いられる

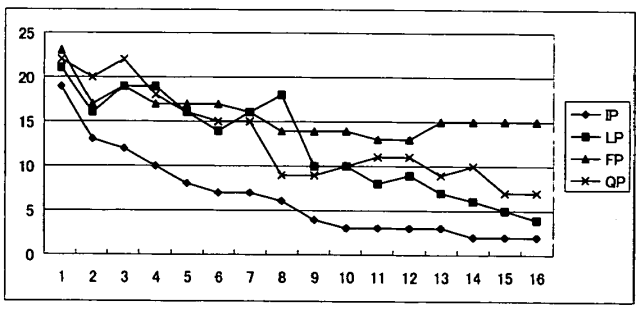

図3 上昇基本系列（16 vars.）

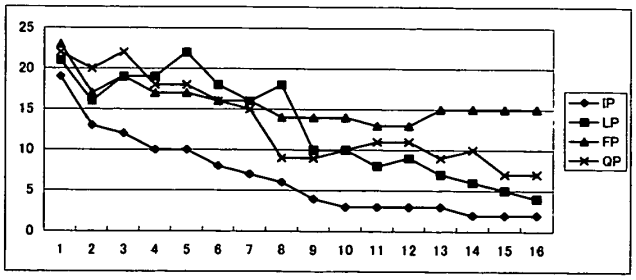

図4 下降基本系列（16 vars.）

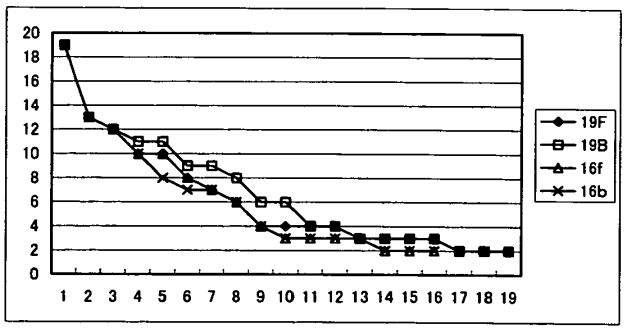

図5 各系列でのIP判別分析の誤分類数比較

表5　AIC最小モデルにおける判別係数

	X9	X12	X15	X18	定数項
IP	-5.363E-2	-6.177E-3	1.341E-2	-1.249E-2	1
LP	-3.228E-3	-6.791E-4	8.071E-4	-3.823E-3	1
F(0.75, 0.25)	-9.349E-3	-1.539E-3	1.602E-3	-3.740E-3	1
F(0.5, 0.5)	-8.592E-3	-1.414E-3	1.472E-3	-3.437E-3	1

表6　AIC最小モデルにおける誤分類数

	誤分類数	ケース
IP	10	38, 58, 113, 174, 204, 206, 207, 218, 220, 227
LP	19	15, 30, 38, 58, 63, 90, 103, 113, 130, 174, 204, 207, 214, 215, 216, 220, 227, 236, 239
F (0.75, 0.25)	17	15, 30, 38, 47, 58, 63, 90, 103, 113, 130, 174, 204, 206, 207, 210, 220, 227
Q (0.75, 0.25)	18	15, 30, 38, 58, 63, 90, 103, 113, 130, 174, 204, 206, 207, 214, 218, 220, 227, 232
F(0.5, 0.5)	22	4, 15, 22, 30, 34, 38, 47, 51, 58, 63, 87, 88, 90, 95, 103, 113, 129, 130, 141, 158, 174, 204
Q(0.5, 0.5)	18	15, 22, 30, 34, 38, 47, 58, 63, 88, 90, 95, 103, 113, 129, 130, 174, 204, 220

決定係数などでは指摘できなかった．基本系列上でモデルを内部標本の誤分類数という尺度で眺めることで明らかになった．

(2)　医師が選んだモデル

松木が選んだ6変数モデルの誤分類数は，11個と13個と悪い．モデル選択の技術が未熟な時代に行われたため，判断を誤ったのだろう．

(3)　AIC最小モデル

AIC最小化基準では，4変数モデルを選ぶことが示唆される．

表5は，この4変数モデルにおける判別係数を示す．判別係数の符号は一致しており，あまり極端な違いはないようだ．表6は，誤分類されたサンプル番号を示す．IP-OLDFで誤分類されたケースが，他の判別関数のそれに含まれることが望ましい．IP-OLDFで誤分類された10ケースは，Fisherでは218番目のケースを除いて含まれる．2次判別関数の18例には，完全に含まれる．この事実は，最適化によって従来の結果とかけ離れたものが選ばれたわけではないことが分かる．

一方，事前確率が0.5の場合は，Fisherでは5例，2次判別関数では4例が含まれないことが分かる．事前確率を用いて，データに比例したほうが，IP-OLDFの内部標本の誤分類数を最小化することに対応しているためかもしれない．

5.　決定係数と誤分類数の食い違い

多重共線性を省いた16変数の下降基本系列（16b）の誤分類数が，他の系列に比べて良いことが分かった．しかし，表2に示したように16変数の下降基本系列の決定係数あるいはそれによる順位は，例えば19Bと比べて決してよくない点である．1変数から3変数では，両方とも一致し1位である．4変数と5変数のみで，16bの順位が良い．6変数以上では，19Bの成績が良い．19位までしか出力しなかったので，それより悪いものは＊にしてある．

すなわち，決定係数の良いモデル（19B）が，誤分類数が少ないことに必ずしも対応していない．あるいは，これまでモデルの良さを表すために用いられてきた決定係数が良くなくても，誤分類数が少ない16bのようなモデルの系列があることを示すことができた．

この理由として次のことが指摘できる．

・これまで回帰診断などで，多重共線性の問題点は色々指摘されてきた．今回は，誤分類数という分かりやすい尺度で，判別分析においても問題点を指摘できた．すなわち，多重共線性のあるモデルでは，決定係数はモデルの良さを表す良い統計量になっていないと考えられる．

・各基本系列上での決定係数とAICを比較すると，多重共線関係のある6変数以上では19Bが良いことが分かる．4変数と5変数では，16bが良くなっている．1変数から3変数では，4つの系列は一致している．

すなわち，多重共線性のあるデータでは，決定係数の値が好ましくないモデルの中にも，今回のような誤分類数が少ないものがあるということがCPDデータから示せた．

今後，多重共線性があったり多次元正規分布から大きく逸脱する他のデータにも適用し，このことがある程度普遍的なことか否か検証する必要がある．

6.　誤分類数の評価

(1)　相関係数

表3の誤分類数と決定係数の値を相関分析すると，

IPの誤分類数は他と強い相関があることが分かる．決定係数は，IPやFPと強い負の相関がある．

(2) 回帰分析による判別成績の評価

FP, LP, QPをIPで単回帰分析を行うと次の回帰式が得られた．

FP＝12.843＋0.469 IP　(r＝0.840)

LP＝4.776＋1.316 IP　(r＝0.861)

QP＝11.172＋0.913 IP　(r＝0.543)

決定係数をIPで回帰すると両変数とも単調増加と減少を示すので，−0.921と強い負の相関がある．

R_Squuare＝0.6−0.003 IP

(3) 4手法の比較

図6は，線形判別関数（実線），LP-OLDF（1点鎖線），2次判別関数（破線）の誤分類数をIP-OLDFで回帰した回帰直線を重ねがきしたものである．横軸はIPの値であり，縦軸は各判別手法の誤分類数の予測値である．

実線の線形判別関数は，IP＝0で約13，IP＝20で22の予測値である．すなわち，誤分類数の変化幅が一番少ないことを表わしている．

仮に，IPをIPで回帰すると，原点とIP＝20で縦軸が20（実線の2目盛り下）を結ぶ直線になる．すなわち，IP-OLDFが1変数から19変数の全てのモデルで3手法より優れていることが分かる．

一方，線形判別関数と2次判別関数はIP＝4（表2から9変数モデルに対応）で交差し，線形判別関数とIP-OLDFはIP＝10（表2からAIC最小の4変数モデルに対応）で交差し，2次判別関数とIP-OLDFはIP＝16（表2から1変数モデルに対応）で交差していることが分かる．

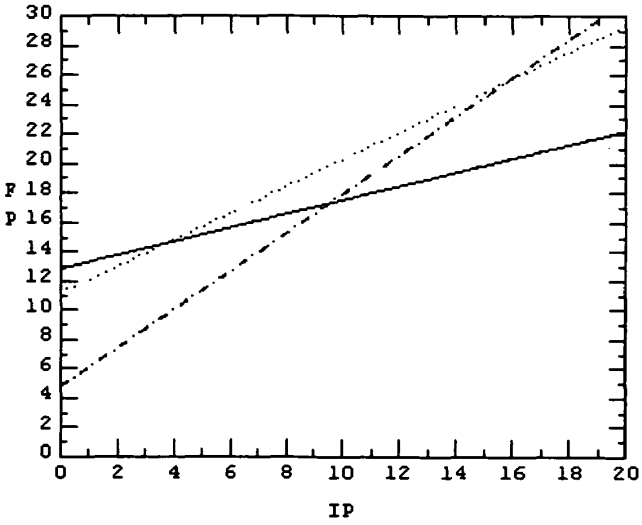

図6　FP，QP，LPの回帰式

すなわち，IPの定義域[10, 16]は，1変数から4変数のモデルに対応している．この区間では，IP-OLDF＜線形判別関数＜LP-OLDF＜2次判別関数の順に誤分類数が増える．

平均値で得られた順位のIP-OLDF＜LP-OLDF＜線形判別関数＜2次判別関数は，IPの区間[4,10]に対応し，4変数以上9変数までのモデルに対応している．

10変数以上では，IP-OLDF＜LP-OLDF＜2次判別関数＜線形判別関数の順に誤分類数が増える．

以上から，多重共線性がある場合は，3手法の優劣は説明変数の数によって異なってくることが分かる．ただし，AIC最小モデルに対応するIP＝10前後の結果を採用すべきであろう．

7. まとめと今後の課題

今回，数理計画法を用いたIP-OLDFとLP-OLDFを現実のデータを用いて既存の手法と比較評価した．

説明変数が少なく，比較的Fisherの線形判別関数の理論的前提を満たしていると思われるアイリスデータでは，それほど優位性は認められなかった．

これに対し，医学データでは，19個と説明変数の数が多く多重共線性があり，離散変量などの種々の変数が混在しているため多次元正規分布でないことは一目瞭然である．このようなデータでは，IP-OLDFは従来の手法に対して，基本系列上で比較評価すると明らかに良い結果が得られた．

すなわち，IP-OLDFの誤分類数は説明変数に対して単調減少であるのに対して，Fisherの線形判別関数は減少しなかった．また，多重共線性の影響で，決定係数の良いモデル系列よりも，決定係数が悪いにもかかわらずより誤分類数の少ないモデル系列があることを例証できた．

このことは，多重共線性やデータが多次元正規分布でないため（正規性からの乖離）と考えられ，それらの存在を視覚的に捕らえることができる指標と考えても良いだろう．

また，今回の誤分類数と決定係数を相関分析すると，IP-OLDFのそれは他と相関が高かった．これまで，事前確率やリスクの設定で誤分類数は変化しモデル評価に困難な状況を引き起こした．そこで，内部標本に対し一意に決まるIP-OLDFの誤分類数を説明変数として，他の誤分類数の単回帰分析を行った．同じ単調減少性を示す決定係数の結果が良いのは当然として，

Fisherの誤分類数のほうが2次判別関数のそれと比較して，IP-OLDFの誤分類数で良く回帰できた．

さらに，基本系列上の全てのモデルは，2群の分散共分散のχ^2検定で1%で棄却された．これまでの理論の教えるところに従えば，Fisherの線形判別関数より2次判別関数を用いた方がよいことになるが，実際のデータでは却って誤分類数が悪い例が多く存在していることが分かった．

しかし，IP-OLDFの単調減少性は，「けちの原理(principal of parsimony)」で表されるモデル決定には直接的に役立たない．本論文では，AIC最小化基準を用いてモデル決定を行い，そのモデルの判別係数や誤分類されたケースを検討したが，IP-OLDFには特に問題はなかった．

今後は人工データや他の分野のデータを用いて検討することで，IP-OLDFがどのような例で成績が特に良いのかあるいは悪いのか，そしてモデル選択に関する具体的な知見を探す必要があろう．

この新しい手法によって，多変量正規分布を前提にした従来の判別分析では分からなかった，新しい知見が得られることを期待したい．

参考文献

[1] 松木玄篤：数理的解析によるCPDの判定，日本産科婦人科学会雑誌，30-12，1727-1736，1978.

[2] 三宅章彦・新村秀一：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，18-1，15-20，1979.

[3] 新村秀一，三宅章彦：重回帰分析と判別分析のモデル決定(1)—19変数をもつC. P. D. データの多重共線性の解消—，医療情報学，3-3，107-123，1983.

[4] 新村秀一：重回帰分析と判別分析のモデル決定(2)—19変数をもつC. P. D. データのモデル決定—，成蹊大学経済学部論集，27-1，180-203，1996.

[5] 新村秀一：数理計画法を用いた最適線形判別関数，計算機統計学，11-2，93-105，1998