分散共分散行列に基づく判別分析の終焉

新村　秀一†

†成蹊大学経済学部 〒180-8633 東京都武蔵野市吉祥寺北町 3-3-1

1. はじめに

Fisher[4]が計算機能力の乏しい時代に分散共分散行列に基づいてFisherの線形判別関数(LDF)を提案し，その後に分散共分散行列で定式化できる 2 次判別関数(QDF)，マハラノビスの汎距離による多群判別や品質管理の MT(マハラノビス田口)理論[36]といった判別分析の中核となる判別手法が開発された.筆者は1971年に大学を卒業後，大阪府立成人病センターで「心電図自動診断解析システム」の診断論理を判別分析で研究したが，プロジェクトリーダーの野村裕医師の開発した「フィードバックのかかった枝分かれ論理」にかなわなかった[34]．その反省に立って，医学における正常と異常の 2 群判別は，1) 正常群からある計測値が連続的に大きく（あるいは小さく）なることで異常群が決定される．すなわち正常群を地球，異常群を山脈と考えれば，地平線にあたる判別境界面上に異常群が一番多く，典型例は山の頂点であるとする「地球モデル」を考えた[19]．そしてBayesの定理を用いたスペクトル診断を提案した[16]．また正常群と複数の異常群の多群判別は最良の説明変数の組み合わせが異なるので，各異常群と正常群の複数の2群判別を行うべきと結論した [10,17]．その後，異常群に属する確率 p を考え，そのオッヅ比の対数を線形回帰するロジスティック回帰が提案された．本手法は医学診断や品質管理における正常と異常は，判別する 2 群を Fisher の仮説を満たす正規分布と考えることは問題であることを示していると考えた．

三宅と新村[9]は，Warmack & Gonzalez[38]の論文に触発され，最小誤分類数(Minimum Number of Misclassifications, MNM) に基づく最適線形判別関数(Optimal Linear Discriminant Function, OLDF)を提案した．しかし，ヒューリスティック手法のため，研究を発展させることができなかった．筆者は日本に統計ソフトの SAS[11-12,20]と数理計画法ソフトの LINDO や LINGO[13-14]を日本に紹介し，大学や企業に普及を図るとともに研究を行ってきた．そこで学位申請のテーマとして整数計画法(IP)を用いて，学習標本の誤分類数を直接最適化するIP-OLDFを研究することは統計と数理計画法を融合する最適なテーマと考えた．そして新しい判別分析の知見が得られた[21-23]．しかし，データが一般位置[6]にない場合に正しい MNM を求めないことがあるので，MNM を求める改定 IP-OLDF を開発した[27]．本研究では 100 重交差検証法で，改定 IP-OLDF をソフト・マージン最大化 SVM (S-SVM)，LDF，ロジスティック回帰と比較し，検証標本で得られた平均誤分類確率最小モデルで，一番成績が良いことを示す．

2. 比較する判別関数

2.1　統計的判別関数

Fisher は判別される 2 群の分散比の最大化から式(1)に示す LDF を提案した．これは制約式なしの非線形計画法(NLP)になる．

$$\text{MIN}=\ b^t\ (x_{m1}-x_{m2})\ (x_{m1}-x_{m2})^t b/b'\ \Sigma b\ ;\quad (1)$$

ただし，m_1 と m_2 は群 1 と群 2 の平均，

Σ:プールした分散共分散行列,b は判別係数．その後，2 群が「Fisher の仮説」を満たせば式(1)で得られる LDF が，容易に 2 群を表す正規分布の確率密度関数の対数尤度比で式(2)のように明示的に定式化されることが分かった．これが統計ソフトに取り入れられ分散共分散行列による判別関数が広く普及したと考えられる．

$$\text{LDF}:\ f(x)=\{x-(m_1+m_2)/2\}^t\Sigma^{-1}(m_1-m_2) \qquad (2)$$

2 群の分散共分散行列が等しくない場合は式(3)の 2 次判別関数(QDF)が明示的に定式化される．

$$\text{QDF}\ :\ f(x)=x^t(\Sigma_2^{-1}-\ \Sigma_1^{-1})x/2 +(m_1^t\Sigma_1^{-1}-m_2^{\ t}\Sigma_2^{-1})x+c \qquad (3)$$

また式(4)のマハラノビスの汎距離を用いて多群判別や品質管理の MT 理論が考えられた．

$$D=\text{SQRT}\ ((x-m)^t\Sigma^{-1}(x-m)) \qquad (4)$$

式(2)から式(4)のように分散共分散行列の逆行列を計算するだけで LDF や QDF や多群判別や 1 クラス判別関数が定式化され，数理計画法を用いた式(1)よりも統計ソフトとして実現しやすく，理論展開しやすいことは一目瞭然である．しかし分散共分散行列の逆行列を計算する必要があり，ある説明変数が一定値をとる場合はランク落ちして判別関数が求まらない問題がある．またこれらの判別関数は MNM=0 のデータを認識できないこともすでに分かっている[33,35]．

これに対して p を基準と考える群に属する確率と考え，賭け事で利用されるオッヅ比の対数 log(p/(1-p))が説明変数の回帰式で表されると考えて式(5)のロジスティック回帰が提案された．

一般的にいってLDFやQDFよりロジスティック回帰の誤分類数が少ないことが経験的に知られている．これは判別される2群のデータに合わせてロジスティック曲線を求めているからと考えられる．また，ある計測値が一方的に大きくなる(あるいは小さくなる）と確率は0から1まで大きくなるので，医学診断（や株や不動産の格づけ）などで正常群から離れていくほど異常群に属する確率が1まで高まっていくという常識に合致する．また，Firth[3]は，収束計算が不安定になると誤分類数が 0 になることを指摘している．しかし，一部の例外でJMPのロジスティック回帰はMNM=0のデータを認識できないことを報告している[33]．これは判別超平面上にケースが来た場合，改定IP-OLDF以外の判別関数はそれを正しくいずれの群に判別するかができない問題（判別分析の未解決問題）のためと考えられる．

$$P=1/(1+\exp(-b_0-b_1x_1\cdots-b_k x_k)) \qquad (5)$$

2.2 数理計画法による線形判別関数

数理計画法は，制約条件のあるなしにかかわらず，関数の最大／最小値問題を扱う学問である[28,32]．このため，重回帰分析や判別分析を数理計画法で定式化する研究が行われてきた．しかし重回帰分析は，誤差の平方和の最小化モデルがQP で，絶対値の和の最小化モデルが線形計画法(LP)で定式化されることが知られているが[13,28]，判別分析ほど多くの研究は行われていない．憶測の域を出ないが，重回帰分析では推測統計学的な知見が得られるので数理計画法による研究は不利であるのに対して，判別分析はクラスター分析に近い記述的な手法であり有利とみなされたためと考えられる．

2.2.1 Stamの嘆き

Stam[15]は，数理計画法による判別関数を分類問題として総括した論文の 5.1 節(5.1 Why have statisticians rarely used Lp-norm methods?)で，これらの研究成果が統計ユーザーに使われていないことを指摘しその理由を分析している．彼の分析は大学の同僚の統計研究家との議論の集約であると述べているが，間違いではないが本質的でないと考える．数理計画法による判別分析の研究は,Fisherなどの統計分野で確立された手法の後追いの研究であり，それらと比較して判別結果が優れていることと新知見をつけ加える必要性を認識していなかためと考える．

2.2.2 SVM

Stamと相前後してVapnik[37]はSVMを提案し，今日でも広く使われている．その理由は，種々の実証研究で成果を出していること，式(6)のハード・マージン最大化 SVM(H-SVM)から式(7)のソフト・マージン最大化 SVM(S-SVM)，そして非線形判別関数である Kernel SVM と異なった視点で魅力的なモデルを展開し，汎化能力(Generalization Ability)という統計ユーザーの要求に合う提案をしている[1]．多くの研究者はKernel SVMに注目しているが，筆者はH-SVMで判別分析を線形分離可能なデータ(MNM=0)の判別を判別分析の出発点とした点が重要と考えている．

$$\text{MIN}=||b||^2/2;\ y_i*(x_i{}^t b+b_0)>=1; \qquad (6)$$
$$y_i=1/-1 \in \text{群 1/群 2}$$

しかし現実のデータはMNM=0であることは稀である．その場合，幾つかのケースが SV の反対側にくる事を認め，その距離の和（Σe_i）を最小化する．そして式(7)のように S-SVM は，目的関数でマージン最大化とΣe_iの最小化という2つの異なった最適化基準をペナルティcと呼ぶ重みで単目的化している [26]．問題は，H-SVM は MNM=0 のデータ以外に適用できず，S-SVM は理論的に MNM=0 の判別の保証がないことと，MNM=0 のデータを見つけることが難しかったために，MNM=0 の判別分析の研究が進まなかった点である．

$$\text{MIN}=||b||^2/+c\Sigma e_i;\ y_i*(x_i{}^t b+b_0)>=1-e_i; \qquad (7)$$

2.2.3 最適線形判別関数[31]

MNM 基準による線形判別関数を最適線形判別関数ということにする．確率分布に基づく統計アプローチでは実現不可能で，どのケースが誤分類されるか否かを直接 0/1 の整数変数 e_i を用いて(混合)整数計画法(IP)で初めて定式化できる．判別規則は，$y_i f(x_i)>0$ であれば x_i を群 1/群 2 に正しく判別し，$y_i f<0$ であれば誤判別するので判別係数を定数倍しても変わらない．そこで線形判別関数 $f(x)=b_0+b_1*x_1+\cdots+b_p*x_p$ の定数項が $b_0\neq 0$ であれば，b_0 で割っても判別規則は影響を受けず定数項を 1 に固定でき($f(x)=b_1*x_1+\cdots+b_p*x_p+1$)、式(8)のように定式化したものが IP-OLDF[21]である．

$$\text{MIN}=\Sigma e_i;\quad y_i(x_i{}^t b+1)>=-\ e_i\ ; \qquad (8)$$

定数項を 1 に固定したことで，式(8)は p 次元線形判別係数の空間で定式化される．n 個のケースで作られた制約式を等号制約にしてn個の線形超平面が定義される．各線形超平面は，p 次元の判別係数空間を$y_i f(x_i)>0$の＋半平面と$y_i f(x_i)<0$の－半平面に分割し，空間は有限個の凸体に分割される．各凸体の内点は，n 個の半平面のうちの－半平面の側にある数が誤分類数になる．凸体の内点の誤分類数がMNMになる凸体を最適凸体と呼ぶ．IP-OLDFは，少なくとも p 個の制約式が等式

制約に拘束される凸体の頂点を求める．e_iは 0/1 の整数変数であり $y_i f(x_i)>=0$ であれば e_i =0 で，$y_i f(x_i)<0$ であれば e_i =1 として $y_i*f(x_i)>=-1$ という制約式が選択される．目的関数はこの e_i の個数を最小化しているが，凸体の頂点や辺に対応した $f(x_i)=0$ になるケースは正しく判別されたものとして扱うので判別分析の未解決問題が生じる．このためデータが一般位置にある場合(頂点が丁度 p 個の線形式の交点)は正しい最適凸体の頂点すなわち MNM を求めるが，一般位置にない場合(頂点が（p＋1）個以上の線形式の交点)は正しい最適凸体の頂点を求めないことがある．また y_i =1/-1 と群 1/群 2 の組み合わせは事前に人間が決めることができずデータが決めるので（この事実はあまり知られていない)，実際は定数項を 1 と 0 と-1 の 3 種類で計算する必要がある．パターン認識の研究では定数項を 1 に固定しないので，定数項を含む(p+1)次元の空間で考えている[6]．この場合はｎ個の線形超平面はすべて原点で交差し，判別係数と誤分類数(NM)の関係が明確でなくなる．その意味で IP-OLDF は実際には利用できないが，次の 3 つの判別分析の新知見が分かった．1) 判別係数と誤分類数の関係が明らかになる．2) 最適線形判別関数は最適凸体の内点を選べばよく，この場合判別分析の未解決問題が解決できる．3) MNM には単調減少性[1]があるので MNM_p =0 であれば $MNM_{(p+1)}=0$ になる．

そこで最適凸体の内点を直接求める式(9)の改定 IP-OLDF を定式化した．b_0は線形判別関数の定数項で自由変数であり，右辺の定数項に 1 を加えて判別超平面の代わりに SV で判別することにした．M は BigM 定数と呼ぶ大きな正の整数値であり，$y_i f(x_i)>=1$ であればケース x_i は e_i =0 として SV で正しく判別され，SV で判別されない x_i は e_i =1 として $y_i f(x_i)>=1-M$ という制約式を選ぶ．目的関数は SV で判別されないケース数の和（Σe_i）を最小化しているので，$|y_i f(x_i)|<1$ になるケースがない場合のみ MNM に等しくなる．M=10000 のような大きな正の値を用いれば，SV で判別されない全てのケースは $y_i f(x_i)=-9999$ という SV の代わりになる超平面に引きずられるので $|y_i f(x_i)|<1$ の範囲にケースがないことを期待している．この場合は，$f(x)\neq 0$ になるケースはないので，IP-OLDF の定義した最適凸体の内点に対応していることが分かる．

$$MIN=\Sigma e_i;\quad y_i(x_i' b+b_0)>=1-M*e_i\ ; \qquad (9)$$

e_iを非負の実数にしたものが改定 LP-OLDF であり，LP で定式化でき計算速度が速い．式(7)で c を大きくした S-SVM とほぼ似た結果になる．$y_i f(x_i)>=1$ であればケース x_i は SV で正しく判別されるが，それ以外は $y_i f(x_i)<1$ であるために e_i =1/M であれば $f(x_i)=0$ になる． 数理計画法を用いた判別関数は，不等式制約の幾つかの制約式を等式制約に固定することで解が求まるので，「判別分析の未解決問題」を回避できない．Fisher が LDF の検証に用いたアイリスデータ[24]の 15 個の判別モデルのうち，2 個の改定 LP-OLDF の誤分類数が改定 IP-OLDF の MNM より小さいが，MNM は線形判別関数の誤分類数の下限値になるので，少なくとも 2 個以上でこの問題が生じていることが分かる．しかしこの問題は，$f(x_i)=0$ になるケース数を直接調べればすむ話である．もし k（>0）個あれば，少なくとも誤分類数はこの値だけ増える可能性がある．

3．分析データと k-重交差検証法

3.1 線形分離可能なデータ

これまで判別分析の研究で，MNM=0 のデータに関する問題の検討が少ないので本研究ではこれに焦点を絞って取り上げる．MNM=0 のデータにスイス銀行紙幣データ[5]がある．本データは，真札と偽札各 100 枚の 6 個の計測値である．IP-OLDF で 2 群判別すると，2 変数（X4,X6）で MNM=0 であり，MNM の単調減少性からこの 2 変数を含む 16 個すべての判別関数が MNM=0 になる．このように意図せず MNM=0 のデータを見つけることは難しい．それに対し，試験の各設問の得点を説明変数とし，合計得点で合否判定を行うことを考えれば，1) 誰もが容易に入手し検証できる．2) 誤分類数が 0 の自明な判別関数が容易に分る．3) 合格最低点を変えることで易しい試験から国家試験のような難しい試験の検討が行える．そこで，2010 年から始めた 3 年間の筆者の成蹊大学経済学部の 1 年次生を対象とした「統計入門（約 130 名)」の中間試験と期末試験データを用いる．必修科目のため，実際の合否判定は 10%点を合格最低点としているが，研究のため 50%点と 90%点も検討する．試験は 10 択 100 問のマークセンス試験である．3 年間の実績で 21 点から 99 点の広い範囲に散らばっている．合否判定を各 3 水準で行い，小問 100 問を説明変数（1/0 の値）とする 18 個の 2 群判別と，100 問を 4 個の大問に分けてそれを説明変数とする 18 個の 2 群判別を行った．例えば大問 4 問の得点を説明変数（T1 から T4）として合計得

[1] p 変数の MNM を MNM_p とし，それに 1 変数追加したものを $MNM_{(p+1)}$ とすれば，追加した変数を a_p ＝0 とすれば（$MNM_p>=MNM_{(p+1)}$)の関係がある．

点が 50 点以上を合格とする合否判定は，線形判別関数を f=T1+T2+ T3+T4-50 として，f≧0 なら合格，f<0 なら不合格とすれば NM=0 で合否判定できる．この場合，説明変数で判別規則が明確に定義できるので f =0 すなわち 50 点をとる学生は合格群に正しく判別できる．しかしスイス銀行紙幣データのような判別では，6 個の計測値から判別規則は決められないので，これが「判別分析の未解決問題」を生む．任意のデータでは改定 IP-OLDF だけが理論的にこの問題の影響を受けない．ロジスティック回帰は，Firth の研究[3]でロジスティック回帰係数が収束計算で不安定になれば NM=0 であるとしているが，筆者の実証研究で改定 IP-OLDF で MNM=0 になることが分かっている場合でもロジスティック回帰は NM=0 にならない場合がある[32]．これはユーザーがどちらを Positive にするかを指定することで，JMP が判別超平面上のケースを Positive に無条件に判別しているためである．この対応は，研究論文でも説明もなく群 1 の側に等号を含める例が多く問題である．この事実は，線形判別関数の判別超平面上にケースが来ることは連続変数であっても稀有でないことを示す．以上から，線形判別関数 f(x)が例えば $|f(x_i)| \leqq 10^{-6}$ で 0 判定している場合，この値を誤分類数の横に表示するのが一番分かり易い解決策と考える．

3.2 小標本のための k－重交差検証法

近年，各種計測器や POS データあるいは Web 上から大量のデータが得られるようになった．しかし，医学診断をはじめ多くの研究では当初小標本のデータしか得られず，その研究結果を踏まえて数多くの研究の中から目的とする研究に収斂させていく必要がある．そのために小標本のために考えられた LOO 法はこれまで多くの貢献をしてきた[7]．しかし，Bootstrap 法[2]を援用して，小標本から大標本（疑似的な母集団）を生成し，それに k-重交差検証法が適用できる．これを用いて LOO 法に代わるモデル決定や，複数の判別手法を個別の判別モデル毎に比較検討できる．さらに，学習標本と検証標本で得られた判別結果から判別係数や誤分類確率の 95%信頼区間の推定に利用できる[31]．

分析したい小標本として，Fisher のアイリスデータ[4]，銀行紙幣データ[5]，CPD データ[9,18]，学生の成績データ[25]を用いて，次のような大標本を作成し 100 重交差検証法を行い，表 1 の結果を得た[31]．

1) n 個の小標本のケースに 1 から n のケース番号を振る．
2) 一様乱数[0,1]の区間を n 等分し，ケース番号の 1 から n に対応させ復元抽出する．
3) 100*n 回一様乱数を発生させて，対応するケースを 100*n 個復元抽出する．そして，n 個ずつ 100 組の部分標本に分割し，100 重交差検証法に用いる．大標本の 100 重交差検証法では，1 組の部分標本を検証標本に用い，残りの 99 組の部分標本を学習標本に用いる．しかし小標本から作られた大標本の場合，取り出した 1 組を学習標本とし，残りの 99 組を検証標本に用いる方が自然である．99 組を学習標本にする理由は見当たらない．

アイリスデータは説明変数が 4 個あり 15 個，銀行紙幣データは 6 変数で 63 個，CPD データは 19 個あるため逐次変数選択法で選んだ 26 個，学生の成績データは 5 変数で 31 個の判別モデルの計 135 個の判別モデルで評価した．LDF とロジスティック回帰は JMP で 100 個の誤分類数から平均誤分類確率を計算した．2010 年当時は LINGO で改定 IP-OLDF を実行するのに時間がかかるので，MNM の推定値を高速で求める改定 IPLP-OLDF[30]を用いて平均誤分類確率を計算した．そして，これらの差の範囲を表 1 に示す．LDF の学習標本では 135 個のモデル全てで，LDF は改定 IPLP-OLDF より平均誤分類確率は大きかった．検証標本ではアイリスデータ（括弧内の数字），銀行データで 10 個，学生データで 3 個の 15 個だけで LDF が良かった．ロジスティック回帰は学習標本で 3 個，検証標本では 33 個で改定 IPLP-OLDF より良かった[31]．

表 1　LDF とロジスティック回帰の平均誤分類確率と改定 IPLP-OLFDF の差

	LDF-改定 IPLP		ロジスティック回帰 - 改定 IPLP	
手法(135)	学習 (0)	検証 (15)	学習 (3)	検証 (33)
アイリス(15)	[0.55,5.23]	[-0.6(2),2.36]	[0.59,5.31]	[-0.84(2),1.85]
銀行(63)	[0,5.32]	[-0.33(10),3.45]	[0,5.40]	[-0.3(24),3.64]
学生(31)	[1.46,8.61]	[-1,29(3),7.11]	[-2.12(3),6.48]	[-2.89(7),5.59]
CPD(26)	[3.05,7.28]	[2.21,6.15]]	[0,13,3.43]	[0.29,1.74]

最初，k=10 で試みたが判別係数や誤分類数の 95%信頼区間を求めるには最低でも k=100 が望ましいと考えるようになった．本研究では，LDF，ロジスティック回帰と S-SVM を改定 IP-OLDF と比

較する．ここ5年ほどで分析に用いているLINGOの整数計画法の速度が100倍以上改善され，統計手法の100重交差検証法より速くなったためである．またBootstrap法で大標本を作成する前回の方法は，操作手順が面倒である．そこで次のように誰もが簡単に利用できるように改善した．

1) JMPで小標本を100個結合し，100*n個のケース数をもつ一つの大標本にする．

2) これに一様乱数の変数値を付け加える．

3) 一様乱数の値で昇順に並べ替えて，n件ずつ100組の部分標本を表す1から100の部分標本番号を与える．

4) 100個の部分標本から順次1組を取り出し学習標本とし，100*n件の大標本全体を検証標本に用いる．検証標本は，検討したい小標本を100倍しただけなので，分析したい小標本の特徴とよく対応している.たとえばケース数が100倍なので，平均値の標準誤差は1/10になり将来の研究に役立つと考える．一方，学習標本は小標本の一部になる．通常の母集団と標本の関係を考えると，今回の方がこの関係をよく表している．前回の方法では，検証標本に学習標本のケースが含まれない可能性を排除できないし，検証標本が全て異なることになる．本研究では検証標本は母集団の役割に近づけ，学習標本はその部分集合と考えている．

4. 合否判定による100重交差検証法

MNM=0なデータとして，スイス銀行紙幣データがある．6個の計測値のうち（X4,X6）の2変数でMNM=0であり,この2変数を含む16個のモデルがMNM=0であることがIP-OLDFの研究で分かった[29]．しかしMNM=0のデータを探すことは難しいが，試験の合否判定を設問の得点を説明変数として判別すれば，自明な判別関数がえられる．そこで筆者が行った2012年の「統計入門」の中間試験のデータを用いる[35]．10択100問の試験をT1（基礎統計量の概念，29点），T2（基礎統計量の計算，12点），T3（正規分布と偏差値の解釈，19点），T4（JMPの解釈，40点）の大問4問に分けた．この4個の得点を説明変数とし，合計得点の10%点(37点)，50%点(63点)，90%点(78点)の3水準で合否判定を行う.10%点では大問2問，50%点と90%点では大問4問でMNM=0である．4個の説明変数の判別モデルは全部で15個あるが，1変数の判別は意味がないので2変数以上の11個で検討する．MNMの単調減少性から，10%点の合否判定では4個の判別モデルがMNM=0で，7個がMNM=0でない.50%点と90%点では1個だけがMNM=0で10個がMNM=0でない．これによって，MNM=0を含む判別分析の問題が検討できる．SVMはH-SVMでMNM=0なデータの判別を判別分析の出発点にしたことは革新的である．しかし，SVMを含めこれまで判別分析はMNM=0のデータの実証研究を行ってこなかった．判別分析の重要な応用分野であるパターン認識では，MNM=0のデータの判別が重要である．このため，今回MNM=0が自明な試験の合否判定データを用いることは適切と考える．試験の合否判定データは，多くの人が入手し容易に検証でき，合格水準を種々に変えることで2群のケース数の種々の割合の違いによる判別への影響を検討できるなどの利点がある[8].

4.1 10%水準の100重交差検証法

表2は100重交差検証法の結果である．1列目は各判別手法の下にモデル番号を示す．2列目は説明変数の数を表す．3列から5列は学習標本の100個の誤分類確率(誤分類数を124で割ったもの)の最小値と最大値そして平均誤分類確率でMIN,MAX,MEANに添え字の1を付けた．6列から8列は検証標本の誤分類確率(誤分類数を12,400で割ったもの)の最小値，最大値と平均値に添え字の2を付けた.検証標本の平均誤分類確率(MEAN2)が最少のモデルを100重交差検証法が選ぶモデルと考える．

改定IP-OLDF（表の「改定IP」）は100重交差検証法の学習標本で元データと同じく4個の判別モデルでMNM=0であるが，検証標本ではSN=2,6だけが0になった．この理由は，学習標本は元データの一部しか含まないため種々の判別超平面になり，元データのケースを全て含む検証標本をMNM=0で判別できないためである.「差」は「MEAN2」と「MEAN1」との差である．判別分析では見かけの誤分類確率のMEAN1に比べてMEAN2がどれだけ悪くなるかで評価されいずれも0である．これは改定IP-OLDFの汎化能力が高いことを示す．一般的に，改定IP-OLDFは学習データで過学習し，検証標本で判別の予測結果は悪くなると考えられたが，実際は異なっていた．ロジスティック回帰はSN=1を選んだ．差は0.77と悪くないが，SN=8の差の0.35が最小値である.SVMはSN=1を選び，差も0.81%と最小である．LDFはSN=6を選んだが差はSN=11の方が最小である．しかし差の値の0.37はロジスティック回帰の0.8とSVMの0.81に比べて小さいので，LDFの予測性能が良いと判断するのは問題である．単にLDFのMEAN1の見かけの平均誤分類確率が悪い結果である．これは，各手法のMEAN1とMEAN2を改定IP-OLDFの対応す

る平均誤分類確率との差を表す「MEAN1 差」と「MEAN2 差」と比較すれば一目瞭然である．ロジスティック回帰の「MEAN2 差」の範囲は[0.39,1.62]であり,最大 1.62%だけ改定 IP-OLDF より悪い．SVM は[0.4,1.19]であり最大 1.19%悪い．これに対して，LDF は[6.23,10.55]であり 2 変数以上の全てのモデルが改定 IP-OLDF より 6.23%以上悪く，ロジスティック回帰や SVM に比べて劣っていることが分かる．MNM=0 の 4 モデルでは 9.91%以上悪いが，残りの 7 モデルでも 6.23%以上悪いことが分かる．また「MEAN1 差」の範囲は[7.98, 11.34]であり，見かけの平均誤分類確率を他と比較するだけで悪いことが分かる．

表 2.　10%水準の 100 重交差検証法による誤分類確率

改定 IP	P	MIN1	MAX1	MEAN1	MIN2	MAX2	MEAN2	差		
1	4	0	0	0	0	1.61	0.07	0.07		
2	3	0	0	0	0	0	0	0		
3	3	0	0	0	0	2.42	0.03	0.03		
4	3	0	2.42	0.79	1.61	4.84	2.44	1.65		
5	3	0	5.65	2.25	3.23	8.06	4.64	2.39		
6	2	0	0	0	0	0	0	0		
7	2	0	4.03	1.78	2.42	6.45	3.40	1.62		
8	2	0	5.65	2.28	2.42	7.26	3.14	0.85		
9	2	1.61	8.87	4.88	5.65	10.5	6.63	1.75		
10	2	1.61	9.68	4.52	4.84	10.5	6.42	1.90		
11	2	1.61	7.26	4.94	6.45	12.1	7.21	2.27	改定 IP-OLDF との差	
Logi	P	MIN1	MAX1	MEAN1	MIN2	MAX2	MEAN2	差	MEAN1 差	MEAN2 差
1	4	0	0	0	0	4.84	0.77	0.77	0	0.7
2	3	0	0	0	0	3.23	1.09	1.09	0	1.09
3	3	0	0	0	0	4.84	0.85	0.85	0	0.81
4	3	0	5.65	1.59	1.60	4.84	2.83	1.24	0.80	0.39
5	3	0	11.3	4.12	3.22	8.03	5.46	1.34	1.87	0.82
6	2	0	0	0	0	4.84	0.91	0.91	0	0.91
7	2	0	8.06	3.25	2.38	6.40	4.26	1.01	1.47	0.85
8	2	0	8.87	3.68	2.40	7.19	4.03	0.35	1.40	0.89
9	2	2.42	12.1	6.94	5.58	11.2	7.78	0.84	2.06	1.15
10	2	1.61	12.9	7.65	5.6	9.6	8.04	0.38	3.14	1.62
11	2	1.61	11.3	7.05	6.4	9.6	7.82	0.78	2.10	0.61
SVM	P	MIN1	MAX1	MEAN1	MIN2	MAX2	MEAN2	差	MEAN1 差	MEAN2 差
1	4	0	0	0	0	4.84	0.81	0.81	0	0.73
2	3	0	0	0	0	3.23	1.15	1.15	0	1.15
3	3	0	0	0	0	4.84	0.89	0.89	0	0.85
4	3	0	5.65	1.36	1.61	5.65	2.84	1.48	0.57	0.40
5	3	0	8.87	3.96	3.23	8.06	5.72	1.76	1.71	1.08
6	2	0	0	0	0	4.84	0.85	0.85	0	0.85
7	2	0	7.26	2.89	2.42	6.45	4.19	1.3	1.10	0.78
8	2	0	8.06	3.08	2.42	8.06	3.73	0.65	0.80	0.60
9	2	1.61	9.68	6.45	5.65	8.87	7.45	1	1.57	0.82
10	2	1.61	10.5	6.61	4.84	9.68	7.60	0.99	2.10	1.19
11	2	1.61	8.87	6.70	4.03	10.5	8.10	1.4	1.76	0.89
LDF	P	MIN1	MAX1	MEAN1	MIN2	MAX2	MEAN2	差	MEAN1 差	MEAN2 差
1	4	4.03	16.1	9.64	6.40	12.8	10.5	0.90	9.64	10.50
2	3	4.03	16.9	9.89	8.01	12.8	10.5	0.66	9.89	10.50
3	3	4.03	16.1	9.48	8	13.6	10.1	0.61	9.48	10.10

4	3	4.84	17.7	11.4	7.96	16.8	12	0.60	10.7	9.60
5	3	5.65	20.2	12.4	9.62	15.2	12.7	0.32	10.1	8.05
6	2	4.03	17.7	9.54	8	13.6	9.91	0.37	9.54	9.91
7	2	4.84	17.7	11.8	9.55	16.8	12.2	0.40	9.98	8.76
8	2	4.84	17.7	10.8	7.96	16	11	0.22	8.52	7.89
9	2	5.65	20.2	13.2	9.62	16	13.5	0.28	8.31	6.84
10	2	6.45	20.2	12.5	11.1	15.2	12.6	0.16	7.98	6.23
11	2	8.06	24.2	16.3	12	21.6	16.5	0.20	11.3	9.27

4.2　50%水準と 90%水準の合否判定

得点分布の 50%点の 63 点を合格最低点として 2 群判別を行う．ロジスティック回帰の「MEAN2 差」の範囲は[0.05, 3.01]，SVM は[-0.1, 3.23]，LDF は[0.65, 5.92]である．10%水準ほど悪くないが，LDF は他の 3 手法が最小になるモデル 1 で改定 IP-OLDF より 5.92%も悪い．

得点分布の 90%点の 78 点を合格最低点として 2 群判別を行う．ロジスティック回帰の「MEAN2 差」の範囲は[-0.15,1.8]で，SVM は[0.23,1.48]であるのに対して，LDF は[7.83, 22.58]と明らかに悪い．ロジスティック回帰の平均誤分類確率は，モデル選択と無関係な 9 番目のモデルで改定 IP-OLDF より少ない．

5. 小問 100 問の分析

5.1　合格群の全てが不合格群に誤判別

表 3 は，2012 年度の中間試験の小問 100 個を説明変数とした合否判定である．左から 10%点，50%点，90%点の 3 水準の判別結果である．合否判定できる最小次元は，全ての説明変数の組み合わせモデルで検討していないので，ここで得られたものより少なくなることがある．一応，10%点では 6 変数，50%点では 19 変数，90%点では 15 変数で *MNM=0* であり，ロジスティック回帰でも誤分類数が 0 になった．LDF と QDF はこの次元で誤分類数は 0 でない．さらに QDF は，10%点は 2 変数で誤分類数が 114 人(=114/121*100=94%)，50%点は 21 変数で 67 人(55%)，90%点は 1 変数で 12 人(10%)の合格群の全てが不合格群に誤判別された．

表 3　2012 年度の中間試験の判別結果

p	VAR	MNM	Logi	pLDF	pQDF	VAR	MNM	Logi	pLDF	pQDF	VAR	MNM	Logi	pLDF	pQDF
1	x85	10	14	14	14	x87	26	57/67	26	26	x92	12	34	12	12
2	x15	6	6	6	114	x77	26	26	26	26	x42	8	34	8	12
3	x68	5	6	8	114	x26	15	15	15	16	x21	5	19	5	12
4	x47	3	8	3	114	x89	11	11	11	13	x54	4	4	8	12
5	x7	1	1	3	114	x35	9	10	11	13	x65	1	7	3	12
6	x32	0	0	0	114	x69	6	6	8	10	x100	1	3	3	12
7	x20	0	0	0	114	x25	5	6	6	8	x83	1	3	3	12
14	x98				114	x73	1	1	1	3	x1	1	1	1	12
15	x5				114	x12	1	1	1	3	x62	0	0	1	12
16	x1				114	x7	1	1	1	4	x3			1	12
18	x38				114	x39	1	1	1	2	x60			0	12
19	x6				114	x95	0	0	1	3	x96				12
20	x89				114	x8			0	1	x22				12
21	x100				114	x38				67	x40				12

5.2 QDF が合格群の全てを誤判別する理由と判別理論の修正

(1) QDF が合格群の全てを誤判別する理由

QDF は 2 群の分散共分散行列の逆行列の計算が式中に含まれている．筆者を含めこの式を見て，「少なくとも一方の群のある変数が一定値すなわちばらつかない場合に QDF は計算できないことに気づいた人は少ない」と考える．これを説明するために，LDF,QDF と 2 群の平均の差の検定で説明する.2 群のある変数が同じ一定値をとる場合，統計ソフトはこの変数を分析から省く．両方の群が異なった一定値をとる場合，この変数だけで MNM=0 になるが LDF だけが正しい結果を出す．しかし，QDF や平均の差の検定は計算できない．今回の問題は，一方の群が一定値をとり，他群がばらつく場合を統計ソフトは想定していなかったように考える．LDF と平均の差は計算できるが，QDF は定式化どおりであれば計算できない．しかし，ダーティーなデータの判別にも対応するため，2 群の分散共分散行列の対角要素などを全体の分散共分散行列を参考にして修正しているものと思われる．また正則化法でもこのデータを正しく判別できなかった．

(2)とりあえずの統計ソフトの対応

筆者の指摘に対して，とりあえず「変数値が一定値になる個数のウオーニングを出して，正則化折衷案という暫定手法の利用を薦める」対応策がとられた．しかし LDF と QDF のどちらに近づけるかのオプション値を LDF 寄りに設定しなければ改善できない．このようなクラスター分析のクラスター数や主成分分析の主成分数を決めるようなあいまいな基準を導入するよりも，QDF と正則化法は現時点で対応できないので LDF を使うように進めるべきであろう．また一定値をとる変数がリストアップされておれば，筆者も無駄な研究を 2010 年以降行うことを避けられた．

実はこの問題を解決する方法として，一定値に SD=1^{-6} ぐらいの小さなばらつきを与えるだけで解消できる．10%点の X15 でこれを試みると，p QDF の誤分類数は,14→28→28→28→9→3→0 になる．また，マハラのビスの汎距離を使う MT 理論で基底空間のある変数の分散が 0 の場合は，それが判別に役立つか否かは別途検討する必要がある．

また伝統的な統計の考え方に従えば，LDF,QDF,ロジスティック回帰は 2 群が等確率をデフォルトとすることは正しいが，出現頻度の方が一般的に誤分類確率は少ない．また異なった出現頻度のデータを集めたことには意味があるので，こちらをデフォルトにすべきであろう．また判別分析は，最終的に誤分類数で評価されるので，この分割表にどの判別境界を選んだのか表示しないと分析結果の解釈に混乱を生む可能性が高い．

表 4　判別分析の結果

p	Var.	Cp	AIC	BIC	LDF	QDF	Logi	t-test
1	排気量	47.8	24.2	29	2	0	0	11.37
2	価格	14.9	4.3	10	1	0	0	5.42
3	定員	11.0	1.7	9	1	29	0	8.93
4	ＣＯ２	7.3	-1.4	7	1	29	0	4.27
5	燃費	5.0	-3.4	6	0	29	0	-4.00
6	台数	7.0	-0.4	10	0	29	0	-0.82

5.3 日本車の判別

この問題を、日本車 44 車種の普通車と小型車の 2 群判別で確認する。台数，価格，総排気量，燃費，CO2 排出量，定員の 6 個の説明変数で 2 群判別する.総排気量は,小型車が 0.657 から 0.658 の範囲であり，普通車とは完全に分離している．小型車の定員は 4 人と一定値で，普通車は 5 人以上であり，両変数単独で MNM=0 の判別ができる．

これに対して平均の差の t 検定は、総排気量が 11.37，定員が 8.93，価格が 5.42，CO2 が 4.27，燃費が-4.00，台数が-0.82 であり，1 変数で考えるとこの順で判別成績が悪くなる．特に定員は小型車が一定の値をとっても t 検定は影響されないが，判別関数は統計ソフトによって種々の異なった対応をとっていて利用には注意が必要である．

表 4 は普通車に 1，小型車に 0 のダミー変数を与えて目的変数とし，6 変数で変数増加法を行い変数欄の順に説明変数が取り込まれた．変数減少法では 6 変数のフルモデルから台数，燃費の順に掃き出される．Cp 統計量は 6 変数，AIC と BIC から 5 変数のモデルが選ばれる．これは単に判別データを 2 群に 1/0 のダミー変数を与えると重回帰分析の回帰係数は LDF の判別係数になる．これは「Fisher の仮説」を考えたことは判別関数が母集団と標本を考えた推測統計学的手法と考えることが間違いであり誤解であることを示す．また判別手法には変数選択に用いる手法がないので，「一つとっておき法（Leave one out，LOO)」が開発された.筆者はこれに代わって「小標本の 100 重交差検証法」を提案している．重回帰分析の Cp 統計量，AIC,BIC も単に参考指標として用いる．LDF，QDF，Logi は Fisher の LDF，QDF，ロジスティック回帰の誤分類数である．ロジスティック回帰は「総排気量」がモデルに入ってくると誤分類

数が 0 と正しく判別し,それ以降も 0 と判別する.しかし LDF は 5 変数で初めて誤分類数が 0 になる.QDF は 1 変数と 2 変数で誤分類数が 0 になるのに,3 変数で「定員」がモデルに入ってくると普通車の 29 人が小型車に誤判別される。

6. まとめ

本研究では，試験の合否判定データを用いて MNM=0 の判別分析の問題を検討した.SVM は H-SVM で MNM=0 の判別分析を判別分析の出発点とした点が重要である．改定 IP-OLDF は理論的に MNM=0 のデータを認識できる．ロジスティック回帰は回帰係数の収束計算が不安定になり，標準誤差が大きくなって誤分類数が 0 になるとき，一部の例外を除いて改定 IP-OLDF で MNM=0 になる最小次元の判別モデルを求めることが分かった．しかし，LDF と QDF は MNM=0 のデータを正しく認識できないばかりか誤分類確率が大きいことが分かった．たとえば，2010 年から 2012 年の 3 年間の統計入門の中間と期末の 3 水準の 18 個の大問の合否判定で，LDF の誤分類確率の範囲は[2.3,16.7]，QDF は[0.8,10.8]と大きい．S-SVM は多くの場合，MNM=0 のデータを判別できることが多いが理論的な保証はない．MNM=0 のデータから 0 でないデータまで，改定 IP-OLDF とロジスティック回帰が対応できる以上，分散共分散行列に基づく LDF や QDF は生命を扱う医学診断やパターン認識に用いることは問題と考えられる．また，MNM 基準に基づく改定 IP-OLDF は学習標本で過学習し，検証標本で汎化能力が悪いと考えられるが，実際には正規分布を仮定する LDF が学習標本と検証標本で確率分布を仮定していない S-SVM やロジスティック回帰よりも悪かった．これは Fisher の仮説を満たす現実のデータが少ないのにそれで理論構築したためと考えられる．

謝意

本研究は，統計ソフトの SAS と JMP，そして数理計画法の LINDO，What’sBest!，LINGO といった理数系のソフトで得た知識と経験が大きく貢献している．また LDF とロジスティック回帰の 100 重交差検証法は SAS ジャパンの JMP 事業部，S-SVM と改定 IP-OLDF はシカゴ大学ビジネススクール教授で LINDO Systems Inc.の創業者の L.Schrage 教授の協力を得て作成した．

文　献

[1] N.Cristianini & J.Shawe-Taylor, サポートベクターマシン入門．共立出版，2005．［大北剛訳］.

[2] B.Efron, Bootstrap Methods -Another Look at the Jackknife-,The Annals of Statistics, 7/1,1－26, 1979.

[3] D.Firth, Bias reduction of maximum likelihood estimates. Biometrika, 80, 27-38, 1993.

[4] R.A.Fisher, The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188, 1936.

[5]B. Flury,& H.Rieduyl, Multivariate statistics -A practical approach-. Cambridge University Press, 1988.［田端吉雄訳，多変量解析とその応用，1900.］.

[6] 石井健一郎，上田修功，前田英作，村瀬洋．分かり易いパターン認識．オーム社,1988.

[7] P.A.Lachenbruch, & M.R.Mickey, Estimation of error rates in discriminant analysis. Technometrics 10, 1-11, 1968.

[8] A.Miyake & S.Shinmura, Error rate of linear discriminant function, F.T. de Dombal & F.Gremy , edit 435-445, North-Holland,1976.

[9]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，18-1，15-20，1980.

[10] Y.Nomura & S.Shinmura, Computer assisted prognosis of acute myocardial infarction, MEDINFO 77, Shires / Wolf, editors IFIP 517-521, North Holland Publishing Company,1978.

[11] J. Sall (新村訳),SAS による回帰分析の実践，朝倉書店，東京,1986.

[12] J.P.Sall, L.Creighton, & A.Lehman(新村秀一監修),JMP を用いた統計およびデータ分析入門(第 3 版). SAS Institute Japan ㈱,2004.

[13] L.Schrage, LINDO -An optimization modeling system-. The Scientific Press,1981. ［新村秀一，高森寛，実践数理計画法．朝倉書店,1992. ］.

[14] L.Schrage, Optimizer Modeling with LINGO. LINDO Systems Inc, 2003.

[15]A.Stam, Nontraditional approaches to statistical classification：Some perspectives on Lp-norm methods. Annals of Operations Research, 74, 1-36,1997.

[16] 新村秀一，北川護，高木義人，野村裕,二段階重みづけによるスペクタル診断，第 12 回日本 ME 学会大会論文集，107-108,1973.

[17] 新村秀一，鈴木隆一郎，中西克己,各種判別手法を用いた医療データ解析の標準化 ― マンモグラフィによる乳癌の診断 ― 医療情報学,3-2,38-50,1983.

[18]新村秀一，三宅章彦,重回帰分析と判別解析のモデル決定（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，3-3，107-124,1983.

[19]新村秀一,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,29-7,415-421,1984.

[20] 新村秀一,体験に基づく汎用統計パッケージの紹介．品質，17-3，261-268,1987.

[21] 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．11-2,89-101,1988.

[22] 新村秀一，垂水共之,2 変量正規乱数データによる IP-OLDF の評価.計算機統計学 12-2,107-123,1999.

[23] S.Shinmura, A new algorithm of the linear discriminant function using integer programming. New Trends in Probability and Statistics 5,133-142,2000.

[24] 新村秀一,数理計画法を用いた最適線形判別関数(1)から(4)．オペレーションズ・リサーチ,47/1 から 47/5,2002.

[25] 新村秀一,JMP による統計学とっておき勉強法，東京，講談社，2004.

[26] 新村秀一,改定IP-OLDFによるSVMのアルゴリズム研究,オペレーションズ・リサーチ,51/11,702-707,2006.
[27] 新村秀一,改定IP-OLDFによるIP-OLDFの問題点の解消, 計算機統計学, 19/1,1-16,2007.
[28] 新村秀一,Excel と LINGO で学ぶ数理計画法, 丸善,2007.
[29] 新村秀一,数理計画法による判別分析の 10 年. 計算機統計学, 20/1&2, 59-94,2007.
[30] 新村秀一,線形計画法による改定 IP-OLDF の計算時間の改善. 計算機統計学, 22/1,37-57,2009.
[31] 新村秀一,最適線形判別関数. 日科技連出版社,2010.
[32] 新村秀一,数理計画法による問題解決法.日科技連出版社,2011.
[33] 新村秀一,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, 157-172,2011.
[34] 新村秀一,SAS/JMP との歩み, SAS Technical Report, 13－16, 2012 春号.
[35] 新村秀一,統計教育における判別分析の改善点. 統計教育実践研究第 5 巻-統計数理研究所研究レポート 293-, 36-45,2013.
[36] 田口玄一,タグチメソッドわが発想法,経済界,1999.
[37] V.Vapnik,The Nature of Statistical Learning Theory. Springer-Verlag,1995.
[38] R.Warmack & R.C.Gonzalez,An Algorithm for the optimal solution of linear inequalities and its application to pattern recognition,IEEE Trans.Computers, 1065-1075,1973.

付録 A:LINGO のモデル

数理計画法ソフトの LINGO は、シカゴ大学ビジネススクールの Linus Schrage 教授が開発し、LINDO Systems Inc. (http://WWW.LINDO.COM)が開発する LINGO、What'sBest!(Excel のアドイン最適化ソルバー)、LINDO API(最適化ライブラリー)の汎用ソルバーの一つである。LP、QP、IP、NLP、確率計画法などの数理計画法ソルバーを、ユーザーが意識することなく利用できる汎用ソフトである。数理計画法モデルを通常の数式モデルで利用する自然形式と、以下で説明する集合表記による使い分けができる。自然表記は数理計画法モデルを通常の数式モデルとして扱える分かりやすさがある。集合表記は、大規模モデルの開発や、複雑な最適化システムを CALC 節で制御し、Excel などの外部 DB とデータの入出力ができる。改定 IP-OLDF のモデルや 100 重交差検証法のプログラムは[32]の付録を参照。

以下では簡単な文法を説明する。

最初に集合節は、「SETS:」で始まり「ENDSETS」で終わる。その中で 1 次元の集合とそれで管理される配列を定義する。例えば、P は 1 次元の集合で、P1 は定数項を含む集合とする。VARK は、P1 の属性をもつ 1 次元の配列になる。N は 1 次元の集合でケース数の数だけの要素をもつ。SCORE は n 件のケースの判別スコアを入れる 1 次元配列、E は 0/1 の整数変数である。2 次元集合 D は、一次元集合の N と P1 で定義される。判別データが 20 件*定数項を含め 10 件であれば、20*10 の 2 次元配列 IS を定義する。このように必要な 1 次元集合を定義し、それによって 2 次元以上の集合とその属性をもつ配列が定義される。そして、数理計画法モデルでこれらの集合の引数でモデルが操作できる。

```
SETS:
 P;P1: VARK;
 N: SCORE,E,CONSTANT;
 D(N,P1):ES;
ENDSETS
```

DATA節は、「DATA:」で始まり「ENDDATA」で終わり、配列に値を与えたり,「@OLE関数」で、外部DBとデータの入出力が行える。例えばPは集合Pが19変数の説明変数を表し、P1は定数項を含んで20個の要素をもつことを示す。Nは集合Nの要素数すなわちケース数が240件を意味し、N2はその100倍のリサンプリング標本に対応している。これに対しPenaltyはSVMの c を1に設定し、改定IP-OLDFのBigM定数を1000に設定している。「@OLE関数」を等号の右に指定すれば、LINGOの配列CHOICEとESにExcelからCHOICEとESというセル範囲名をもつデータを入力する。逆に「@OLE=VARK;」とすれば、最適化計算で求まった判別係数をExcelの同名のセル範囲名に出力する。

```
DATA:
  P=1..19;P1=1..20; N=1..240;N2=1..24000;MS=1..19;G100=1..100;
  penalty=1;BIGM=1000;
  CHOICE,ES=@OLE();
ENDDATA
```

SUBMODEL節は、例えば「SUBMODEL SVM;で始まり、ENDSUBMODEL」で終わるS-SVMを定義したり、「SUBMODEL RIP;」で改定IP-OLDFを定義する。「MIN=」は右辺を最小化する。右辺の「Penalty*@SUM(N(i):E(i))」は「MIN= ＊＋c*Σei」を表す。すなわち「@SUM(N(i):E(i))」の「@SUM(N(i):)は集合Nの要素数(240件)だけ「E(i)」を合計しなさいを表す。「@SUM(P(J1): VARK(j1)^2)/2」はSV間距離の逆数の最小化を定義する。次の「@FOR(N(i):)は240件のケースに対し、「@SUM(P1(j):IS(i,j)*VARK(j)*CHOICE(k,j)) > 1-E(i));」という制約式「yi*f(xi)>=1-ei;」を定義する。「@FOR(P1(j): @FREE(VARK(j)));」で定数項を含む20変数は自由変数（正にも負にもなる）であることを宣言している。

```
SUBMODEL SVM:
 MIN=@SUM(P(J1):VARK(j1)^2)/2+Penalty*@SUM(N(i):E(i));
 @FOR(N(i):@SUM(P1(j):IS(i,j)*VARK(j)*CHOICE(k,j)) > 1-E(i));
 @FOR(P1(j):@FREE(VARK(j)));
 ENDSUBMODEL
SUBMODEL RIP:
 MIN=@SUM(N(i):E(i));
 @FOR(N(i):@SUM(P1(j):IS(i,j)*VARK(j)*CHOICE(k,j) )  > 1-BIGM*E(i));
 @FOR(P1(j):@FREE(VARK(j))); @FOR(N(i):@BIN(E(i)));
ENDSUBMODEL
```

CALC節は、「CALC:で始まりENDCALC」で終わり、最適化システムを制御し、データの入出力を行うプログラミング機能をもっている。「@WHILE」で100重交差検証法のためのループ計算を行い、「@SOLVE(SVM);」はSUBMODELで定義したS-SVMの最適化をおこなっている。

```
CALC:
 K=1;Lend=@SIZE(MS);  @SET('TERSEO',1);
@WHILE(K#LE#Lend: f=1;
 @WHILE(f#LE#100:@FOR(D(i,j):
    IS(i,j)=ES( @SIZE(N)*(f-1)+i,j));MNM=0;ER1=0 ;
    @SOLVE(SVM);
ENDCALC
DATA:
@OLE( )=VARK100;@OLE( )=IC;@OLE( )=EC;
ENDDATA
```