日本オペレーションズ・リサーチ学会
2005年春季研究発表会

整数計画法による判別分析の新世紀　－1000スイスフラン偽札紙幣の分析－

1202720　　成蹊大学　新村秀一 SHINMURA Shuichi

1980年代以降、数理計画法を用いた重回帰分析と判別分析のモデルが数多く提案されているが、それらが統計分野で注目されたことはない。その理由は、これらの分野の研究者の以下のような認識誤りである。

- 数理計画法の多くのモデルは、数理計画法で始めてアプローチできたものが多い。これに対して、統計モデルを数理計画法で記述できたとしても、それらが既存の統計手法に比べて、優れた点を主張できないことには意味がないということを理解していない点である。
- 新村は、整数計画法を用いた新しい判別関数 IP-OLDF を提案した。本手法は、データ空間でなく判別係数の空間で、標本誤分類数は最小化する基準を用いているが、この方法によって初めて統計的線形判別関数に対して、優位性を主張できる。

1　分析データと統計分析

判別分析の評価データとしては、フィッシャーのアイリスデータが有名である。ただし、説明変数が4個と物足りない。これに対して少し難易度の高い、本物と偽者の1000スイスフラン紙幣それぞれ100枚の6個の測定値からなる「スイス銀行紙幣データ」がある（Flury,B. & Rieduyl.H.(1988)）。本データを用いて、IP-OLDF と Fisher の線形判別関数の比較評価を行うことで、IP-OLDF の優位性を示したい。

6個の説明変数を用いて逐次変数選択を行った。変数増加法では、「diagonal、bottom、top、right、left」の順にモデルに取り込まれ Fin 基準で停止した。変数減少法でも、Fout 基準でこの5変数モデルが選ばれた。しかも AIC は－781.1 で最小である。珍しいことであるが、変数増加法も変数減少法も Fin 基準と Fout 基準と AIC 最小化基準で同じ5変数モデルを選んだ事がわかる。主成分回帰分析では、4変数モデルを選んだ。

2．　総当り法による IP-OLDF と Fisher の判別分析の比較

2・1　統計と IP-OLDF の分析結果

表1は、重回帰分析の説明変数の「総当り法」の結果に、IP-OLDF と Fisher の線形判別分析の結果を加えたものである。分析を見やすくするために、「length、left、right、bottom、top、diagonal」は「モデル」列で X1 から X6 で表わしてある。「数」は説明変数の数である。同じ数のモデルでは、3列目の「R2乗」の値で降順に並べてある。「Fisher」列は、通常の線形判別関数の誤分類数である。1から100の間でばらついていることがわかる。以上は、統計ソフトの JMP で計算した結果である。

この後は、数理計画法ソフトの What's Best!で計算した結果である。「1」と「－1」列は、IP-OLDF で定数項を「+1 と-1」の両方で計算した最小誤分類数である。(x1,x4)の「1」列の値 14(17)は、判別境界点上に9個のケースがあり、IP-OLDF の解 14 を修正した内点解が 17 であることを示す。「OLDF」列は、「1」と「－1」の小さい方の値を表示してある。

最後の「差」列は、「Fisher」列と「OLDF」列の差である。すなわち、IP-OLDF の誤分類数が Fisher の線形判別関数のそれよりも、どれだけ少ないかを表わしている。この値のヒストグラムと分位点から、Q1 と Q2 が1で、Q3 が4である。すなわち、全体で63個のモデルのうち、38個のモデルが1以下、9個が2と3、そして16個が4以上の差があることが分かる。

2・2　単調減少性

IP-OLDF の誤分類数は、元のモデルに変数を追加した場合、必ず誤分類数が同じか少なくなるという単調減少性を示す。その理由は追加した変数の係数を0としたモデルは、元のモデルと同じであり、誤分類数も同じになる。IP-OLDF は標本の最小誤分類数を求めているので、追加したモデルの誤分類数は、元のモデルと等しいかそれ以下になることが分る。

例えば、2変数モデル（X4,X6）は、X4 と X6 の誤分類数2と16の最小値 MIN(2,16)=2 に等しいか小さくなる。実際には0であり、2例改善されているので枠で囲んである。（X4,X5）も同様に MIN(16,48)=16 より小さい3であり、1変数では判別力がなかったのに、2変数で判別力が高まっている。(X3,X5)、(X2,X5)、(X1,X5)、(X1,X2) も誤分類数が少なくなっているので枠で囲んである。

3変数モデル（X3,X4,X5）では、部分モデルの誤分類数の最小値と等しいかそれ以下である。しかし、1変数モデルは、2変数モデルのサブモデルであるので調べる必要はない。すなわち2変数の部分モデルの最小値は、MIN(3,29,16)=3 であるが、実際は2と改善されている。この他、(X2,X4,X5)、(X1,X4,X5)、(X1,X3,X5)、(X2,X3,X5)、(X1,X2,X5)、(X1,X2,X3) も改善されている。しかし、4変数以上のモデルは、判別成績の悪い（X1,X2、X3、X4）と（X1,X2,X3,X5）以外は3変数の部分モデルよりも改善されていない。すなわち、IP-OLDF

では次の2点が結論付けられる。

・1変数モデル（X6）の誤分類数は2であり、2変数モデル（X4,X6）の誤分類数は0になる。すなわち、「bottomとdiagonal」の2変数で偽札と真札を完全に判別できることが分る。この点は、従来の回帰分析や判別手法では、より説明変数の数の大きなモデルを選ぶ傾向にあるが、IP-OLDFではより説明変数の小さいモデルを選ぶことができる。

・63個のモデル全体で考えても、4変数以上のモデルを用いる必要がないことが表1で部分モデルよりも判別成績が良くなるモデルが2例以外は無い事から断定できる。

・お札のような品質管理が厳しく、変動係数が小さなデータでは、計測誤差を除けば、外部標本が内部標本と大きく異なる事は考えられない。すなわち、IP-OLDFの2変数モデルで予測しても、判別成績は良いものと考えられる。

表1　「総当り法」の結果とIP-OLDFとFisherの線形判別分析の結果(一部)

モデル	数	R2乗	Cp	ＡＩＣ	Fisher	1	-1	OLDF	差
X1,X2,X3,X4,X5,X6	6	0.92	7.00	-779.4	1	0	0	0	1
X2,X3,X4,X5,X6	5	0.92	5.32	-781.1	1	0	0	0	1
X3,X4,X5,X6	4	0.92	10.26	-776.0	1	0	0	0	1
X4,X5,X6	3	0.92	10.66	-775.6	1	8	0	0	1
X4,X6	2	0.88	107.00	-698.5	3	86	0	0	3
X4,X5	2	0.84	203.72		8	3	70	3	5
X1,X4	2	0.60	831.84		18	14(17)	16	16	2
X3,X5	2	0.51	1067.58		29	25(29)	47	29	0
X2,X5	2	0.45	1204.03		33	30(32)	47	32	1
X1,X3	2	0.42	1275.02		36	34(43)	31(33)	33	3
X1,X5	2	0.38	1380.23		45	42	3(36)2	36	9
X2,X3	2	0.35	1457.03		44	34(43)	72	43	1
X1,X2	2	0.34	1485.72		47	41(44)	47	46	3
X6	1	0.81	292.02	-603.85	2	100	1(2)	2	0
X4	1	0.60	833.49		17	14(16)	100	16	1
X5	1	0.36	1427.77		49	42(48)	100	48	1
X3	1	0.34	1475.23		43	34(43)	99	43	0
X2	1	0.25	1726.54		53	42(48)	97	48	5
X1	1	0.03	2270.94		100	97	67(77)	77	23

3．まとめ

判別データとして統計学で有名な「偽札データ」を用いて、従来の統計手法とIP-OLDFとの比較を行なった。

主成分回帰分析の変数選択法は4変数モデルを、元の説明変数による回帰分析では5変数モデルを選んだ。しかし、従来の統計手法だけでは、はっきりと決着をつけることができないが、IP-OLDFは2変数モデルで十分であり、しかも誤分類数が0と良い事がわかった。また、パーティションは、IP-OLDFと同じくbottomとdiagonalの2変数を用いて分類できたが、誤分類数は2と判別分析の3よりは良かった。本データは、もともと誤分類数の少ない扱いやすいデータのため、2変数「bottomとdiagonal」を用いたIP-OLDFと、パーティションとFisherの線形判別関数では、誤分類数が0と2と3の僅差でしかないが、この傾向は誤分類数の大きなデータでもいえそうだ。

今後、重回帰分析や判別分析がIP-OLDFより多くの説明変数を選ぶ傾向にあることを、一般的に検討したい。また、63個のモデル全体で考えれば、FisherはIP-OLDFよりも11%ほど多くなることが分かった。ただし、これは、16個ある4例以上差のあるモデルに影響されていると考えられる。

文献

1）Flury,B. & Rieduyl,H.(1988). Multivariate Statistics: A Practical Approach. Cambridge University Press.［田端吉雄(1900).多変量解析とその応用.現代数学社］.

2）新村秀一(2004).JMP活用統計学とっておき勉強法.講談社.

3）John Sall,Lee Creighton, & Ann Lehman（新村監修)(2004). JMPを用いた統計およびデータ分析入門（第3版）. SAS Institute Japan㈱.