中央大学理工学研究所論文集　第12号　2006年
Journal of the Institute of Science and Engineering. Chuo University

高次元多変量2値データに基づく判別の変数選択とDNAフィンガープリントデータへの応用

藤越康祝*，櫻井哲朗*，安部友紀†，薬真寺　裕‡，杉山高一*

Variable Selection in Discriminant Analysis with High-Dimensional Binary Data and Its Application to DNA Fingerprint Data

Yasunori FUJIKOSHI*, Tetsuro SAKURAI*, Yuuki ABE†,
Yutaka YAKUSHINJI‡, Takakazu SUGIYAMA*

abstract

This paper is concerned with the problem of selecting variables in multiple discriminant analysis with high-dimensional multivariate binary data. Let $\boldsymbol{x} = (x_1, \ldots, x_p)'$ be the variables considered, and it is assmed that x_i's are $0-1$ variables. Here we consider a high-dimensional situation where p is large or similar compared to the sample size n. In fact, the DNA fingerprint data (Nakatsu et al. (2000)) considered consist of $p = 84$ and $n = 89$. One of two variable selection methods proposed by Wilbur et al. (2003) is based on the marginal discriminant powers $D_i, i = 1, \ldots, p$, where D_i is (the sum of squares due to within-groups)/(the sum of squares due to between-groups) for the i-th variable. They proposed to select the variabeles whose D_i's are significant, assuming that x_i's are independent. In this paper we first give a model selection criterion AIC_B by introducing a class of appropriate parametric models. Next we propose a selection method without assuming the independence of x_i's which is a sequential procedure based on the conditional discriminant powers. The stopping rule is based on the probability of misclassifications and the model selection criterion AIC_B. We also propose a modified selection method by starting from a reduced set of variables based on the marginal dicriminant powers. Our methods are applied to the DNA fingerprint data to find some better subsets of variables.

1　はじめに

多変量解析において，取り扱う変数が標本数より大きい場合，あるいは同程度な場合は，高次元多変量解析とよばれる．最近，ゲノムデータやファイナンスデータを初めとして，この種のデータが増えている．本論文では，DNA フィンガープリントデータの分析に関連して，高次元の場合における多変量 2 値データの多群への判別について考える．とくに，群間の違いを特徴付ける変数の組を見つけるための変数選択問題に焦点を当てている．

*中央大学理工学研究科
†広島県立高陽東高等学校
‡愛媛県立新居浜工業高等学校

このような問題に対して，Wilber et al. [1] は 2 つの変数選択法を提案している．また，トウモロコシの 4 種類の栽培法の違いを DNA フィンガープリントデータから特徴付けることを試みている．この多変量データでは，変数の数は $p = 84$ で，データは 4 群からなり全標本数は $n = 89$ である．選出された変数の組の評価基準としては，判別的中率を用いている．ここでは独立 2 項モデルを想定しているが，提案されている 2 つの変数選択法は正準判別関数の係数による方法と，各周辺毎の $D_i = b_{ii}/w_{ii}$ (b_{ii} は第 i 変数の群間平方和，w_{ii} は第 i 変数の群内平方和) に基づく方法であって，考えられているモデルに特有なものではないことを注意したい．

本論文は，Wilber et al. [1] によって考察された問題に対して，新たな発展を与えることを目的とする．まず，独立 2 項モデルのもとで変数選択のためのモデルを導入し，モデル選択基準 AIC を適用する方法を提案する．高次元の場合には，すべての変数の組に対して AIC 基準を求めるのは困難であり，考慮すべき変数の組を限定することが重要となる．その方法の 1 つとして D_i が大きい変数から逐次選択する方法を提案する．次に，変数間の相関を考慮し，条件付 D_i 統計量を用いた方法も与える．後者の場合，選ばれた変数の評価基準として多変量正規モデルでの AIC を用いる方法も提案する．条件付 D_i 統計量の適用に当たっては，D_i 基準によって，変数をある程度絞り込んでから適用する方法も提案する．さらに，これらの変数選択法を DNA フィンガープリントデータに応用し，有効性を検証した．

本論文は次のように構成されている．まず 2 節においては，今回取り扱う DNA フィンガープリントデータについて簡単な説明を与える．次に 3 節では，本論文で用いられる正準判別法と最大尤度法，および，誤判別率の推定法について説明する．4 節においては，Wilber et al. [1] によって提案された変数選択法を解説するとともに，その問題点を指摘する．さらに，それらの問題点を考慮した新たな変数選択法を提案する．提案された変数選択法においては，選ばれた変数の評価基準として誤判別率による評価に加えてモデル評価基準による評価を与えている．最後に，5 節において提案した変数選択法をトウモロコシの栽培法についての DNA フィンガープリントデータに適用し，有効な変数の組を見い出す．

2　DNA フィンガープリントデータ

DNA フィンガープリントとは，DNA を切断したときの断片のことであって，このような断片は，通常，制限酵素や耐熱性 DNA 合成酵素などを用いて作られる．これらの作られた切断面のバンドパターンはヒトの指紋と同様に個体を識別する標識として利用できるため，DNA フィンガープリントとよばれている．このようにして DNA を見ることにより，生物の活動を決定する遺伝子を調べる研究も行われている．しかし，遺伝子は膨大な DNA 配列の中に点在しているため，遺伝子そのものを確認することは非常に多くの労力を必要とする．

そこで，近年、遺伝子そのものを確認しないで，個体を識別する方法の開発が進められてきている．それは遺伝子の近くに存在し，遺伝子とともに変化する DNA の配列を探し出し，それによって識別する方法である．このような DNA 配列は DNA マーカーとよばれている．

具体的には，ある特定の配列で切断して得られた DNA フィンガープリントを 2 進法で数値に変換し，DNA マーカーの有無を確認することによって個体の識別が行われる．本論文で，扱うデータ分析では，トウモロコシの栽培方法の違いから，特有の特徴をもつ微生物群生を DNA フィンガープリントで見つけることを目的とする．栽培方法は，無耕農法であるかあるいは耕作農法であるか，さらに単一耕作であるか二毛作であるかによって，全部で 4 つのグループに分けられている．データは，得られた DNA フィンガープリントと DNA マーカーを比較することで，DNA フィンガープリントと DNA マーカーが共通して反応しているところを「1」，またそうでないところを「0」として数値化したものである．

したがって，ここで扱う DNA フィンガープリントデータは次のように表せる．p 個のベルヌーイ変数 $x_1, \ldots, x_p$ が q 個のグループ $G_1, \ldots, G_q$ において観測されており，第 g グループ G_g における p 次元変数 $\boldsymbol{x} = (x_1, \ldots, x_p)'$ の観測値を

$$
\begin{aligned}
&G_g;\ \boldsymbol{x}_1^{(g)}, \ldots, \boldsymbol{x}_{n_i}^{(g)}, \quad g = 1, \ldots, q, \\
&\boldsymbol{x}_j^{(g)} = (x_{1j}^{(g)}, \ldots, x_{pj}^{(g)})', \quad j = 1, \ldots, n_g, \qquad\qquad x_{ij}^{(g)} : 0 \text{ または } 1 \text{ の値をとる}
\end{aligned}
$$

と表わす．ここでは，変数の次元 p が全標本数 $n = n_1 + \cdots + n_q$ に近いか，あるいは，それよりも大きい高次元データの場合における判別分析の変数選択問題を考えていく．Wilber et al. [1] によって分析された，トウモロコシの DNA フインガープリントデータの場合は，数千ある変数のなかからあらかじめ特異性のある変数の組が選ばれており，実際のデータは，

$$p = 84, \quad q = 4, \quad n_1 = 23, \quad n_2 = n_3 = n_4 = 22,$$
$$n = n_1 + n_2 + n_3 + n_4 = 89$$

である．このデータは以下のように表せる．具体的な値 (Nakatsu et al. [2]) については，付録の Table 5 を参照されたい．

	$n_1 = 23$	$n_2 = 22$	$n_3 = 22$	$n_4 = 22$
G_1	$x_{11}^{(1)}, \cdots, x_{p1}^{(1)}$	$x_{11}^{(2)}, \cdots, x_{p1}^{(2)}$	$x_{11}^{(3)}, \cdots, x_{p1}^{(3)}$	$x_{11}^{(4)}, \cdots, x_{p1}^{(4)}$
	$x_{12}^{(1)}, \cdots, x_{p2}^{(1)}$	$x_{12}^{(2)}, \cdots, x_{p2}^{(2)}$	$x_{12}^{(3)}, \cdots, x_{p2}^{(3)}$	$x_{12}^{(4)}, \cdots, x_{p2}^{(4)}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$x_{1n_1}^{(1)}, \cdots, x_{pn_1}^{(1)}$	$x_{1n_2}^{(2)}, \cdots, x_{pn_2}^{(2)}$	$x_{1n_3}^{(3)}, \cdots, x_{pn_3}^{(3)}$	$x_{1n_4}^{(4)}, \cdots, x_{pn_4}^{(4)}$
計	$y_1^{(1)}, \cdots, y_p^{(1)}$	$y_1^{(2)}, \cdots, y_p^{(2)}$	$y_1^{(3)}, \cdots, y_p^{(3)}$	$y_1^{(4)}, \cdots, y_p^{(4)}$

G_1： 耕作農法で単一耕作のトウモロコシ
G_2： 無耕作農法で単一耕作のトウモロコシ
G_3： 耕作農法で大豆との二毛作のトウモロコシ
G_4： 無耕作農法で大豆との二毛作のトウモロコシ

変数の数が $p = 84$ であると，考えられる変数の組は，$2^{84} - 1 \fallingdotseq 1.93 \times 10^{25}$ 通りある．これら全ての変数の組に対して誤判別率やモデル評価基準を計算することは到底できず，総当りによる変数選択の手法は適用できない．そこで

$$\text{"}p \sim n\text{"}, \quad \text{"変数の組 } 2^p\text{"}$$

という状況で，如何にして良い判別を与える変数の組を見い出すかの問題を考える．

3 判別法

多くの判別法が考えられているが，ここでは多群の高次元 2 値データの判別に適用可能な正準判別法と最大尤度法を用いるので、これらの方法について簡単にまとめておく．詳しくは，Krzanowski and Marriot [3], McLachlan [4] などを参照されたい．

3.1 正準判別法による判別

p 次元変数 $\boldsymbol{x} = (x_1, \ldots, x_p)'$ が q 個の群 $G_1, \ldots, G_q$ で観測され，群 G_g における標本を $\boldsymbol{x}_1^{(g)}, \ldots, \boldsymbol{x}_{n_g}^{(g)}$ とする．このとき，群間平方和積和行列 B および群内平方和積和行列 W は次のように定義される．

$$B = \sum_{g=1}^{q} n_g (\bar{\boldsymbol{x}}^{(g)} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}^{(g)} - \bar{\boldsymbol{x}})', \qquad W = \sum_{g=1}^{q} \sum_{j=1}^{n_g} (\boldsymbol{x}_j^{(g)} - \bar{\boldsymbol{x}}^{(g)})(\boldsymbol{x}_j^{(g)} - \bar{\boldsymbol{x}}^{(g)})'.$$

ここで，$\bar{\boldsymbol{x}}^{(g)}$ は群 G_g の平均，$\bar{\boldsymbol{x}}$ は全平均であって

$$\bar{\boldsymbol{x}}^{(g)} = \frac{1}{n_g} \sum_{j=1}^{n_g} \boldsymbol{x}_j^{(g)}, \qquad \bar{\boldsymbol{x}} = \frac{1}{n} \sum_{g=1}^{q} \sum_{j=1}^{n_g} \boldsymbol{x}_j^{(g)}$$

である．

正準判別関数 $\boldsymbol{h}_i'\boldsymbol{x}, i=1,\ldots,m$ の係数ベクトル $\boldsymbol{h}_i$ は $W^{-1}B$ のゼロでない固有値 $\ell_1>\ldots>\ell_m>0$ に対応する固有ベクトルとして定義される．ここに，$m=\min(p,q-1)$．より正確には ℓ_i と $\boldsymbol{h}_i$ は固有方程式

$$B\boldsymbol{h}_i=\ell_i W\boldsymbol{h}_i,\qquad \boldsymbol{h}_i'W\boldsymbol{h}_j=n\delta_{ij}$$

の解であ．ここに，δ_{ij} はクロネッカーのデルタであって，$\delta_{ii}=1$，i と j が異なれば $\delta_{ij}=0$ である．また，$\ell_{m+1}=\ldots=\ell_p=0$ とする．

新たに得られた所属不明の観測値 $\boldsymbol{x}=(x_1,\ldots,x_p)'$ の判別には，まず $\boldsymbol{x}$ および各標本 $\boldsymbol{x}_j^{(g)}$ の正準判別得点

$$\boldsymbol{z}=[\boldsymbol{h}_1,\ldots,\boldsymbol{h}_{q-1}]'\boldsymbol{x},\quad \boldsymbol{z}_j^{(g)}=[\boldsymbol{h}_1,\ldots,\boldsymbol{h}_{q-1}]'\boldsymbol{x}_j^{(g)}$$

を計算する．このとき，各群の平均正準判別得点は

$$\bar{\boldsymbol{z}}^{(g)}=\frac{1}{n_g}\sum_{j=1}^{n_g}\boldsymbol{z}_j^{(g)}=[\boldsymbol{h}_1,\ldots,\boldsymbol{h}_{q-1}]'\bar{\boldsymbol{x}}^{(g)}$$

と表せる．$\boldsymbol{x}$ の判別は，$\boldsymbol{z}$ と各群の平均正準判別得点との距離を調べ最も近い群へ判別する．すなわち，判別方式は

$$\min_g\|\boldsymbol{z}-\bar{\boldsymbol{z}}^{(g)}\|^2=\|\boldsymbol{z}-\bar{\boldsymbol{z}}^{(c)}\|^2\Rightarrow\boldsymbol{x}\in G_c$$

と表せる．

判別の有効性は誤判別率あるいは判別的中率によって評価される．これらの推定法として，各標本を実際に判別したときの誤判別率あるいは判別的中率によって推定することができる．このような推定法の改良法として，交差確認法 (Cross-varidation Method) がある．これは，$\boldsymbol{x}_j^{(g)}$ を判別するときは，データから $\boldsymbol{x}_j^{(g)}$ を取り除いて正準判別変数を構成して判別する方法である．

この正準判別法においては，判別関数を求めるとき，群内平方和積和行列 W の逆行列を求める必要があるが，$p>n-q$ のときは W が特異になり計算が困難になるという問題点がある．この点，次に述べる最大尤度法は n,p がどのような場合でも判別法が定義される．

3.2　最大尤度法による判別

p 次元変数 $\boldsymbol{x}=(x_1,\ldots,x_p)'$ は，$\boldsymbol{x}$ が群 G_g のもとでは $x_1,\ldots,x_p$ は互に独立で，$x_i\sim B(1,\theta_i^{(g)})$ に従うものとする．ここで $x\sim B(n,\theta)$ は成功確率 θ の n 回の試行を繰り返したときの 2 項分布を表す．今 $\boldsymbol{x}$ を $G_1,\ldots,G_q$ のいずれかに判別したいとしよう．$\boldsymbol{x}$ が G_g に属するときの尤度は

$$L_g=\prod_{i=1}^{p}\left(\hat{\theta}_i^{(g)}\right)^{x_j^{(g)}}\left(1-\hat{\theta}_i^{(g)}\right)^{1-x_j^{(g)}}$$

と推定できる．ここに $\hat{\theta}_i^{(g)}$ は $\theta_i^{(g)}$ の最尤推定量であって

$$\hat{\theta}_i^{(g)}=\frac{1}{n_g}\sum_{j=1}^{n_g}x_{ij}^{(g)}$$

と表せる．最大尤度法は L_g が最大となる群へ判別する方法であって，

$$\max_g L_g=L_c\Rightarrow\boldsymbol{x}_j^{(g)}\in G_c$$

と判別する．また，このようにして判別したときの誤判別率の推定法は正準判別のときと同様にして，各データを実際に判別し，そのときの誤判別率で推定する．この場合にも，$\boldsymbol{x}_j^{(g)}$ を判別するときにはデータから $\boldsymbol{x}_j^{(g)}$ を取り除いて構成した尤度推定を用いて判別する交差確認法による方法がある．

4 変数選択法

ここでは，はじめに Wilber et al. [1] で提案された変数間の独立性を仮定した場合の変数選択法を述べ，その問題点を指摘する．次に，変数間の独立性を仮定しない場合の変数選択法を提案する．また，どの変数の組が最適な判別であるかを調べるためモデル評価基準を与える．

4.1 変数間の独立性を仮定した変数選択

Wilber et al. [1] による DNA フインガープリントデータの解析では，p 次元変数 $\boldsymbol{x} = (x_1, \ldots, x_p)'$ について次のことを仮定して分析を行っている．

1. p 個の変数は互い独立，
2. $x_i^{(g)} | \boldsymbol{x} \in G_g \sim B(1, \theta_i^{(g)}), \qquad i = 1, 2, \cdots, p$

このとき，群間平方和積和行列 $B = (b_{ij})$ と群内平方和積和行列 $W = (w_{ij})$ を用いて次の 2 つの変数選択法を提案している．

4.1.1 変数選択法 (1)

第 i 成分の基準化された群間の変動の大きさを表している統計量

$$D_i = \frac{b_{ii}}{w_{ii}} = \frac{\displaystyle\sum_{g=1}^{q} n_g (\bar{x}_i^{(g)} - \bar{x}_i)^2}{\displaystyle\sum_{g=1}^{q} \sum_{j=1}^{n_g} (x_{ij}^{(g)} - \bar{x}_i^{(g)})^2}$$

を考える．ここで，統計量 D_i は，$x_{ij}^{(g)}$ が 0 または 1 であることにより，

$$D_i = \frac{\displaystyle\sum_{g=1}^{q} n_g (\bar{x}_i^{(g)} - \bar{x}_i)^2}{\displaystyle\sum_{g=1}^{q} n_g \bar{x}_i^{(g)} (1 - \bar{x}_i^{(g)})}$$

となる．なぜならば，

$$\begin{aligned} \sum_{j=1}^{n_g} (x_{ij}^{(g)} - \bar{x}_i^{(g)})^2 &= \sum_{j=1}^{n_g} \left\{ x_{ij}^{(g)^2} - 2 x_{ij}^{(g)} \bar{x}_i^{(g)} + \bar{x}_i^{(g)^2} \right\} \\ &= n_g \left(\bar{x}_i^{(g)} - 2 \bar{x}_i^{(g)^2} + \bar{x}_i^{(g)^2} \right) \\ &= n_g \bar{x}_i^{(g)} (1 - \bar{x}_i^{(g)}). \end{aligned}$$

これにより，D_i は全データの平均と群平均だけに依存していることが分かる．

このようにして求められた $D_1, D_2, \ldots, D_p$ を各変数の重要度と考え，変数を選択していく．実際，Wilber et al. [1] では各変数毎に有意性検定を行い，有意な変数を選択する方法を提案している．具体的には，第 i 成分についての有意性仮説

$$H_i \colon \theta_i^{(1)} = \cdots = \theta_i^{(q)}$$

を統計量 $D_i = b_{ii}/w_{ii}$ を用いて検定し，有意ならば変数 x_i を選ぶ．棄却点 d_i は，

$$P(D_i > d_i|H_i) = 0.05$$

をみたすものであるが，並べ替え検定により決めている．

4.1.2　変数選択法 (2)

また Wilber et al. [1] では，標本数 n と次元数 p が近いかあるいは $p > n - q$ 場合において W^{-1} を直接計算することが困難であるため，$W^{-1}B$ を

$$V = (v_{ij}), \quad v_{ij} = \frac{b_{ij}}{\sqrt{w_{ii}w_{jj}}}$$

で近似したときの判別関数の係数に基づく方法も提案している．V の固有値を $\tilde{\ell}_i$，対応する固有ベクトルを $\tilde{\boldsymbol{h}}_i = (\tilde{h}_{1i}, \cdots, \tilde{h}_{pi})'$ とすると，

$$V\tilde{\boldsymbol{h}}_i = \tilde{\ell}_i\tilde{\boldsymbol{h}}_i, \qquad i = 1, \cdots, m = \min\{p, q-1\}.$$

選択する変数の集合を

$$M = \left\{ x_k| \text{ 少なくとも 1 つの } i \text{ に対して, } \tilde{h}_{ki} \notin \left(c_{ki}\left(\frac{1}{2}\alpha\right), c_{ki}\left(1 - \frac{1}{2}\alpha\right)\right), \ \ i = 1, \cdots, m \right\}$$

とする．ここに，棄却点 c_{ki} は，

$$P\left(\tilde{h}_{ki} < c_{ki}\left(\frac{1}{2}\alpha\right)\right) = P\left(\tilde{h}_{ki} > c_{ki}\left(1 - \frac{1}{2}\alpha\right)\right) = \frac{1}{2}\alpha$$

となるものであるが，これらの点も並べ替え検定を利用して定めている．

変数間に独立性を仮定した場合の変数選択は以上により与えることができるが，一般に，生物学的見地から，隣接塩基間においては何らかの影響を及ぼし合い，独立性が認められないことが指摘されている．

このため，独立性の仮定が妥当であるかどうかを調べてみた．高次元の場合のおける独立性の検定法の 1 つとして Schott [5] によるものがある．これは母集団が正規性の仮定もとでの結果であるが，ある程度の標本数と次元数があれば 2 値データのような離散分布でも適用できることがわかっている (青木等 [6])．

その検定は相関行列の非対角成分の 2 乗和に基づくものであるが，

$$t_{n,p} = \sum_{i=2}^{p}\sum_{j=1}^{i-1} r_{ij}^2 - \frac{p(p-1)}{2(n-1)}$$

が漸近的に $N(0, \sigma_{n,p}^2)$ であることを用いる．ここに

$$\sigma_{n,p}^2 = \frac{p(p-1)(n-2)}{(n-1)^2(n+1)}.$$

今のデータでは

$$\text{標本数} : n = 89, \text{ 次元数} : p = 84,$$
$$\sum_{i=2}^{p}\sum_{j=1}^{i-1} r_{ij}^2 = 62.47258, \qquad \frac{p(p-1)}{2(n-1)} = 39.61364, \qquad \sigma_{t_{n,p}}^2 = \frac{p(p-1)(n-2)}{(n-1)^2(n+1)} = 0.8702996$$

となり

$$t_{n,p}/\sigma_{n,p} = 24.50315$$

であるから独立性の仮説は棄却される．

また，Wilber et al. [1] において変数が多くなると独立性の仮定が崩れるやすくなることが示されている．したがって，独立性の仮定を外した場合の変数選択法のアルゴリズムを考える必要がある．

4.2 変数間の独立性の仮定を外した場合の変数選択

4.2.1 変数選択法 (3)

まず独立性を仮定した場合と同様に統計量 $D_1,\dots,D_p$ を B と W を用いて求める．これらを大きさの順に並べる．

$$D_{i_1} \geq D_{i_2} \geq \cdots \geq D_{i_p}$$

とする．以下記号簡単のため

$$i_1 = 1, i_2 = 2, \dots, i_p = p$$

とする．つまり $(x_1, x_2, \dots, x_p)$ については，

$$D_1 \geq D_2 \geq \cdots \geq D_p$$

が成立している．これより

$$\max_{1\leq i\leq p} D_i = D_1^2$$

から変数 x_1 が選ばれる．

次に

$$B = \left(\begin{array}{c|c} b_{11} & \boldsymbol{b}'_{21} \\ \hline \boldsymbol{b}_{21} & B_{22} \end{array}\right), \quad \begin{array}{l} \boldsymbol{b}'_{21} : 1\times(p-1), \\ B_{22} : (p-1)\times(p-1) \end{array} \qquad W = \left(\begin{array}{c|c} w_{11} & \boldsymbol{w}'_{21} \\ \hline \boldsymbol{w}_{21} & W_{22} \end{array}\right), \quad \begin{array}{l} \boldsymbol{w}'_{21} : 1\times(p-1), \\ W_{22} : (p-1)\times(p-1) \end{array}$$

と分割して，変数 x_1 の影響を除いた群間平方和積和行列

$$B_{22\cdot 1} = B_{22} - \boldsymbol{b}_{21} b_{11}^{-1} \boldsymbol{b}'_{21} = (b_{ij\cdot 1}), \qquad i,j = 2,\dots,p$$

を求める．同様に変数 x_1 の影響を除いた群内平方和積和行列

$$W_{22\cdot 1} = W_{22} - \boldsymbol{w}_{21} w_{11}^{-1} \boldsymbol{w}'_{21} = (w_{ij\cdot 1}), \qquad i,j = 2,\dots,p$$

を求める．x_1 の影響を除いた，あるいは，x_1 が与えられたときの条件付 D_i 統計量を

$$D_{i|1} = \frac{b_{ii\cdot 1}}{w_{ii\cdot 1}} \qquad i = 2,\dots,p$$

と定義し，$D_{2|1}, D_{3|1}, \dots, D_{p|1}$ が最大になる変数を選ぶ．

以下順次繰り返して，k 個の変数 $x_{i_1},\dots,x_{i_k}$ が順次選ばれたとする．ここで，記号簡単のためあらためて $i_1 = 1, i_2 = 2, \dots, i_k = k$ とする．k 個の変数 $\{x_1,\dots,x_k\}$ が選択されると

$$B = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array}\right), \quad \begin{array}{l} B_{12} : k\times(p-k), \\ B_{22} : (p-k)\times(p-k) \end{array} \qquad W = \left(\begin{array}{c|c} W_{11} & W_{12} \\ \hline W_{21} & W_{22} \end{array}\right), \quad \begin{array}{l} W_{12} : k\times(p-k), \\ W_{22} : (p-k)\times(p-k) \end{array}$$

と分割して，変数 $\{x_1,\dots,x_k\}$ の影響を除いた群間平方和積和行列

$$B_{22\cdot 1} = B_{22} - B_{21} B_{11}^{-1} B_{12} = (b_{ij\cdot 1\cdots k}), \qquad i,j = k+1,\dots,p$$

を求める．同様に変数 $\{x_1,\ldots,x_k\}$ の影響を除いた群内平方和積和行列

$$W_{22\cdot 1}=W_{22}-W_{21}W_{11}^{-1}W_{12}=(w_{ij\cdot 1\cdots k}),\qquad i,j=k+1,\ldots,p.$$

を求める．これらの平方和積和行列を用いて，$x_1,\ldots,x_k$ が与えられたときの条件付 D_i 統計量

$$D_{i|1\cdots k}=\frac{b_{ii\cdot 1\cdots k}}{w_{ii\cdot 1\cdots k}},\quad i,j=k+1,\ldots,p$$

を計算し，その中から最大ものに対応する変数を選ぶ．

上で提案された変数選択法のアルゴリズムは次のようにまとめられる．

(1) $\max_i D_i=D_{i_1}$

(2) $\max_i D_{i|i_1}=D_{i_2|i_1}$

(3) $\max_i D_{i|i_1,i_2}=D_{i_3|i_1,i_2}$

以下，同様

また，記号簡単のため，選ばれた変数の添え字を，

$$(i_1,\cdots,i_k)=(1,\cdots,k)$$

とし，変数 $x_{[k]}=(x_{i_1},\ldots,x_{i_k})=(x_1,\ldots,x_k)$ と表す．

この方法は W の部分行列の逆行列を用いるため，ここで扱うデータのように n と p が近かったり，あるいは $n-q\leq p$ の場合には計算が困難という問題点がある．これを解消するため次の修正変数選択法を提案する．

4.2.2　変数選択法 (4)

ここでは変数選択法 (3) の修正を考える．まず D_i を大きさの順に並べ

$$D_{i_1}\geq D_{i_2}\geq\cdots\geq D_{i_p}$$

それに対応する変数 $\{x_{i_1},\ldots,x_{i_p}\}$ の中から初期変数として k 個の変数 $\{x_{i_1},\ldots,x_{i_k}\}$ を選ぶ．この $k(\leq p)$ 個の変数に対応する群間平方和積和行列 B_*，群内平方和積和行列 W_* を求め，ここで，変数選択法 (3) を適用して初期変数の中から変数を選び出す．

このように修正することで，まずはじめに，各周辺ごとであらかじめ重要な変数を選んでおき，その中から独立性の仮定を外した変数選択により，相関の高い変数を除くことによって真に重要な変数を選び出すことが可能になる．またあらかじめ，p よりも小さい k を選ぶことにより標本数と次元数の問題を回避できることになる．

4.3　モデル評価基準

このようにして選ばれた変数の組の中で，どの変数の組が最適な判別かを調べるために誤判別率が用いられる．ここでは別な評価基準として変数選択モデルに基づくモデル評価基準を導入する．

4.3.1　多変量正規モデルを想定した場合

p 次元変数 $\boldsymbol{x}=(x_1,\ldots,x_p)'$ に対して，

$$G_g:N_p(\boldsymbol{\mu}^{(i)},\Sigma),\qquad i=g,\cdots,q$$

とする．G_g からの大きさ n_g の標本にもとづく，群間平方和積和行列を B，群内平方和積和行列を W，全平方和積和行列を $T=B+W$ とする．全標本数を $n=n_1+\ldots+n_q$ とする．最初の k 個の変数の $\boldsymbol{x}_1=(x_1,\cdots,x_k)'$ が判別に関しての全情報をもち，残りの $(p-k)$ 個の変数 $\boldsymbol{x}_2=(x_{k+1},\cdots,x_p)'$ は追加情報をもたないというモデルを

$$M_{1\cdots k}:\boldsymbol{\mu}_{2\cdot 1}^{(1)}=\cdots=\boldsymbol{\mu}_{2\cdot 1}^{(q)}$$

と定義する．ここに

$$\boldsymbol{\mu}^{(i)} = \begin{pmatrix} \boldsymbol{\mu}_1^{(i)} \\ \boldsymbol{\mu}_2^{(i)} \end{pmatrix}, \quad \begin{matrix} \boldsymbol{\mu}_1^{(i)} : k \times 1, \\ \boldsymbol{\mu}_2^{(i)} : (p-k) \times 1, \end{matrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \begin{matrix} \Sigma_{11} : k \times k, \quad \Sigma_{12} : k \times (p-k), \\ \Sigma_{21} : (p-k) \times k, \quad \Sigma_{22} : (p-k) \times (p-k), \end{matrix}$$

$$\boldsymbol{\mu}_{2\cdot 1}^{(i)} = \boldsymbol{\mu}_2^{(i)} - \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\mu}_1^{(i)}$$

である．

このとき，モデル $M_{1\cdots k}$ に対する AIC 基準は

$$AIC_M = -n \log \frac{|W_{22\cdot 1}|}{|T_{22\cdot 1}|} + n \log \left| \frac{1}{n} W \right| + np(1 + \log 2\pi) + 2 \left\{ qk + p - k + \frac{1}{2} p(p+1) \right\}$$

で与えられる (Fujikoshi [7])．ここに

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},\ W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix},\ T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix},\ B_{12}, W_{12}, T_{12} : k \times (p-k),$$

$$W_{22\cdot 1} = W_{22} - W_{21}W_{11}^{-1}W_{12}, \qquad T_{22\cdot 1} = T_{22} - T_{21}T_{11}^{-1}T_{12}.$$

ここで n と p が接近または $n \leq p$ の場合には $|\frac{1}{n}W| = 0$ の値をとる場合がある．または $|\frac{1}{n}W|$ は全てのモデルで共通なため，これを除いた次の基準量

$$AIC_M^* = -n \log \frac{|W_{22\cdot 1}|}{|T_{22\cdot 1}|} + np(1 + \log 2\pi) + 2 \left\{ qk + p - k + \frac{1}{2} p(p+1) \right\}$$

を用いることにする．

一般に，逐次法の各段階で，上記の冗長性仮説モデルを考え，AIC_M^* を求めて，各段階で求められた変数の組のよさを評価することができる．

4.3.2 独立 2 項モデルを想定した場合

ここでは，p 次元変数 $\boldsymbol{x} = (x_1, \ldots, x_p)'$ の各成分は互いに独立で，群 G_g のもとでの $\boldsymbol{x}$ を $\boldsymbol{x}^{(g)} = (x_1^{(g)}, \ldots, x_p^{(g)})'$ と表す．さらに

$$x_1^{(g)} \sim B(1, \theta_1^{(g)}) \ , \ldots , \ x_p^{(g)} \sim B(1, \theta_p^{(g)})$$

とする．

$$y_1^{(g)} = \sum_{j=1}^{n_g} x_{1j}^{(g)}, \ \ldots , \ y_p^{(g)} = \sum_{j=1}^{n_g} x_{pj}^{(g)}$$

とおく．このとき

$$y_1^{(g)} \sim B(n_g, \theta_1^{(g)}) \ , \ldots , \ y_p^{(g)} \sim B(n_g, \theta_p^{(g)})$$

である．$\boldsymbol{x}_1 = (x_1, \cdots, x_k)'$ が判別に関しての全情報をもち，$\boldsymbol{x}_2 = (x_{k+1}, \cdots, x_p)'$ は追加情報をもたないという仮説を

$$\tilde{M}_{1\cdots k} : \boldsymbol{\theta}_2^{(1)} = \ldots = \boldsymbol{\theta}_2^{(q)}$$

と定義する．ここに，

$$\boldsymbol{\theta}^{(g)} = \begin{pmatrix} \theta_1^{(g)} \\ \vdots \\ \theta_p^{(g)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\theta}_1^{(g)} \\ \boldsymbol{\theta}_2^{(g)} \end{pmatrix}, \quad \boldsymbol{\theta}_1^{(g)}; k \times 1$$

と分割すると，AIC 基準は次のように与えられる．

$$\begin{aligned} AIC_B &= -2\max_{M_{1\dots k}} \log L(\Theta) + 2\,(\text{独立パラメーター数}) \\ &= -2\sum_{g=1}^{q}\sum_{j=1}^{p}\log\binom{n_g}{y_j^{(g)}} - 2\sum_{g=1}^{q}\sum_{j=1}^{k}\left\{y_j^{(g)}\log\hat{\theta}_j^{(g)} + (n_i - y_j^{(g)})\log(1-\hat{\theta}_j^{(g)})\right\} \\ &\quad - 2\sum_{j=k+1}^{p}\left\{y_j^{(\cdot)}\log\hat{\theta}_j^{(\cdot)} + (n - y_j^{(\cdot)})\log(1-\hat{\theta}_j^{(\cdot)})\right\} + 2\,\{qk + p - k\} \end{aligned}$$

と表せる．ここに，

$$y_j^{(\cdot)} = \sum_{i=1}^{q} y_j^{(i)}, \quad \hat{\theta}_j^{(i)} = \frac{1}{n_i}y_j^{(i)},\ (i = 1,\dots,q, j = 1,\dots,k), \quad \hat{\theta}_j^{(\cdot)} = \frac{1}{n}y_j^{(\cdot)},\ (j = k+1,\dots,p)$$

である．

多変量正規性を仮定した場合と同様に，各変数の組に対して，モデル $\tilde{M}$ を考え，AIC_B を用いてその変数の組の重要度を評価することができる．なお，条件付 D_i 統計量を用いることは暗に変数間の独立性は成り立たないことを仮定しているが，AIC による評価基準は独立性を仮定したものになっている．

5　応用

付録の Table 5 で与えられるトウモロコシのフィンガープリントデータに対して，変数選択法 (1)～(4) を適用し，最適な変数の組を見い出すことを試みる．なお，ここでは変数 x_j を $x(j)$ と表している．

5.1　変数選択方法 (1)，(2) の結果と考察

Wilber et al. [1] による変数選択法 (1)，(2) ではそれぞれ次のような変数が選ばれている．

$$\begin{aligned} J_1 &= \{x(13), x(54), x(34)\} \\ J_2 &= \{x(9), x(12), x(13), x(16), x(19), x(32), x(34), x(36), x(39), x(43), x(45), x(46), x(48), \\ &\qquad x(49), x(51), x(53), x(54), x(55), x(84)\} \end{aligned}$$

そこで，D_i の大きい変数から逐次変数を選ぶ方法を考え，選ばれた変数の組に対して評価基準を求めたのが Table 1 である．

この表における，それぞれの項目は次のとおりである．

変数　　：　選ばれた変数
AIC_B　：　2 項分布を仮定した場合の AIC
CD_0　　：　正準判別法による判別的中率
ML_0　　：　最大尤度法による判別的中率
CD_1　　：　CD_0 の交差確認 (CV) 版
ML_1　　：　ML_0 の交差確認 (CV) 版

また，これらの値を求めるにあたって以下の処理あるいは修正を行っている．

(1) 正規分布を仮定した場合の AIC において
　　標本数と次元数が接近してくると，計算が不定となったため，その欄は掲載しなかった．
(2) AIC_B において $\hat{\theta}_{ij} = 0$ のとき $\log\hat{\theta}_{ij} = 0$ とした．

選ばれた変数の組 J_2 は Step19 に対応している．変数の組 J_1, J_2 よりもよい変数の組が多くあることを注意したい．

これらの結果から次のことがわかった．

Table 1　独立性を仮定した場合の変数選択

Step	変数	AIC_B	CD_0	ML_0	CD_1	ML_1
1	$\boldsymbol{x}(13)$	1509.17	75.28	75.28	75.28	75.28
2	$\boldsymbol{x}(54)$	1443.44	66.29	66.29	66.29	62.92
3	$\boldsymbol{x}(34)$	1395.27	74.16	76.40	74.16	76.40
4	$\boldsymbol{x}(48)$	1369.60	75.28	75.28	75.28	64.04
5	$\boldsymbol{x}(32)$	1337.82	77.53	79.78	77.53	71.91
6	$\boldsymbol{x}(9)$	1304.96	80.90	83.15	68.54	80.90
7	$\boldsymbol{x}(36)$	1282.93	80.90	83.15	71.91	83.15
8	$\boldsymbol{x}(12)$	1258.46	84.27	82.02	79.78	77.53
9	$\boldsymbol{x}(84)$	1236.40	84.27	89.89	84.27	85.39
10	$\boldsymbol{x}(19)$	1219.41	84.27	88.76	84.27	84.27
11	$\boldsymbol{x}(39)$	1205.28	85.39	89.89	83.15	85.39
12	$\boldsymbol{x}(45)$	1193.12	84.27	88.76	82.02	86.52
13	$\boldsymbol{x}(49)$	1180.95	84.27	92.13	84.27	87.64
14	$\boldsymbol{x}(55)$	1168.78	88.76	94.38	83.15	88.76
15	$\boldsymbol{x}(46)$	1155.48	94.38	93.26	86.52	89.89
16	$\boldsymbol{x}(16)$	1146.56	92.13	94.38	85.39	92.13
17	$\boldsymbol{x}(51)$	1137.64	93.26	93.26	91.01	92.13
18	$\boldsymbol{x}(53)$	1128.52	93.26	93.26	92.13	92.13
19	$\boldsymbol{x}(43)$	1117.02	93.26	93.26	93.26	92.13
20	$\boldsymbol{x}(14)$	1103.73	95.51	94.38	94.38	92.13
21	$\boldsymbol{x}(40)$	1093.21	95.51	95.51	94.38	92.13
22	$\boldsymbol{x}(22)$	1087.44	95.51	94.38	94.38	93.26
23	$\boldsymbol{x}(28)$	1081.66	95.51	94.38	94.38	93.26
24	$\boldsymbol{x}(70)$	1075.89	95.51	93.26	93.26	93.26
25	$\boldsymbol{x}(64)$	1070.11	95.51	93.26	94.38	93.26
26	$\boldsymbol{x}(71)$	1064.34	95.51	94.38	95.51	93.26
27	$\boldsymbol{x}(77)$	1058.57	95.51	94.38	95.51	93.26
28	$\boldsymbol{x}(42)$	1052.34	94.38	95.51	94.38	95.51
29	$\boldsymbol{x}(7)$	1046.95	94.38	95.51	94.38	95.51
30	$\boldsymbol{x}(78)$	1041.57	94.38	95.51	94.38	95.51
31	$\boldsymbol{x}(50)$	1035.84	94.38	95.51	94.38	95.51
32	$\boldsymbol{x}(30)$	1027.98	96.63	96.63	94.38	95.51
33	$\boldsymbol{x}(56)$	1025.27	96.63	96.63	94.38	95.51
34	$\boldsymbol{x}(8)$	1022.56	100.00	96.63	98.88	95.51
35	$\boldsymbol{x}(63)$	1019.85	100.00	96.63	100.00	95.51
36	$\boldsymbol{x}(67)$	1017.14	98.88	97.75	97.75	96.63
37	$\boldsymbol{x}(75)$	1014.42	98.88	97.75	97.75	96.63
38	$\boldsymbol{x}(21)$	1011.24	100.00	97.75	98.88	96.63
39	$\boldsymbol{x}(73)$	1007.74	100.00	97.75	98.88	96.63
40	$\boldsymbol{x}(18)$	1004.32	100.00	98.88	97.75	97.75
41	$\boldsymbol{x}(37)$	1001.89	100.00	98.88	98.88	97.75
42	$\boldsymbol{x}(35)$	996.20	100.00	98.88	97.75	97.75
43	$\boldsymbol{x}(24)$	993.71	100.00	98.88	98.88	98.88
44	$\boldsymbol{x}(17)$	992.73	100.00	98.88	100.00	98.88
45	$\boldsymbol{x}(11)$	993.00	100.00	98.88	100.00	98.88
46	$\boldsymbol{x}(27)$	993.27	100.00	98.88	100.00	98.88
47	$\boldsymbol{x}(41)$	993.53	98.88	98.88	98.88	98.88
48	$\boldsymbol{x}(61)$	993.80	98.88	98.88	98.88	98.88
49	$\boldsymbol{x}(82)$	994.07	98.88	98.88	98.88	98.88
50	$\boldsymbol{x}(4)$	994.34	98.88	98.88	98.88	98.88
51	$\boldsymbol{x}(58)$	994.60	98.88	98.88	98.88	98.88
52	$\boldsymbol{x}(81)$	994.87	98.88	98.88	98.88	98.88
53	$\boldsymbol{x}(10)$	994.82	98.88	98.88	98.88	97.75
54	$\boldsymbol{x}(47)$	995.27	98.88	98.88	98.88	97.75
55	$\boldsymbol{x}(66)$	995.73	100.00	98.88	98.88	97.75
56	$\boldsymbol{x}(3)$	995.28	100.00	98.88	100.00	97.75
57	$\boldsymbol{x}(52)$	995.15	100.00	98.88	100.00	97.75
58	$\boldsymbol{x}(15)$	995.32	100.00	98.88	100.00	98.88
59	$\boldsymbol{x}(65)$	994.68	100.00	98.88	100.00	97.75
60	$\boldsymbol{x}(44)$	995.04	100.00	98.88	100.00	97.75
61	$\boldsymbol{x}(80)$	997.00	100.00	98.88	100.00	97.75
62	$\boldsymbol{x}(26)$	998.30	100.00	98.88	100.00	97.75
63	$\boldsymbol{x}(62)$	999.60	100.00	98.88	100.00	97.75
64	$\boldsymbol{x}(5)$	1001.00	100.00	98.88	100.00	98.88
65	$\boldsymbol{x}(6)$	1004.17	100.00	98.88	100.00	97.75
66	$\boldsymbol{x}(20)$	1007.34	100.00	98.88	100.00	97.75
67	$\boldsymbol{x}(38)$	1010.51	100.00	98.88	100.00	97.75
68	$\boldsymbol{x}(59)$	1013.68	100.00	98.88	100.00	97.75
69	$\boldsymbol{x}(68)$	1016.85		98.88		97.75
70	$\boldsymbol{x}(83)$	1020.02		98.88		97.75
71	$\boldsymbol{x}(2)$	1023.19		98.88		97.75
72	$\boldsymbol{x}(23)$	1026.36		98.88		97.75
73	$\boldsymbol{x}(25)$	1029.53		98.88		97.75
74	$\boldsymbol{x}(31)$	1032.70		98.88		97.75
75	$\boldsymbol{x}(72)$	1035.87		98.88		97.75
76	$\boldsymbol{x}(74)$	1039.04		98.88		97.75
77	$\boldsymbol{x}(76)$	1042.21		98.88		97.75
78	$\boldsymbol{x}(79)$	1045.38		98.88		97.75
79	$\boldsymbol{x}(69)$	1048.01		98.88		97.75
80	$\boldsymbol{x}(1)$	1051.23		98.88		97.75
81	$\boldsymbol{x}(29)$	1054.46		98.88		97.75
82	$\boldsymbol{x}(33)$	1057.69		98.88		97.75
83	$\boldsymbol{x}(60)$	1061.99		98.88		97.75
84	$\boldsymbol{x}(57)$	1067.49		98.88		97.75

- AIC_B について
  Step44 において最小の AIC_B をとり，また判別的中率においても CD_0，CD_1 とともに 100%になっている．

- CD_0 について
  Step34 において判別的中率が 100%になっている．また，それ以降の Step35, Step38～46, Step55～68 においても 100%の判別を実現している．しかし，Step69 以降において共分散行列が特異になり，固有値や固有ベクトルが数値的に不安定なるため計算が困難となった．

- ML_0 について
  判別的中率が 100%になる変数の組は現れなかったが，90%以上となる変数の組は CD_0 より少ない変数の組; Step13 が求められている．

- CD_1 について
  Step35 において 100%の判別的中率を実現している．また，それ以降の Step44～46，Step56～68 に

Table 2　独立性を外した場合の変数選択

Step	変数	AIC_B	CD_0	ML_0	CD_1	ML_1
1	$\boldsymbol{x}(13)$	1509.17		75.28		75.28
2	$\boldsymbol{x}(54)$	1443.44		66.29		62.92
3	$\boldsymbol{x}(48)$	1417.78	75.28	75.28	75.28	71.91
4	$\boldsymbol{x}(34)$	1369.60	75.28	75.28	75.28	64.04
5	$\boldsymbol{x}(19)$	1352.61	80.90	80.90	80.90	78.65
6	$\boldsymbol{x}(40)$	1342.09	82.02	80.90	69.66	80.90
7	$\boldsymbol{x}(39)$	1327.97	82.02	82.02	70.79	82.02
8	$\boldsymbol{x}(45)$	1315.80	84.27	84.27	79.78	76.40
9	$\boldsymbol{x}(36)$	1293.77	84.27	85.39	79.78	74.16
10	$\boldsymbol{x}(50)$	1288.05	84.27	85.39	83.15	74.16
11	$\boldsymbol{x}(14)$	1274.76	84.27	86.52	78.65	83.15
12	$\boldsymbol{x}(16)$	1265.83	86.52	87.64	79.78	85.39
13	$\boldsymbol{x}(56)$	1263.12	86.52	87.64	79.78	85.39
14	$\boldsymbol{x}(30)$	1255.26	88.76	87.64	79.78	85.39
15	$\boldsymbol{x}(46)$	1241.96	89.89	88.76	85.39	85.39
16	$\boldsymbol{x}(18)$	1238.54	89.89	89.89	84.27	88.76
17	$\boldsymbol{x}(22)$	1232.77	92.13	92.13	87.64	91.01
18	$\boldsymbol{x}(32)$	1200.98	92.13	93.26	87.64	93.26
19	$\boldsymbol{x}(80)$	1202.94	93.26	93.26	88.76	89.89
20	$\boldsymbol{x}(44)$	1203.30	93.26	93.26	88.76	89.89
21	$\boldsymbol{x}(9)$	1170.44	95.51	94.38	88.76	92.13
22	$\boldsymbol{x}(12)$	1145.97	94.38	93.26	92.13	91.01
23	$\boldsymbol{x}(84)$	1123.92	93.26	96.63	91.01	93.26
24	$\boldsymbol{x}(49)$	1111.75	94.38	95.51	89.89	92.13
25	$\boldsymbol{x}(55)$	1099.58	93.26	94.38	91.01	92.13
26	$\boldsymbol{x}(51)$	1090.66	94.38	95.51	93.26	94.38
27	$\boldsymbol{x}(53)$	1081.54	94.38	95.51	94.38	93.26
28	$\boldsymbol{x}(43)$	1070.04	96.63	96.63	96.63	95.51
29	$\boldsymbol{x}(28)$	1064.26	96.63	96.63	96.63	95.51
30	$\boldsymbol{x}(70)$	1058.49	96.63	95.51	96.63	95.51
31	$\boldsymbol{x}(64)$	1052.72	96.63	96.63	96.63	95.51
32	$\boldsymbol{x}(71)$	1046.94	96.63	96.63	96.63	95.51
33	$\boldsymbol{x}(77)$	1041.17	96.63	96.63	96.63	95.51
34	$\boldsymbol{x}(42)$	1034.94	97.75	97.75	95.51	96.63
35	$\boldsymbol{x}(7)$	1029.56	97.75	97.75	95.51	96.63
36	$\boldsymbol{x}(78)$	1024.17	97.75	97.75	95.51	96.63
37	$\boldsymbol{x}(8)$	1021.46	100.00	97.75	98.88	97.75
38	$\boldsymbol{x}(63)$	1018.75	100.00	97.75	100.00	97.75
39	$\boldsymbol{x}(67)$	1016.04	100.00	98.88	98.88	98.88
40	$\boldsymbol{x}(75)$	1013.32	100.00	98.88	97.75	98.88
41	$\boldsymbol{x}(21)$	1010.14	100.00	98.88	98.88	98.88
42	$\boldsymbol{x}(73)$	1006.64	100.00	98.88	98.88	97.75
43	$\boldsymbol{x}(37)$	1004.21	100.00	98.88	98.88	97.75
44	$\boldsymbol{x}(35)$	998.52	100.00	98.88	98.88	97.75
45	$\boldsymbol{x}(24)$	996.02	100.00	98.88	98.88	98.88
46	$\boldsymbol{x}(17)$	995.05	100.00	98.88	98.88	98.88
47	$\boldsymbol{x}(11)$	995.32	100.00	98.88	98.88	98.88
48	$\boldsymbol{x}(27)$	995.58	100.00	98.88	100.00	98.88
49	$\boldsymbol{x}(41)$	995.85	100.00	98.88	98.88	97.75
50	$\boldsymbol{x}(61)$	996.12	100.00	98.88	98.88	97.75
51	$\boldsymbol{x}(82)$	996.39	100.00	98.88	98.88	97.75
52	$\boldsymbol{x}(4)$	996.65	100.00	98.88	98.88	97.75
53	$\boldsymbol{x}(58)$	996.92	100.00	98.88	100.00	97.75
54	$\boldsymbol{x}(81)$	997.19	100.00	98.88	100.00	97.75
55	$\boldsymbol{x}(10)$	997.14	100.00	97.75	98.88	97.75
56	$\boldsymbol{x}(47)$	997.59	100.00	97.75	100.00	97.75
57	$\boldsymbol{x}(66)$	998.04	100.00	97.75	100.00	97.75
58	$\boldsymbol{x}(3)$	997.60	100.00	98.88	100.00	97.75
59	$\boldsymbol{x}(52)$	997.46	100.00	98.88	100.00	97.75
60	$\boldsymbol{x}(15)$	997.64	100.00	98.88	100.00	97.75
61	$\boldsymbol{x}(65)$	997.00	100.00	98.88	100.00	97.75
62	$\boldsymbol{x}(26)$	998.30	100.00	98.88	100.00	97.75
63	$\boldsymbol{x}(62)$	999.60	100.00	98.88	100.00	97.75
64	$\boldsymbol{x}(5)$	1001.00	100.00	98.88	100.00	98.88
65	$\boldsymbol{x}(6)$	1004.17	100.00	98.88	100.00	97.75
66	$\boldsymbol{x}(20)$	1007.34	100.00	98.88	100.00	97.75
67	$\boldsymbol{x}(38)$	1010.51	100.00	98.88	100.00	97.75
68	$\boldsymbol{x}(59)$	1013.68	100.00	98.88	100.00	97.75
69	$\boldsymbol{x}(68)$	1016.85		98.88		97.75
70	$\boldsymbol{x}(83)$	1020.02		98.88		97.75
71	$\boldsymbol{x}(2)$	1023.19		98.88		97.75
72	$\boldsymbol{x}(23)$	1026.36		98.88		97.75
73	$\boldsymbol{x}(25)$	1029.53		98.88		97.75
74	$\boldsymbol{x}(31)$	1032.70		98.88		97.75
75	$\boldsymbol{x}(72)$	1035.87		98.88		97.75
76	$\boldsymbol{x}(74)$	1039.04		98.88		97.75
77	$\boldsymbol{x}(76)$	1042.21		98.88		97.75
78	$\boldsymbol{x}(79)$	1045.38		98.88		97.75
79	$\boldsymbol{x}(69)$	1048.01		98.88		97.75
80	$\boldsymbol{x}(1)$	1051.23		98.88		97.75
81	$\boldsymbol{x}(29)$	1054.46		98.88		97.75
82	$\boldsymbol{x}(33)$	1057.69		98.88		97.75
83	$\boldsymbol{x}(60)$	1061.99		98.88		97.75
84	$\boldsymbol{x}(57)$	1067.49		98.88		97.75

おいても 100%になっている．しかし，CD_0 と同様に Step69 以降において共分散行列が特異になり，固有値や固有ベクトルが数値的に不安定なるため計算が困難となった．

- ML_1 について
  判別的中率が 100%となる変数の組は ML_0 と同様に現れなかったが，90%以上となる変数の組は CD_0 よりの少ない変数の組，Step16 において現れている．

5.2　変数選択方法 (3) の結果

ここでは独立性の仮定を外した場合の変数選択法 (3) の結果を Table 2 に与えている．また，選ばれた変数の組とその変数の組に対する AIC と判別的中率も与えている．

なお，数値計算を行うにあたって先ほどの点に加えて以下の処理あるいは修正を行った．

(3) 変数選択において，$D(i|i_1,\ldots,i_k)$ を求める場合において分子の値がとても小さくなるときは値が不定となるため，$0/0 \to 1$ として計算を行った．

これらの結果から次のことがわかった．

- AIC_B について
  Step46 において最小の AIC_B をとり，また判別に関しても CD_0 で 100%の判別的中率になっている．また，判別的中率が CD_0 と CD_1 の両方で 100%となる変数の組のなかで，AIC_B が最小になるものは Step48 である．

- CD_0 について
  Step37 において判別的中率が 100%になっている．また，それ以降の Step37～68 においても 100%である．しかし，Step69 以降において共分散行列が特異になり，固有値や固有ベクトルが数値的に不安定なるため計算が困難となった．

- ML_0 について
  判別的中率が 100%になる変数の組は現れなかったが，90%以上となるものについては CD_0 と同時に，Step17 において現れている．

- CD_1 について
  Step38 において判別的中率が 100%になっている．また，それ以降の Step48, Step53, Step54, Step56～68 においても 100%になっている．しかし，CD_0 と同様に Step69 以降において共分散行列が特異になり，固有値や固有ベクトルが数値的に不安定なるため計算が困難となった．

- ML_1 について
  判別的中率が 100%になる変数の組は ML_0 と同様に現れなかったが，90%以上となるものについては CD_0 よりの少ない変数の組，Step17 において現れている．

また全体的に判別は周辺からの結果の方が，1 つぐらいの変数が少ない場合において 100%や 90%以上の結果を得ている．

5.3 変数選択方法 (4) の結果

ここでは，あらかじめ初期の変数を選び，それらの中から方法 (3) に従って変数を逐次選ぶ方法を適用する．このとき，初期の変数の選び方は，統計量 D_i の大きい方から選んだ．具体的には，初期変数を 1 個から 84 個まで増やしていき，それぞれの場合で変数選択法 (3) を適用した．このとき，次のような変数の組に注目した．それは，初期の変数を $\{x_1, \ldots, x_k\}$ としたときの選ばれた変数を順に $\{x^k_{(1)}, \ldots, x^k_{(k)}\}$ と表し，初期の変数に 1 個追加した $\{x_1, \ldots, x_k, x_{k+1}\}$ と選ばれた変数を順に $\{x^{k+1}_{(1)}, \ldots, x^{k+1}_{(k)}, x^{k+1}_{(k+1)}\}$ とする．ここで，それぞれの結果の最初から k 個までの変数すなわち $\{x^k_{(1)}, \ldots, x^k_{(k)}\}$ と $\{x^{k+1}_{(1)}, \ldots, x^{k+1}_{(k)}\}$ に注目した．

ここで追加された変数 x_{k+1} が他の変数と関連性が無ければ独立性の仮定を外した場合における変数選択でも選ばれる順番に変化はなく，最初から k 番目までに選択された変数に違が生じないであろう．しかし，x_{k+1} が他の変数と関連性がある場合にはそれぞれの変数の選ばれる順番に影響を及ぼす．つまり，そのような変数の組が独立性の仮定を外した場合での意味のある変数の組となる．後者の変数の組は Table 3 で与えてある．その表で用いられている $S_1, \ldots, S_9$ は共通に選択される変数の組であって，変数の番号だけを以下のように用いて与えられる．

$$S_1 = \{13, 54, 48, 34\},\ S_2 = \{19, 46, 39, 45, 36\},\ S_3 = \{55, 9\},\ S_4 = \{32, 49, 84, 12\},\ S_5 = \{22, 14, 16\},$$
$$S_6 = \{12, 9, 53, 43\},\ S_7 = \{32, 9, 84, 49, 55, 46, 43\},\ S_8 = \{56, 30, 46\},\ S_9 = \{18, 22, 32\}.$$

また，J_1 における最後の 21 個の変数をまとめて S_* と表しているが，次のように与えられる．

$$S_* = \{21, 73, 37, 35, 24, 17, 11, 27, 41, 61, 82, 4, 58, 81, 10, 47, 66, 3, 52, 15, 65\}$$

初期変数として 40 個用いたときの AIC と判別的中率が Table 4 である．これらの結果から次のことがわかった．

Table 3　選ばれた変数

J4	S_1
J7	S_1,36,9,32
J8	S_1,12,9,32,36
J10	S_1,19,9,32,36,84,12
J11	S_1,19,9,32,36,39,12,84
J12	S_1,19,9,45,32,84,39,12,36
J13	S_1,19,9,45,49,12,84,36,32,39
J15	S_1,S_2,S_3,S_4
J17	S_1,S_2,S_3,S_4,51,16
J18	S_1,S_2,S_3,S_4,53,16,51
J19	S_1,S_2,S_3,43,84,12,32,51,49,16,53
J21	S_1,S_2,S_3,53,84,12,14,46,32,49,16,51,43
J22	S_1,S_2,S_5,S_6,32,84,49,55,46,51
J25	S_1,S_2,S_5,S_6,64,32,84,49,55,46,51,28,70
J26	S_1,S_2,S_5,12,71,51,53,64,S_7,28,70
J31	S_1,S_2,S_5,12,64,51,53,70,S_7,22,28,71,77,42,7,78
J33	S_1,S_2,S_5,S_8,51,28,32,9,12,84,49,55,53,43,22,70,64,71,77,42,7,78
J40	S_1,S_2,S_5,S_8,S_9,9,12,84,49,55,51,53,43,28,70,64,71,77,42,7,78,8,63,67,75,21,73
J61	S_1,S_2,S_5,S_8,S_9,80,44,9,12,84,49,55,51,53,43,28,70,64,71,77,42,7,78,8,63,67,75,S_*

- AIC_B について
  Step39 において最小の AIC_N をとり，また判別においても CD_0 について判別的中率が 100%になっている．また，CD_0 と CD_1 の両方で判別的中率が 100%になる変数の組のなかで，AIC_N が最小になるものは Step36 である．

- AIC_B について
  Step40 において最小の AIC_B をとり，また CD_0 について判別的中率が 100%になる変数の組が現れている．また，CD_0 と CD_1 の両方で 100%となる変数の組のなかで，AIC_B が最小になるものは Step36 である．

- CD_0 について
  Step35 において判別的中率が 100%となっている．また，それ以降の Step36～40 においても 100%になっている．

- ML_0 について
  判別的中率が 100%となる変数の組は現れなかったが，90%以上となるものについては CD_0 と同時に，Step17 において現れている．

- CD_1 について
  Step36 において判別的中率が 100%になっている．

- ML_1 について
  判別的中率が 100%となる変数の組は ML_0 と同様に現れなかったが，90%以上となるものについては CD_0 よりの少ない変数の組，Step17 において現れている．

Table 4　初期変数を 40 個選んだ場合の変数選択法

Step	変数	AIC_N	AIC_B	CD_0	ML_0	CD_1	ML_1
1	$\boldsymbol{x}(13)$	9290.62	1028.54	75.28	75.28	75.28	75.28
2	$\boldsymbol{x}(54)$	9234.33	962.81	66.29	66.29	66.29	62.92
3	$\boldsymbol{x}(48)$	9203.84	937.14	75.28	75.28	75.28	71.91
4	$\boldsymbol{x}(34)$	9175.47	888.97	75.28	75.28	75.28	64.04
5	$\boldsymbol{x}(19)$	9167.45	871.97	80.90	80.90	80.90	78.65
6	$\boldsymbol{x}(40)$	9161.08	861.46	82.02	80.90	69.66	80.90
7	$\boldsymbol{x}(39)$	9162.63	847.34	82.02	82.02	70.79	82.02
8	$\boldsymbol{x}(45)$	9161.13	835.17	84.27	84.27	79.78	76.40
9	$\boldsymbol{x}(36)$	9163.78	813.14	84.27	85.39	79.78	74.16
10	$\boldsymbol{x}(50)$	9166.59	807.41	84.27	85.39	83.15	74.16
11	$\boldsymbol{x}(14)$	9166.57	794.12	84.27	86.52	78.65	83.15
12	$\boldsymbol{x}(16)$	9160.01	785.20	86.52	87.64	79.78	85.39
13	$\boldsymbol{x}(56)$	9165.74	782.49	86.52	87.64	79.78	85.39
14	$\boldsymbol{x}(30)$	9165.19	774.63	88.76	87.64	79.78	85.39
15	$\boldsymbol{x}(46)$	9163.94	761.33	89.89	88.76	85.39	85.39
16	$\boldsymbol{x}(18)$	9162.28	757.91	89.89	89.89	84.27	88.76
17	$\boldsymbol{x}(22)$	9145.51	752.14	92.13	92.13	87.64	91.01
18	$\boldsymbol{x}(32)$	9140.51	720.35	92.13	93.26	87.64	93.26
19	$\boldsymbol{x}(9)$	9136.75	687.49	94.38	92.13	89.89	89.89
20	$\boldsymbol{x}(12)$	9129.20	663.02	94.38	93.26	91.01	89.89
21	$\boldsymbol{x}(84)$	9131.63	640.96	93.26	94.38	89.89	93.26
22	$\boldsymbol{x}(49)$	9123.08	628.80	94.38	94.38	89.89	92.13
23	$\boldsymbol{x}(55)$	9116.70	616.63	93.26	94.38	91.01	92.13
24	$\boldsymbol{x}(51)$	9096.82	607.71	94.38	94.38	93.26	93.26
25	$\boldsymbol{x}(53)$	9095.59	598.59	94.38	94.38	94.38	93.26
26	$\boldsymbol{x}(43)$	9089.14	587.09	96.63	96.63	96.63	95.51
27	$\boldsymbol{x}(28)$	9093.40	581.31	96.63	96.63	96.63	95.51
28	$\boldsymbol{x}(70)$	9088.73	575.54	96.63	95.51	95.51	94.38
29	$\boldsymbol{x}(64)$	9091.07	569.77	96.63	95.51	95.51	94.38
30	$\boldsymbol{x}(71)$	9082.90	563.99	96.63	96.63	95.51	95.51
31	$\boldsymbol{x}(77)$	9084.37	558.22	96.63	96.63	95.51	95.51
32	$\boldsymbol{x}(42)$	9074.70	551.99	96.63	97.75	95.51	96.63
33	$\boldsymbol{x}(7)$	9072.03	546.60	96.63	97.75	95.51	96.63
34	$\boldsymbol{x}(78)$	9073.85	541.22	96.63	97.75	95.51	96.63
35	$\boldsymbol{x}(8)$	9044.43	538.51	100.00	97.75	98.88	96.63
36	$\boldsymbol{x}(63)$	9039.12	535.80	100.00	97.75	100.00	96.63
37	$\boldsymbol{x}(67)$	9034.51	533.08	100.00	98.88	98.88	97.75
38	$\boldsymbol{x}(75)$	9033.73	530.37	100.00	98.88	97.75	97.75
39	$\boldsymbol{x}(21)$	9030.58	527.19	100.00	98.88	98.88	97.75
40	$\boldsymbol{x}(73)$	9699.11	523.69	100.00	98.88	97.75	97.75

以下のように本論文で提案された変数選択法 (3)，(4) により，誤判別確率が従来の方法と比べ小さくなる変数の組が見い出されている．一方，できるだけ少ない変数の組で，群間の違いを説明することも重要である．なお，ここで提案された方法ではないが，次の変数の組

$$J_* = \{x(8), x(9), x(13), x(16), x(22), x(24), x(27), x(32), x(35), x(37), x(42), x(48), x(54), x(55), x(57), x(58), x(60), x(67), x(70), x(71), x(84)\}$$

で，どの判別法でも的中率 100%の判別が可能であることを注意したい．J_* の変数の数は 21 個である．ここで考えられた変数選択法で，このような性質をもつ変数の組は，たとえば変数選択法 (3) で与えられた 38 個である．J_* が自動的に選ばれる変数選択法にも興味がある．

6　結論

本論文では高次元 2 値データの判別問題における変数選択問題を扱った．Wilber et al. [1] は，変数間の独立性を想定して変数選択法 (1)，(2) を提案している．本論文では独立性の仮定を取り除いた変数選択法 (3) とその修正版 (4) を提案した．また，選ばれた変数の組の妥当性を測る基準として誤判別率に基づく評価基準と

は異なった，変数の冗長性モデルに基づく評価基準を提案した．

実際の DNA フィンガープリントデータに提案した変数選択法とモデル評価基準を適用した．従来の方法では，判別的中率が 100%になる変数の組を見つけることができなかった．しかし，提案した方法により，選ばれた変数の数は多くなるが，判別的中率が 100%になる変数の組を見つけることができた．また，判別的中率に基づく最適な変数の組とモデル評価基準に基づく最適な変数の組は必ずしも一致しないことが確認された．

高次元 2 値データの場合，判別的中率が 100%近くになると，選ばれた変数の組は多くの変数を含んでいる傾向が見られる．一方，ある程度の判別的中率を持った変数の組は，比較的少ない変数で目的を達することができる．どちらを優先するかは，解析の目的に依存する．しかし，本論文では提案した独立性を前提としない変数選択法はより一般的で，これを用いて分析することが求められる．

謝辞

査読者の方には貴重なコメントを頂きました．ここに記して，お礼申し上げます．また本研究は，理工学研究所，共同研究第 2 類「多変量高次元データ解析の理論と応用」(2006 年度) から研究助成を受けております．

参考文献

[1] Wilbur, J. D., Ghosh, J. K., Nakatsu, C. H., Brouder, S. M., and Doerge, R. W.: Variable selection in high-dimensional multivariate binary data with application to the analysis of microbial community DNA fingerprints. *Biometrics*, **58**, 378–386. (2002)

[2] Nakatsu, C. H., Brouder, S. M., Wilbur, J. D., Wanjau, F., and Doerge, R. W.: Impact of tillage and crop rotation on corn development and its associated microbial community. *Proceedings of the 15th Conference of the International Soil Tillage Research Organization (ISTRO)*. Fort Worth, Texas: ISTRO. (2000)

[3] Krzanowski, W. J. and Marriott, F. H.: *"Multivariate Analysis; Part 2 Classification, Covariance Structures and Repeated Measurments."* John Wiley & Sons Inc, New York. (1995)

[4] McLachlan, G. J.: *"Discriminant Analysis and Statistical Pattern Recognition."* John Wiley & Sons Inc, New York. (1992)

[5] Schott, J. R.: Testing for complete independence in high dimensions. *Biometrika*, **92.4**, 951–956. (2005)

[6] 青木　誠, 櫻井哲朗, 藤越康祝: 高次元独立性検定のロバストネスについて. 2006 年度統計関連学会連合大会講演報告集, 252. (2006)

[7] Fujikoshi, Y.: Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria. *J. Multivariate. Anal.*, **17**, 27–37. (1985)

[8] 藤越康祝, 薬真寺裕, 安部友紀: 高次元多変量 2 値データの判別における変数選択. 2003 年度統計関連学会連合大会講演報告集, 519–520. (2003)

付録：フィンガープリントデータ

Table 5　トウモロコシのフィンガープリントデータ (Nakatsu et al. [2] より引用)

1	00000000110011000010000001000001000100010000001001000000000000000000000000000000000000
1	00000000110011000010000001000000000100000001000001000100000000000000000000000000000000
1	00000000110011000010000001000001000100000000000000000100000000000000000000000000000000
1	00000000000011000010000101000001000100100000000100100010010000000000000000000000010000
1	00000000100011000000000001000000000110000000010101000100000100000000000000000000010000
1	00000000110011000000000001001001000100000000000000000100000000000000000000000001010000
1	00000000110011000000000001000000000100000000010000001000001000000000000000000001000000
1	00001010000011000000000001000001000000000000000000000100000000000000000000000000000000
1	10001010000011000000000001000001001000000000000000000100000000000000000000000000000000
1	00101010000011000000000001000001001100000001000000000100100100000100100000000000000000
1	00000000000011000000000001000001000100000000000000000100000000000000000000000001000000
1	00100010000011000000000001000001001000000000000000000000000000000000000000000000000000
1	00000000000011000000000001000001000100100000000000000100000000000000000000000010001
1	00000000000011000000000001000001100000000000000000000100000000000000000000000010001
1	00000000000011000000000101000001000000000000000000000100000000000010010000000001010001
1	00000000000011000000000001000001000010000000000000000100000000000000000000000010001
1	00000000000011000000000001000001000010000000000000100010000000000000000000000010001
1	00000000000011000100000001000001000000000000000000000100000000000000000000000010001
1	00000000000011000000000001000000010000000000000000000000000000000000000000000010001
1	00000000000011000000000001000001000100000000000000000100000000000000000000000010001
1	00000000000011000000000001000001000000000000000000000100000000000000000000000010001
1	00000000000011000000000001000000000000000000000000000100000000000000000000000010001
2	00000000110101000010000001000000000001000000000100001000000010000000010000000010000
2	00000000110101001000000001010000000000000001000100000000000100000000010000000010000
2	00000000110101000010010000010000000000000001000000000000000100000000000000000010000
2	00000000110101000010000011000000001000000001000000000000000000000000000000000010000
2	00000000110000000010000001000000000100000010010000001000000000000001100000000010000
2	00000000110001000010000001100000000000000000000000000000000100000000010000000010100
2	00001000100001000000000001000000000000000000000000000000000000000001000000000010000
2	00001000100001000010100001000000000000000010000100001001000000001000100000000010000
2	00000000100000000010010001000000001000000100100000000000000000000000000000000000000
2	00000000100001000000000001010000000001000000000000000001000000000000000000000000000
2	00000000100101000000010001000001000000010000000000010000000000000000000000000000000
2	00000000100101000001010001001001000000010000000000000000000000000010000000000010000
2	00000000101000000000000001000000000000000000000000000000001001000000000000000010010
2	00000000100000000010000001010001000001000000101000000001001010000000000000000010100
2	00000000110101000000000001000001000001000000101000010010000000000000000000000000000
2	00000000110101000000000001000001000000010000010100001001000000000000000000000000000
2	00000000110101000010000001000001000000001000001000000000000000000000000000000000000
2	00000000111000000000000001000000000100100000010100000000000000001000000000000010000
2	00000000100001000000000001000000000000110000010100001001000000000000000000000000000
2	00000100100001000000000001000001000000111000000100010000000000000000000000000000000
2	00000000100100001000000010000000000001000000101000000000000000000000000000000000000
2	00000000100000001000000010000001000000000000010000000000000000000000000000000000000
3	00000000110001100001000010001000100000001000100000000001001000100000000000000000000
3	00000000110001001000100001000100010000000000000000000100010000011000000000000010001
3	00000000110001000000100001000000010000000000000000000100010000000000000000000000000
3	00000000100001000000100001000000010000000000000000000000000000000000000000000010001
3	01000000100001000000100001000100100000001000000000000000000000000000000000000010001
3	10010000100001100000000101000000010000000001000000000001000000000000000000000000000
3	00010001000001000000000001000000010000000100010000000000000000000000000000000000000
3	00100000100001000000000001000000010000000000000000000100000000000000000000000000000
3	00100001000001000000000001000000010000000000000000000000000000010000000000000000000
3	00100001000001000000000001000000010000000100000000010000000000000000000000000000000
3	00100000000001000000000001000000010000000100000000000100000000000000000000000000000

3	0000000000000100000000000100000000000000000000000010100000100000000000010000010001
3	0000000000000100000000000100000000000000000010000000000000000000100000001000000010001
3	0000000000000100000000000100000001000000000010000000010000000000100000000000000010001
3	0000000000000100000000000100000001000000000010000000010000000001000010000000000010001
3	0000000000000100000000000100000001000000000010000000010000000001000010000000000010001
3	0000000000000100000000000100000000000000000000000000010000000000000000000000000010001
3	0000000000000100000000000100000000000000000000000000010000000000000000000000000010001
3	0000000000000100000000000100000001000000000010000000010000000000000000000000000100010001
3	0000000000000100000000000100000001000000000010000000010000000000000000000000000100010001
3	0000000000000100000000000100010001000000000010000000010000000001000000000000000010001
3	0000000000000100000000000100000001000000000000000000010000000000000000000000000010001
3	0000000000000100000000000100000000000000000000000000010000000000000000000000000010001
4	0000000011000010010000000100010101100000001000000000001000010000100000000000000000000
4	0000000000001100000000000100000000000000010000000000000000000000000000000000000000000
4	0000000011000010000000010000010000000001000000001000001000000000100000000010000000000
4	0000000010001100000000000100000000000000000000000000000100000000010000010000010010000
4	0000000011000001010010000000010000000001001000000000001000000000000000001010101000000
4	0000000000001010001000000010001000000000000000000000110000000100000010000000010000000
4	0000000010000100000000001010000000110000000000000000100000000100000000000000000000000
4	0000000010000100000000001010000010000000100100000000000000000000000000000000000000000
4	0000000000001000000000001010000010100000100000000100000000000000000000000000000000000
4	0000000010000100010000010100000001000000000000000000000000000000000000000000000000000
4	0000000010001000000000001000000011000000000000000000000000000000000100001000000000001
4	0010100001000101000000000100000000000001000000000001000000000010000000000000000000001
4	0000000000001010100000000010000000000000000000001000000010000000000000000000000010000
4	0000000010010101000000101100000001000000010000000110000000000100000000000000000000000
4	0000000000000100000000000100000001000000010000000000000000100100010000000000010000000
4	0000000000000100000000000100000001000000000000000000000000000000000000000100000010001
4	0000000000001101000000000100000000000000001000000000100000010000000000000100000000000
4	0000000011000000000000000100001001000000000000000001000100000000000000000000000000000
4	0000000000001010000000000100000001000010010000000110000000000000000000000010011001
4	0000000010010100000000000100010000000001000000001000001000000000000000010001000010001
4	0000000010010100000000000100000000000001000000001000000000000000010000010001000010001
4	0000000010010100000000000100000000000001000000001000000000000000000000010000000011001