高次元データにおけるナイーブな正準相関を用いた特徴選択手法

玉　谷　　充

1. はじめに

特徴選択手法とは，パターン認識において有効な特徴を抽出する方法の1つである。そもそもパターン認識とは，認識対象がいくつかのクラスに分類できるとき，観測されたパターンをそれらのクラスのうちのひとつに対応させる処理のことである。単純なパターン認識としては，重さや大きさによって対応させる硬貨の認識，0から9の数字のいずれかに対応させる数字の認識などが挙げられるが，キーボードからの入力に代わる文字の入力方法として，ヒトの話す音声をコンピュータによって何らかの処理をし，話している内容を文字データとして取り出す音声認識や，画像データの画像内容を分析し，その形状を認識する画像認識などがパターン認識の応用として挙げられる。近年における技術においても，音声認識だとiPhoneのSiriであったり，画像認識においてもGoogleアプリであったりと，日常生活においてもパターン認識という技術が広く用いられていることが分かる。

パターン認識の枠組みとしては，2値化やノイズ除去などのような認識過程を容易にするために前もって行う処理（前処理）をし，何らかの構造を持った特徴ベクトルを選択し，クラスが既知であるパターンを学習し，クラスが未知であるパターンを，前もって学習して得た評価式をもとに決定・出力するのがおおまかな流れである。本論文では，確率的な対応関係

に基づいて判別関数をどのように構築すればいいのかという枠組みを紹介し，高次元データにおいて特徴選択手法をどう組み込めば良いのかを考える。一般に，クラスが既知のデータに基づいて学習をする手法を教師あり学習と呼び，与えられたクラスが未知であるデータに対して誤って判別してしまう確率をなるべく小さくすることが目的である。その目的を得るために，k-Nearest Neighbor Rule や経験リスク最小化，カーネル法，ニューラルネットワークなど，様々な手法が提案されてきた（各手法については，例えば Devroye, et al. (1996) を参照）。特に，クラス間の距離を大きく，クラス内の分散を小さくするような基準から導かれる Fisher の線形判別関数はよく用いられ，正規性かつ等分散の条件のもとではその最適性も証明されている。

しかしながら，Fisher の線形判別関数は少ない特徴（次元）に対して，既知のデータがある程度手元にあれば構築可能であるが，次元が手元にあるデータ数よりも大きい場合，線形判別関数に含まれている分散共分散行列の推定量を求める際に特異行列となってしまい，従来の手法では構築することができないという問題が生じてしまう。そこで一つの改良として，得られた推定量の対角成分だけを取り出して判別関数を構築する。こうすることにより，各成分の分散が 0 でない限り，判別関数を構築することが可能となる。このような手法を Naive Bayes といい，Bickel and Levina (2004)，Fan and Fan (2008) はこの判別関数の性能について考察を与えた。

一般に，データ数よりも次元が極端に大きいようなデータを高次元小標本 (HDLSS) というが，近年では DNA マイクロアレイデータの解析やゲノムに関する研究，すなわち HDLSS に関する研究が盛んに行われている。しかしながら，このようなデータはコストが非常にかかるため，多くの標本を収集することは難しく，そのようなもとでの解析・理論は発展途上の段階である。Naive Bayes の手法は，データ数と次元の関係に依らず構築することができるが，その反面で共分散構造を落とし過ぎているのが一つ

の欠点であり，HDLSS においては推定量の一致性についても議論がなされる。

本論文は次のように構成されている。第 2 節では，特徴選択を実行した後に行う判別分析を正準相関分析という観点で構築をする。その際，Fisher の評価基準というものに帰着するため，その辺りの部分について議論する。第 3 節においては HDLSS の設定においての判別手法について述べる。特に，HDLSS の設定のもとだと従来の手法を使うことができない問題に直面してしまうが，その問題を回避する手法の 1 つであるナイーブな正準相関に基づく判別関数という手法を紹介する。第 4 節では，特徴選択手法を紹介したもとで，先行研究を改善する特徴選択手法を新たに提案する。そして，第 5 節においては新たに提案した特徴選択手法の性能を見ていくために，適当な設定のもとで先行研究との比較を行っていく。

2. 正準相関分析に基づく判別分析

判別分析とは，クラスへの属性が未知の n 個のデータがあり，どのクラスに属しているのかが未知の観測値が新たに得られたときに，n 個のデータを利用してその新たな観測値のクラスへの属性を決定する手法である。

数学的な枠組みとしては，M をクラス数としたとき，クラスを与える関数

$$g : \mathbb{R}^d \to \{1, 2, \cdots, M\}$$

を定義し，新しい観測値 $x \in \mathbb{R}^d$ に対するクラスの割り当てを行う。このような関数 g を判別関数といい，観測値 x のクラスを y としたとき，なるべく $g(x) \neq y$ とならないように判別関数 g を構築する必要がある。

本論文では，$M=2$ として，2 値 $(y=0, 1)$ の判別における最良の判別関数 g^* の構築方法について述べる。特に，ここでは各クラスの確率分布を d 変量正規分布に従っていると仮定する。一般に，確率変数ベクトル X

が平均ベクトル μ，分散共分散行列 Σ の d 変量正規分布に従っているとき，X の確率密度関数 f_X は，

$$f_X(x)=\frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

と書ける。これを $X\sim N_d(\mu, \Sigma)$ と書く。この分布のもとで最良の判別関数 g^* を考えた場合，Fisher の線形判別関数が最適であると言われている。線形判別関数とは，$\mathbb{R}^d$ 空間における新たな観測値ベクトル $x=(x_1, \cdots, x_d)^T$ に対して，重みベクトル $a=(a_1, \cdots, a_d)^T$ と定数 a_0 によって重み付けされた量 $a^Tx+a_0=\sum_{i=1}^d a_ix_i+a_0$ の符号によって判別する関数のことである。このような関数は，複雑な超曲面を当てはめる手法よりも解釈しやすく，判別関数の精度評価をする際にも解析しやすいのが特長である。ここで問題となってくるのが，$\mathbb{R}^d$ における重みベクトル a をどのように決めれば良いのかであるが，実際に重みベクトル a の決定は，様々な評価基準によって実行できる。最初に提唱された手法は直感的な概念に基づく手法であり，Fisher (1936) によって提唱された。この手法について理解するために，いくつかの記号を導入する。

定義 2.1．クラス k のもとでの条件付き平均ベクトル μ_k を以下で定義する：

$$\mu_k=E[X|Y=k]$$

定義 2.2．同じクラス内の散らばりの尺度を測る量をクラス内変動行列 S_W とし，以下で定義する：

$$S_W=E[(X-\mu_0)(X-\mu_0)^T|Y=0]+E[(X-\mu_1)(X-\mu_1)^T|Y=1]$$

定義 2.3．クラス間の散らばりの尺度を測る量をクラス間変動行列 S_B とし，以下で定義する：

$$S_B=(\mu_1-\mu_0)(\mu_1-\mu_0)^T$$

これらの尺度を用いて，観測値ベクトル X を重みベクトル a の向きで射影した $Z(a)=a^TX$ の量についての尺度を考える。このとき，$Z(a)$ はスカラー量となり，この $Z(a)$ についての定義 2.1 から定義 2.3 によって定義される量をそれぞれ $\tilde{\mu}_k(a)$，$\tilde{S}_W(a)$，$\tilde{S}_B(a)$ とすると，

$$\tilde{\mu}_k(a)=a^T\mu_k,\qquad \tilde{S}_W(a)=a^TS_Wa,\qquad \tilde{S}_B(a)=a^TS_Ba$$

で与えることができる。Fisher は，この変換後に得られた $\tilde{S}_W(a)$ と $\tilde{S}_B(a)$ に対する比

$$J(a)=\frac{\tilde{S}_B(a)}{\tilde{S}_W(a)}=\frac{a^TS_Ba}{a^TS_Wa}$$

を最大にすることを考えた。つまり，クラス間変動をできるだけ大きくし，なおかつクラス内変動を小さくするようにしたのが Fisher の与えた概念である。この評価基準 $J(a)$ を最大にするベクトルを $\tilde{a}$ としたとき，いくつかの線形代数における固有値問題を解くことによって

$$\tilde{a}=S_W^{-1}(\mu_1-\mu_0)$$

を導出することができることが知られている。ゆえに，互いのクラスを分ける分岐点を l とし，得られた $\tilde{a}$ を用いて $l=(\tilde{\mu}_1(\tilde{a})+\tilde{\mu}_0(\tilde{a}))/2$ とする。このとき，$\tilde{g}_a(x)=Z(\tilde{a})-l$ とすれば，Fisher の判別基準に基づく判別関数を得ることができる。

次に，正準相関分析について述べ，判別分析との関連性についてまとめる。正準相関分析を考える際に，改めて設定を表記する。まず，p 変量の中心化された確率変数ベクトル $X\equiv X-E[X]$ と q 変量の中心化された確率変数ベクトル $Y\equiv Y-E[Y]$ を考え，$s=\min\{p,q\}$ とする。また，重みベクトルを $a_i\in\mathbb{R}^p$，$b_i\in\mathbb{R}^q$ $(i=1,\cdots,s)$ としたとき，以下の合成変量

$$u_i = a_i^T X, \quad v_i = b_i^T Y$$

を考え，u_i と v_i の相関が最も高くなるような重みベクトル a_i と b_i を求める。つまり，u_i と v_i の相関係数を $r(u_i, v_i)$ とし，この量が最大となるような重みベクトル a_i，b_i と定めれば良いが，これは $V[u_i] = V[v_i] = 1$ という制約のもとで $r(u_i, v_i)$ を変形すると，

$$r(u_i, v_i) = a_i^T E[(X - E[X])(Y - E[Y])^T] b_i \equiv a_i^T \sum_{XY} b_i$$

となる。あとはラグランジュの未定乗数法を用いて

$$a_i^T \sum_{XY} b_i + \lambda\left(a_i^T \sum_X a_i - 1\right) + \nu\left(b_i^T \sum_Y b_i - 1\right) = 0$$

を解くと，固有値問題

$$\left|\sum_X^{-1/2} \sum_{XY} \sum_Y^{-1} \sum_{YX} \sum_X^{-1/2} - \lambda_i^2 I\right| = 0$$

に帰着し，固有値が0でないものに対する固有ベクトルは s 個定まる。このときの u_i，v_i を第 i 正準変量といい，u_i，v_i の間の相関係数 $\rho_i = \sqrt{\lambda_i^2}$ を第 i 正準相関係数という。特に，$B = \sum_X^{-1/2} \sum_{XY} \sum_Y^{-1/2}$ とおけば，固有値問題は $|BB^T - \lambda_i^2 I| = 0$ と書き表すことができる。ここで，正準相関に基づいて判別関数を構築することを考える。すなわち，先の正準相関分析では X，Y を共に中心化された p，q 変量確率変数ベクトルと定めていたが，特に Y を $q = 2$ とし，2値をとる離散型確率変数ベクトルとする：

$$Y = \begin{bmatrix} Y_0 \\ Y_1 \end{bmatrix}, \quad Y_1 = \begin{cases} 1 & Probability = \pi \\ 0 & Probability = 1 - \pi \end{cases}, \quad Y_0 = 1 - Y_1.$$

そして，Y に関しては中心化しない以下の量を考える：

$$C = \sum_X^{-1/2} E[(X - \mu_X) Y^T] E[YY^T]^{-1/2}.$$

このとき，各期待値部分について評価をすると，最終的に

$$C=\sum_X^{-1/2}(\mu_1-\mu_0)\left[\sqrt{\pi}(1-\pi)\quad -\pi\sqrt{1-\pi}\right]$$

が得られ，

$$CC^T=\rho\sum_X^{-1/2}(\mu_1-\mu_0)(\mu_1-\mu_0)^T\sum_X^{-1/2}$$

となる。ただし，$\rho=\pi(1-\pi)$ である。ここで，固有値問題 $CC^Tp=\nu^2p$ について考え，$b=\sum_X^{-1/2}p$ とおくと，

$$\rho(\mu_1-\mu_0)(\mu_1-\mu_0)^Tb=\nu^2\sum_X b$$

となり，両辺左から b^T を掛ければ，Fisher の評価基準に帰着する：

$$J(b)=\frac{b^T(\mu_1-\mu_0)(\mu_1-\mu_0)^Tb}{b^T\sum_X b}.$$

3. HDLSS における判別関数の構築

前節においては，母集団における判別関数の構築方法について述べてきた。しかしながら，実際には d 変量正規分布の場合だと平均ベクトル μ_k や分散共分散行列 Σ は未知であり，これらのパラメータは既に得られたデータに基づいて推定量を与えていかなければならない。従来，標本サイズがある程度得られており，次元が標本サイズよりも大きくない場合には先で議論した判別関数を構築することは可能であるが，次元が標本サイズよりも大きい場合だと判別関数を構築することすら不可能な状況となる。ここで本節では，なぜそのような問題が生じてしまうのかを述べ，先行研究で実際に行われている手法 Naive Bayes と正準相関に基づく判別関数の構築方法について述べる。

まず，標本に基づく従来の手法について与える。各クラスにおける観測値ベクトル X_k は，平均 μ_k，分散共分散行列 Σ の d 変量正規分布 $N_d(\mu_k, \Sigma)$ に従っていると仮定する。また，ここでは各クラスで n_k 個のデ

ータがあると仮定し，クラス k における観測値ベクトルを X_{k1}，…，X_{kn_k} とする。このとき，各パラメータの推定量としてよく用いられるのは最尤推定量である以下の標本平均 $\widehat{\mu_k}$ と標本分散共分散行列 $\widehat{\Sigma}$ である：

$$\widehat{\mu_k}=\frac{1}{n_k}\sum_{i=1}^{n_k}X_{ki},$$

$$\widehat{S_k}=\frac{1}{n_k-1}\sum_{i=1}^{n_k}(X_{ki}-\widehat{\mu_k})(X_{ki}-\widehat{\mu_k})^T,$$

$$\widehat{\Sigma}=\frac{1}{n-2}\left\{(n_0-1)\widehat{S_0}+(n_1-1)\widehat{S_1}\right\}.$$

このようにして得られた推定量を，対応するパラメータに代入した判別関数を，標本に基づく判別関数と呼ぶ。従来の手法では，得られたデータの標本サイズはある程度大きく，特徴（次元）が標本サイズを超えないような設定のもとで推定量を求めれば，データが互いに独立であるという仮定がある限り，判別関数を構築することができる。

一方，標本サイズ n と次元 d の関係が逆の関係になったとき，すなわち，データと次元が $n<d$ という関係になってしまったとき，本来 $\widehat{\Sigma}$ は n 本のベクトルを用いて行列を生成していることから，$rank\widehat{\Sigma}=\min\{n, d\}$ となる。$\widehat{\Sigma}$ 自体は $d\times d$ の行列なので，$n\geq d$ であれば $rank\widehat{\Sigma}=d$ となり，$\widehat{\Sigma}$ は正則行列であることが分かる。しかし，$n<d$ だと $rank\widehat{\Sigma}=n<d$ となってしまい，$\widehat{\Sigma}$ は正則行列でなくなってしまう。すなわち，これは $\widehat{\Sigma}$ の逆行列が存在しないことを意味し，Fisher の評価基準で用いられていた分散共分散行列に代入することができないといった問題が生じてしまう。

では，標本サイズ n と次元 d の関係が $n<d$ のとき，どのように推定量 $\widehat{\Sigma}$ の逆行列が存在しない問題を回避すれば良いのだろうか。その問題解決の手法の 1 つとしては，対角成分だけを残し，非対角成分は全て 0 にするという手法が挙げられる。すなわち，従来の分散共分散行列の推定量 $\widehat{\Sigma}$ の代わりに，$\widehat{D}=diag\,\widehat{\Sigma}$ を用いることを考える。このようにして判別

関数を構築すれば，$\widehat{D}=(\widehat{\sigma_{ij}})$ の対角成分 $\widehat{\sigma_{ii}}$ が非零である限り，逆行列は存在することから，判別関数を構築することができる。

次に，正準相関に基づく判別分析において，Naive Bayes を適用することができないかどうかについて述べる。母集団においては $b=\sum_X^{-1/2}p$ を用いることによって Fisher の判別基準に帰着することを述べていたが，ここでは C の推定量を与え，$\widehat{C}\widehat{C}^T$ の最大固有値に対する固有ベクトル $\widehat{p}$ を求め，この推定量を用いることにより Naive Bayes が導けることを示す。その際に，$n<d$ だと推定量 $\widehat{b}=\widehat{\sum}_X^{-1/2}\widehat{p}$ は，Fisher の線形判別関数と同様 $\widehat{\sum}_X$ の逆行列が得られないという問題が生じる。ゆえに，X，Y をそれぞれデータ行列とクラス行列とし，行列 C の推定量を

$$\widehat{C}=\widehat{D}^{-1/2}\left(\frac{1}{n}XY^T\right)\left(\frac{1}{n}YY^T\right)^{-1/2}$$

によって与える。ただし，$X=[X_{11},\cdots,X_{1n_1},X_{01},\cdots,X_{0n_0}]$，$Y=[Y_{11},\cdots,Y_{1n_1},Y_{01},\cdots,Y_{0n_0}]$ である。このとき，$\widehat{C}\widehat{C}^T$ の最大固有値に対する固有ベクトルの推定量 $\widehat{p}$ を求めることができ，$\widehat{b}=\widehat{D}^{-1/2}\widehat{p}$ を得ることができる。つまり，通常の判別分析では Fisher の判別基準を満たすベクトル $\vec{a}$ を求めるために，固有値問題を考えていくのに対して，正準相関に基づく判別手法においては，$\widehat{C}$ を用いてその行列の最大固有値に対する固有ベクトルを求め，データ行列 X とクラス行列 Y の間で相関が大きくなるような $\widehat{p}$ を与える。特に，$\widehat{C}$ の中でも XY^T/n という量が X と Y の間の相関に関連している。一方，$\widehat{C}$ と $\widehat{C}^T$ を入れ替えた $\widehat{C}^T\widehat{C}$ の最大固有値に対応する固有ベクトルの推定量 $\widehat{p}_0$ は，$\widehat{p}_0=(\sqrt{n_0/n_1},-1)^T$ と求められる。したがって，求める $\widehat{p}$ は $\widehat{C}\widehat{C}^T$ の最大固有値に対する固有ベクトルのため，固有値・固有ベクトルの関係式

$$\widehat{C}^T\widehat{C}\widehat{p}_0=\nu\widehat{p}_0 \Longrightarrow \widehat{C}\widehat{C}^T(\widehat{C}\widehat{p}_0)=\nu(\widehat{C}\widehat{p}_0)$$

より，$\widehat{C}\widehat{C}^T$ の最大固有値に対する固有ベクトル $\widehat{p}=\widehat{C}\widehat{p}_0$ を得る。この $\widehat{p}$

を整理すると，

$$\widehat{p}=\widehat{C}\widehat{p}_0=\sqrt{\frac{n_0}{n_1}}\widehat{D}^{-1/2}(\widehat{\mu}_1-\widehat{\mu}_0)$$

という厳密な等号が成立し，最終的に求める $\widehat{b}$ は

$$\widehat{b}=\widehat{D}^{-1/2}\widehat{p}=\sqrt{\frac{n_0}{n_1}}\widehat{b}_{NB}$$

となる。すなわち，$\widehat{b}$ の長さは Naive Bayes と $\sqrt{n_0/n_1}$ 倍異なるだけであり，判別に影響を与えないため，正準相関に基づく判別関数から Naive Bayes が導ける。

4. 特徴選択手法

前節では HDLSS に対する理論的な側面で判別関数の構築方法を述べてきたが，ここでは HDLSS における実用的な側面についてまずは見ていくことにする。HDLSS とは High Dimension Low Sample Size の略称であり，標本サイズが数十，数百というのに対して，次元が数千，数万あるようなデータのことを指す。HDLSS の例としては白血病 (Leukemia) データ，肺がん (Lung Cancer) データ，前立腺がん (Prostate Cancer) データ等が挙げられ，オープンデータとしても公開されている。これらのデータは次元が 10,000 前後あるようなデータであり，各データの訓練データ (Training Data) とテストデータ (Test Data) を各クラスごとで平均をとり，その差の値をプロットした図を以下に与える：

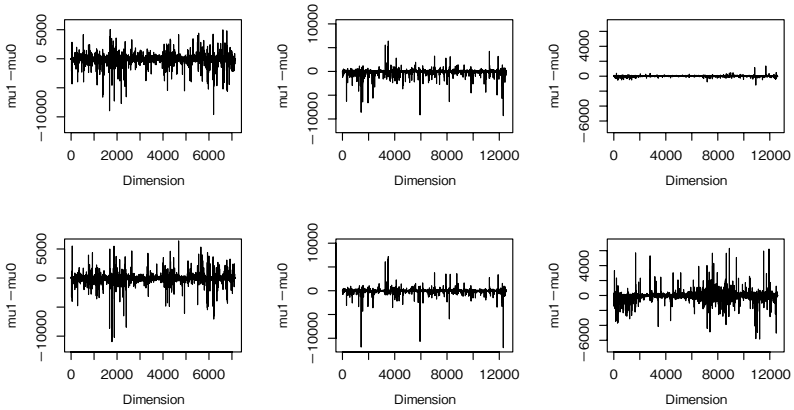

図 2.1：HDLSS の可視化

図 2.1 を見てもわかるように，各特徴には大きな差が見られたり，ほとんど差の無い特徴もまた存在することが分かる。特に，ほとんど差の無い特徴については，判別を行うときに情報を得られないことからノイズとなってしまい，そのようなノイズが蓄積してしまうと，判別自体に弊害が生じてしまう可能性がある。実際には特徴選択をすることによって弊害を避けることができ，HDLSS における特徴選択手法の先行研究としては Tibshirani, et al. (2002) の Nearest Shrunken Centroids method (NSC) や Fan and Fan (2008) が提案した Feature Annealed Independence Rules (FAIR) 等といったものが挙げられる。

本節では，Fan and Fan (2008) が提案した HDLSS における特徴選択手法 FAIR を簡単に紹介し，そこからナイーブな正準相関を用いた特徴選択手法というものを新たに提案する。また，以下では特に断らない限り“次元(特徴)”とは，説明変数の個数を表すことにする。HDLSS における特徴選択手法 FAIR は 2 標本 t 統計量に基づく特徴選択手法であり，j 番目の特徴を

$$T_j=\frac{\widehat{\mu}_{1j}-\widehat{\mu}_{0j}}{\sqrt{\widehat{\sigma}_{jj}(1/n_0+1/n_1)}}$$

としたときに，この値が大きくなるような特徴の番号 j を上から順に取ってくる手法である。一般に，特徴選択手法は以下の 4 つの手順によって構築される。

Step 1. 特徴選択のベースとなるベクトル $\widehat{c}$ を計算する。
Step 2. $\widehat{c}$ の成分の絶対値が降順となるように並べ替えを行う。

$$|\widehat{c}_{i_1}|\geq|\widehat{c}_{i_2}|\geq\cdots\geq|\widehat{c}_{i_m}|\geq\cdots\geq|\widehat{c}_{i_d}|\geq 0.$$

Step 3. 上記で得られた並べ替えに基づいて，データ X とベクトル $\widehat{c}$ を並べ替える。
Step 4. 1 番目から $\widehat{m}$ 番目までの特徴を用いて判別を行う。

先に挙げた FAIR の場合，各成分において 2 標本 t 統計量を計算し，その値が大きい順から取ってくることになるが，前節で用いられた $\widehat{p}=(\widehat{p}_1,\cdots,\widehat{p}_d)^T$ から $T_j=c_n\widehat{p}_j$ として表現することができる。この先行研究に対して，本論文では前節で用いられた $\widehat{b}=(\widehat{b}_1,\cdots,\widehat{b}_d)^T$ に基づいて並べ替えをする手法を提案する。この手法を Naive Canonical Correlation (NACC) といい，詳細については Tamatani, Koch and Naito (2012) にてまとめている。一方，Step 4 において，$\widehat{m}$ 番目までの特徴を用いて判別を行うとしているが，$\widehat{m}$ というのをどのようにして決めるのかという点も 1 つの重要な点となる。Fan and Fan (2008) では，FAIR を提案する際に Naive Bayes における誤判別確率というのを導出しており，その結果に基づいて以下の選択個数 $\widehat{m}$ を与えている。

$$\widehat{m}=\underset{1\leq m\leq d}{\arg\max}\frac{1}{\widehat{\lambda}_{max}(R_m)}\frac{\left[\sum_{j=1}^{m}(\widehat{\mu}_{1j}-\widehat{\mu}_{0j})^2/\widehat{\sigma}_{jj}+m(1/n_0-1/n_1)\right]^2}{nm/(n_1n_0)+\sum_{j=1}^{m}(\widehat{\mu}_{1j}-\widehat{\mu}_{0j})^2/\widehat{\sigma}_{jj}}$$

ただし，R_m は相関行列であり，$\hat{\lambda}_{max}$ は行列に対する最大固有値である。本論文においても，この $\widehat{m}$ に基づいて構築するが，いくつか改良をしたもとで新たな特徴選択手法を提案する。まず，上記の $\widehat{m}$ については点推定によるものであり，これを区間推定の形で与え直すことによって特徴選択手法の改良を加える。Tamatani and Naito (2019) では，$\widehat{C}\widehat{C}^T$ の最大固有値 $\hat{\lambda}_{n,d}$ における分布収束性について議論した。これは，適当な正則条件のもとで

$$\sqrt{\frac{n}{\lambda}}\left\{\hat{\lambda}_{n,d}-\frac{n_1 n_0}{n^2}\frac{n-2}{n-4}\alpha^T D^{-1}\alpha-\frac{nd}{n^2}\frac{n-2}{n-4}\right\} \xrightarrow{D} N(0,\zeta)$$

となることを導いた。ただし，$\alpha=\mu_1-\mu_0$ である。また，ζ については特定の条件下であれば $\zeta=4\alpha^T D^{-1}\Sigma D^{-1}\alpha \,/\, \alpha^T D^{-1}\alpha$ に収束することを示すことができる。その結果を用いることにより，以下の $\lambda_{n,m}$ の $100(1-\alpha)\%$ の信頼区間を得ることができる：

$$U_m=\left[\hat{\lambda}_{n,m}-\frac{m}{n}\frac{n-2}{n-4}-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\lambda}_{n,m}}{n}}\hat{\zeta},\ \hat{\lambda}_{n,m}-\frac{m}{n}\frac{n-2}{n-4}+z_{\alpha/2}\sqrt{\frac{\hat{\lambda}_{n,m}}{n}}\hat{\zeta}\right]$$

ここで，$z_{\alpha/2}$ は標準正規分布 $N(0,1)$ の上側 $100(\alpha/2)$ パーセント点である。一方，固有値 $\hat{\lambda}_{n,d}$ については $\hat{\lambda}_{n,d}=(n_1 n_0/n^2)\widehat{\alpha}^T\widehat{D}^{-1}\widehat{\alpha}$ という形で表すことができ，$\widehat{m}$ の式の中にある $\sum_{j=1}^{m}(\widehat{\mu}_{1j}-\widehat{\mu}_{0j})^2/\widehat{\sigma}_{jj}$ に依存していることが分かる。ゆえに，$\widehat{m}$ を最大化にする量を $\Psi(\hat{\lambda}_m)$ としたとき，

$$m^*=\underset{1\le m\le d}{\arg\max}\ \max_{u_m\in U_m}\Psi(u_m)$$

という形で新たに特徴の選択個数 m^* を提案する。これにより，信頼区間内で最大にするような点を見つけるために区間 U_m 内で数値計算を行わなければならず，$\widehat{m}$ と比べると計算回数が多くなってしまうのが欠点ではあるが，実際の判別の精度を改善してくれる見込みはある。以下では，選択個数 m^* を用いて特徴選択する手法を FAIR* 並びに NACC* と命名す

る。次節では通常の特徴選択手法との性能を比較するために，数値実験を行うことにする。

5. 数値実験

ここでは次のような形で数値実験を行う。まず，パラメータの設定として各観測値ベクトルは d 変量正規分布 $N_d(\mu_l, \Sigma)$ $(l=0, 1)$ に従っていると仮定し，クラス 1 の平均ベクトル μ_1 は零ベクトル，クラス 0 の平均ベクトル μ_0 は最初の 10 個の成分は 2 であり，それ以外の成分は 0 といった形で与えておく。また，分散共分散行列 Σ については単位行列 I_d の場合と $A=diag(1, 2, \cdots, d)$ としたとき，$\Sigma=A^{1/2}AR(0.5)A^{1/2}$ の場合の 2 パターンについて実行する。ここで，$AR(0.5)$ は AR 構造のことであり，$AR(\rho)=\left(\rho^{|i-j|}\right)_{1\leq i,j\leq d}$ によって定義される。つまり，(i, j) 成分の添字 i と j の値が近ければ相関が高く，遠ければ相関が無くなっているような構造となっている。また，予め各クラスのテストデータを 5,000 個生成し，後の手順で出てくる誤判別率の推定値で用いることにする。次に，数値実験の手順について述べる。まず，トレーニングデータとして各クラス n_l 個生成する。この実験では特に偏りなく，$n_1=n_0$ としておくため，$n_l=n/2$ として考える。そして，そのトレーニングデータをもとに，前節で述べた特徴選択手法を適用する。すなわち，FAIR と NACC，そして，FAIR* と NACC* の 4 種類を実行することになるが，実際に特徴選択が有用であるのかを確かめるために Naive Bayes も実行するため計 5 種類の性能を比較することになる。性能については誤判別率によって比較し，これを 1,000 回繰り返すことによって推定値を与える。以上の手順によって得られた数値実験の結果が以下の表でまとめられる：

表 4.1：$\Sigma = I_d$ としたときの誤判別率の推定値

(n, d)	$(200, 200)$	$(200, 1000)$
FAIR	1.75%	4.45%
FAIR*	2.96%	5.90%
NACC	1.23%	2.00%
NACC*	2.24%	2.81%
Naive Bayes	2.17%	12.33%

表 4.1 から分かることは，次元が 200 の場合は通常の特徴選択手法の精度が良いことが分かり，Naive Bayes の精度もそれほど悪くない傾向が見られる。しかし，次元が 1,000 の場合，NACC と FAIR の間で差が生じ，NACC の方が精度が良いことが分かる。さらに，Naive Bayes については判別精度が明らかに悪くなっている傾向が見られる。これはパラメータの設定より，各クラスの平均ベクトル間で最初の 10 個の成分以外は全て同じ値であり，次元が 200 の場合は 190 次元，次元が 1,000 の場合は 990 次元がいわゆる無駄な特徴を示している。今回の数値実験においては無駄な特徴が増えれば増えるほど判別性能が劣化してしまう傾向が顕著に表れている。

表 4.2：$\Sigma = A^{1/2} AR(0.5) A^{1/2}$ としたときの誤判別率の推定値

(n, d)	$(200, 200)$	$(200, 1000)$
FAIR	20.45%	26.45%
FAIR*	18.22%	19.57%
NACC	23.36%	31.64%
NACC*	22.55%	29.87%
Naive Bayes	28.60%	38.73%

次に，分散共分散が単位行列ではないような設定 $\Sigma = A^{1/2} AR(0.5) A^{1/2}$ を得られた**表 4.2** から考察する。先ほどの**表 4.1** の設定とは違い，誤判別率が少し高めの値となっているが，Naive Bayes は**表 4.1** と同様どの特徴選択手法よりも精度が低く，次元が上がれば精度も悪くなっている様子が分かる。一方で，各特徴選択間の精度については FAIR の方が精度が良い傾向が見られる。さらに，通常の特徴選択手法と今回提案した特徴選択手法を比べると，FAIR と NACC いずれも提案した手法の方が良い傾向が得ら

れている。

ここではさらに，特徴選択の精度を考察するために，誤判別率の推定値以外にも有効な特徴，すなわち，最初の 10 個の成分を取ってきているかどうかを確認したり，選択個数 $\widehat{m}(m^*)$ の挙動を見ていくことにする。まず，有効な特徴をどの程度取ってきているのかについては，$(n,d)=(200,1000)$ かつ $\Sigma=I_d$ の設定において以下のような表を得ることができた：

表 4.3：$\Sigma=I_d$ としたときの有効な特徴を取ってきた割合

(n,d)	$(200,1000)$
FAIR	61.05%
FAIR*	46.96%
NACC	76.47%
NACC*	61.33%
Naive Bayes	100.00%

ただし，Naive Bayes に関しては常に全ての特徴を取ってきているため，最初の 10 個の成分は常に取ってきていることを意味するため 100% としている。Naive Bayes 以外の特徴選択だと NACC が一番有効な特徴を取ってきている傾向があり，次点で NACC*，FAIR，FAIR* と続いている。一方，**表 4.1** と比較してみると NACC が最も誤判別率が低いため，有効な特徴を取ってきている割合と誤判別率には関連があることが分かる。次に，選択個数 $\widehat{m}(m^*)$ の挙動を $(n,d)=(200,1000)$ かつ $\Sigma=A^{1/2}AR(0.5)A^{1/2}$ の設定において見ていくことにする。こちらも表としてまとめておく：

表 4.4：$\Sigma=A^{1/2}AR(0.5)A^{1/2}$ としたときの選択個数 $\widehat{m}(m^*)$ の推定値

(n,d)	$(200,1000)$
FAIR	43.900
FAIR*	8.457
NACC	24.792
NACC*	11.481
Naive Bayes	1000.000

ただし，Naive Bayes に関しては常に全ての特徴を取ってきているため，1,000 としている。各特徴選択手法同士を比較してみると，FAIR* が最も特徴を選択せず，続いて NACC*，NACC，FAIR の順で選択個数が増えている。**表 4.2** で得られた誤判別率と比較すると，選択個数が多いからといって精度が最も低いというわけでなく，例えば FAIR については誤判別率は 2 番目に低いにも関わらず，選択個数は最も多く取ってきているため，$\Sigma = I_d$ の設定と比べると単純な結果となっていないことが分かる。実際に各成分がどれだけの割合で取ってきているのかを可視化を行うと以下の図が得られた：

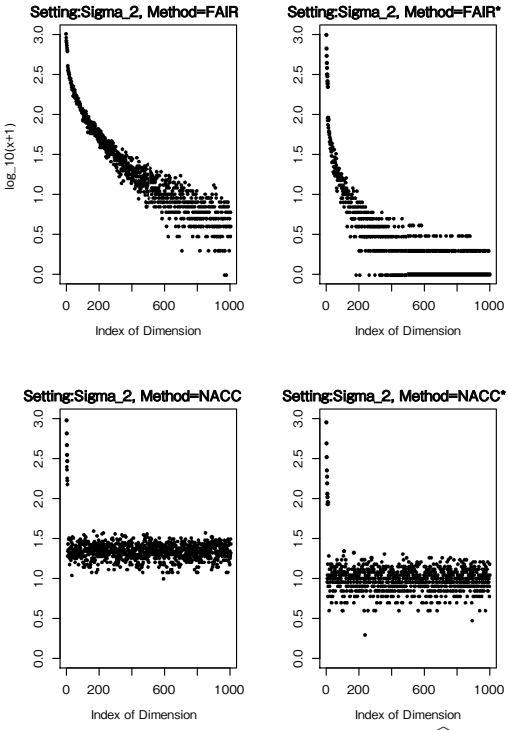

図 4.1　$\Sigma = A^{1/2}AR(0.5)A^{1/2}$ としたときの選択個数 $\widehat{m}$ (m^*) の割合

図 4.1 は横軸が特徴の番号，縦軸が対数化 $(\log_{10}(1+x))$ した数値であり，値が大きければ取ってきている割合も高いことを意味する。例えば FAIR, FAIR* に関しては最初の 10 個の成分はそれ以外の成分よりも大きい割合で取ってきているが，徐々に成分が高くなるほど取ってきている割合が少なくなっている様子が分かる。特に FAIR* に関しては 1,000 回の数値実験の中で全く取ってきていない成分も多くあることが分かり，これが誤判別率の推定値が低くなっている要因であることが分かる。一方，NACC, 並びに NACC* については FAIR や FAIR* と同様最初の 10 個の成分はそれ以外の成分よりも大きい割合で取ってきているが，それ以外の成分はある一定の割合で取り続けている様子が分かる。特に，NACC については NACC* よりもより高い割合で取り続けているため，それが誤判別率の推定値に影響していると考えられる。どちらにも共通していることは，新たに提案した特徴選択手法は $\Sigma = A^{1/2}AR(0.5)A^{1/2}$ としたときの設定上ではうまく特徴の選択個数が機能していることが分かり，それが誤判別率にも影響を及ぼしていることが考察される。

【参考文献】

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, *10*(6), 989-1010.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, Springer, New York.

Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Annals of statistics*, *36*(6), 2605-2637.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, *7*(2), 179-188.

Tamatani, M., Koch, I. and Naito, K. (2012). Pattern recognition based on canonical correlations in a high dimension low sample size context. *Journal of Multivariate Analysis*, *111*, 350-367.

Tamatani, M. and Naito, K. (2019). High dimensional asymptotics for the naive Hotelling

T^2 statistic in pattern recognition. *Communications in Statistics-Theory and Methods*, *48*(22), 5637-5656.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, *99*(10), 6567-6572.