統計数理(2003)
第 51 巻 第 2 号 389–406

［原著論文］

Kernel Flexible Discriminant Analysis による高次非線形データの判別とその応用

安道 知寛[†]

（受付 2003 年 3 月 19 日；改訂 2003 年 6 月 10 日）

要　　旨

判別分析とは，複数の特性について観測・測定された多次元データに基づいて，判別関数を構築し，将来のデータを分類する手法を総称する．これまで，実際問題に対して最も適用を試みられてきた Fisher の線形判別法は，科学・社会システムの発展に大きく貢献してきた．しかし，近年，計算機のハードウェア・ソフトウェア両面にわたる急速な発展は，様々な種類の大規模データの蓄積を容易にし，複雑な非線形構造を有するデータから有効かつ効率的に情報を抽出する手法の必要性が急速に高まっている．

本稿では，カーネル関数に基づく非線形回帰分析の枠組みにより，Fisher の線形判別関数を非線形へと拡張し，文字認識など実際問題への応用に対して十分機能する手法を提案する．また，非線形判別関数の信頼性，汎化能力を向上させるため，パラメータ推定においては平滑化法を適用する．モデル構築に当たっては，平滑化パラメータ等の選択が本質的となるが，これらの選択を情報量，及びベイズ理論の観点から考察したモデル評価規準をそれぞれ導出する．また，諸分野で蓄積されつつある実データおよび人工データの解析を通して，提案する手法の有効性を検証する．

キーワード： Fisher の線形判別法，カーネル法，最適スケーリング，平滑化，統計的モデル選択．

1. はじめに

現在の高度に発展した計算機の利用環境の下で，様々なモデリング手法が開発され，これまでは十分な解析が難しかった複雑な非線形構造を内在する現象の分析が可能となりつつある．本来，人間は高度な統計的情報処理能力を備えており，この知的処理機能を高速計算機で実現しようとする魅力的な試みは，情報工学分野のみならず，社会システム構築に関わる様々な分野でなされてきた．その知的処理機能に関する中心的話題の一つとして“判別分析”があり，筆跡鑑定，郵便番号の数字認識，航空券の音声予約，指紋・網膜照合によるセキュリティシステム，債券信用格付け，倒産予測など現実社会の様々な場面で貢献している．

判別分析とは，観測・測定されたデータを，あらかじめ定められた複数の集団のうちの一つに対応させる処理であり，その潜在的な実用価値は，一般に広く認識されている．従来，判別分析の手法として，Fisher の線形判別関数(Mardia et al. (1979))が主に用いられてきたが，複雑な構造を有するデータの分析に対しては，十分に機能していないことがしばしば指摘されて

[†] 九州大学大学院 数理学府： 〒812–8581 福岡市東区箱崎 6–10–1

きた．この問題は，判別分析のみならず，回帰分析など他の統計手法に関しても認識されており，近年では，スプラインやニューラルネットワークなどに代表される“非線形モデル”と平滑化(正則化)法を融合し，複雑な自然現象・社会現象を解明しようという研究の方向性が生まれ，現在，理論・応用の両側面で数多くの非線形モデリング手法が提案されている．

Fisher の線形判別分析を非線形へ拡張する手法の一つとして，Hastie et al. (1994)は，スプライン加法モデル(Buja et al. (1989))を利用した Flexible Discriminant Analysis (FDA)を提唱した．Hastie et al. は，正準相関分析(Anderson (1984))の理論に基づいて，Fisher の線形判別が線形回帰分析の枠組みで実行できることを示し，線形回帰モデルをノンパラメトリック回帰モデルに拡張することで非線形判別関数を構成している．加法モデルの利点として，各変数の寄与が容易に解釈できるということが挙げられる．しかし，実際問題への適応に当たり，しばしば直面する問題として，十分な数のデータが観測・測定できない場合や，極めて高次元のデータの分析を必要とする場合，たとえ十分大きい平滑化パラメータを用いたとしても，加法モデルの自由度が標本数を上回り，計算上の困難が生じることがある．

この様な背景をふまえて，Roth and Steinhage (1999)は，カーネル関数に基づく FDA を考案し，平滑化パラメータ等の選択を VC 次元(Vapnik (1998))に基づき選択する手法を提唱した．しかし，Roth and Steinhage の手法は，判別関数のパラメータ数と標本数の関係を考慮しておらず，統計的問題が否めない．さらに，VC 次元は非常に粗い最悪評価規準であることから，実際的にはさらに精密な評価が望まれる．加えて，VC 次元には，交差検証法(Stone (1974))を優越する明白な利点は無いとの報告もある(Hastie et al. (2001))．

本稿では，標本数とモデルのパラメータ数を考慮したカーネル関数に基づく Kernel Flexible Discriminant Analysis (KFDA)を提案する．KFDA の数理的背景には，Support Vector Machine (Vapnick (1998))で利用されるカーネルトリックが存在し，観測データを超高次元空間(特徴空間)へ写像して特徴抽出することにより，極めて判別能力の高い非線形判別関数を構成することができる．さらに，非線形判別関数の信頼性，汎化能力を向上させるため，パラメータ推定においては平滑化法を適用する．非線形判別手法の構築において本質的となる点は，平滑化パラメータ等の選択である．一つの方法として，交差検証法を用いる方法も考えられるが，大規模なデータ解析では，計算量が膨大となってしまう．Hastie et al. (1994)は，計算量を考慮して一般化交差検証法(Generalized Cross Validation (GCV)，Wahba (1990))を利用している．しかし，Hastie et al. の提案する GCV には未知の量が含まれていることから，GCV 規準そのものを最初に最適化する必要があり，利用に際しては慎重な検討を要する．

そこで，本稿では KFDA モデルと尤度概念の関係を明らかにし，平滑化パラメータなどの選択を情報量，及びベイズ理論の観点から考察してモデル評価規準を解析的に与える．

本稿の構成は以下の通りである．2 節では，Fisher の線形判別分析の概略を述べ，線形回帰分析の枠組みで Fisher の線形判別関数を構成する．3 節では，Fisher の線形判別分析を非線形へと拡張した KFDA を提案し，その数理的解釈を与える．また，人工データの解析を通して，KFDA の性質について述べる．4 節では，平滑化法によって構成された KFDA の評価をおこなうためのモデル評価規準を，情報量，及びベイズ理論の二つの異なる概念に基づいて導出する．5 節では，音声認識，文字認識など各分野で蓄積されつつある実データの解析，および人工データの解析を通して，これまでに提唱された様々な手法と提案する手法を比較して，その有効性を検証する．

2. 線形判別分析についての概略

2.1 Fisher の線形判別関数

本節では，p 次元特徴ベクトル $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)' \in R^p$ と群のラベル $G \in \{1, 2, \ldots, L\}$ に関する n 個の観測データ $\{(\boldsymbol{x}_\alpha, g_\alpha); \alpha = 1, \ldots, n\}$ に基づいて Fisher の線形判別関数を構成する．Fisher の判別分析法は，p 次元特徴ベクトル $\boldsymbol{x}$ をパラメータ $W = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_q)$, $\boldsymbol{w}_j = (w_{j1}, \ldots, w_{jp})'$ に基づく線形変換により，q 次元部分空間 $W'\boldsymbol{x} \in R^q$ へ射影し，この空間で判別をおこなう．その本質的な点は，q 次元部分空間での群間分散を可能な限り大きく，逆に，群内分散を可能な限り小さくしようとすることにある．通常，部分空間の次元は，視覚化が可能な $q = 2, 3$ とすることが多い．

いま，第 k 群に属する観測データ数を n_k とすると，観測データが得られた p 次元空間での群間分散 S_B と群内分散 S_W は以下で与えられる統計量である．

$$S_B = \sum_{k=1}^{L} \frac{n_k}{n} (\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{m}_k - \boldsymbol{m})',$$

$$S_W = \frac{1}{n} \sum_{k=1}^{L} \sum_{g_\alpha = k} (\boldsymbol{x}_\alpha - \boldsymbol{m}_k)(\boldsymbol{x}_\alpha - \boldsymbol{m}_k)',$$

$$\boldsymbol{m}_k = \frac{1}{n_k} \sum_{g_\alpha = k} \boldsymbol{x}_\alpha, \quad k = 1, 2, \ldots, L, \quad \boldsymbol{m} = \frac{1}{n} \sum_{\alpha=1}^{n} \boldsymbol{x}_\alpha.$$

ただし，$\sum_{g_\alpha = k}$ は第 k 群に属するデータに対して和をとるものとし，$\boldsymbol{m}_k$ は群 k の平均ベクトル，$\boldsymbol{m}$ は全体の平均ベクトルである．本稿では，あらかじめ，全体の平均ベクトルは $\boldsymbol{m} = \boldsymbol{0}$ と基準化されているものとする．

このとき，q 次元部分空間での群間分散 S_B^*，群内分散 S_W^* はそれぞれ，$S_B^* = W'S_BW$, $S_W^* = W'S_WW$ となり，パラメータ W は次の制約付き最適化により推定される．

$$\text{maximize} \quad J(W) = \mathrm{tr} S_B^* \quad \text{s.t.} \quad S_W^* = I_q \,. \tag{2.1}$$

この解は，次の固有値問題

$$S_W^{-1} S_B W = W\Lambda, \quad \Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_q\}$$

に帰着し，最適解 $\hat{W}$ は $S_W^{-1}S_B$ の固有値のうち，大きいほうから q 個の固有値 $\lambda_1, \ldots, \lambda_q$ に対応する固有ベクトル $\hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_q$ となる．部分空間の次元 q は，最大限 $\min\{p, L-1\}$ 次元までとることができ，将来観測されるデータ $\boldsymbol{x}$ は，部分空間へと射影され(つまり，$\hat{W}'\boldsymbol{x}$)，部分空間での各群の平均 $\hat{W}'\boldsymbol{m}_k$, $(k = 1, 2, \ldots, L)$ との距離が最も小さい群へ判別される．

以上が Fisher の線形判別分析法の概略であるが，この手法については，これまで多数の優れた書物が出版されているので(Mardia et al. (1979)，柳井・高根(1985)，石井 他(1998)など)，これ以上は深入りせず，線形回帰分析を利用した Fisher の線形判別へと話題を移す．

2.2 線形回帰分析に基づく Fisher の線形判別分析

本節では，Fisher の線形判別関数が回帰分析の枠組みで構成できることを紹介する．いま，$\boldsymbol{\theta}(k) : G \to R^q$ を第 k 群の中心を表す関数とする．前節では，n 個の観測データ $\{(\boldsymbol{x}_\alpha, g_\alpha); \alpha = 1, \ldots, n\}$ が与えられたとき，(2.1) の制約付き最適化問題を解くことで Fisher の線形判別関数 $W'\boldsymbol{x}$ を 構成した．あらかじめ，全体の平均ベクトルは $\boldsymbol{m} = \boldsymbol{0}$ とされているとき，回帰分析の枠組みにおいては，次の二乗誤差和を最小にするようにパラメータ W は推定される．

$$ASR(\Theta, W) = \sum_{\alpha=1}^{n} ||\boldsymbol{\theta}(g_\alpha) - W'\boldsymbol{x}_\alpha||^2 = \mathrm{tr}||Z\Theta - XW||^2 \,. \tag{2.2}$$

ここで，$X=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)'$, $\Theta=(\boldsymbol{\theta}(1),\ldots,\boldsymbol{\theta}(L))'$ とし，$n\times L$ 行列 Z は，データがどの群に属するかを表す行列で，その (α,k) 成分 $z_{\alpha,k}$ は，データ $\boldsymbol{x}_\alpha$ が k 群に属するとき $z_{\alpha,k}=1$ となり，それ以外の群に属するときは $z_{\alpha,k}=0$ となる．つまり，$Z\Theta$ の第 α 列には，データ $\boldsymbol{x}_\alpha$ が属する群の中心がくることになる．

群の中心を表す行列 Θ を固定したとき，(2.2) 式を最小とする $\hat{W}$ は，$ASR(\Theta,W)$ を W で微分し，$X'XW-X'Z\Theta=O$ を解くことで与えられる．

$$\hat{W}=(X'X)^{-1}X'Z\Theta\,.$$

推定量 $\hat{W}$ は q 次元部分空間での基底を構成するが，部分空間内での群の中心を表す Θ を決定する必要がある．そこで，射影子行列 $P_X=X(X'X)^{-1}X'$ と推定量 $\hat{W}$ を用いて (2.2) 式を変形すると，以下の式となる．

$$ASR(\Theta,\hat{W})=\mathrm{tr}\{\Theta'Z'Z\Theta\}-\mathrm{tr}\{\Theta'Z'P_XZ\Theta\}\,. \tag{2.3}$$

Hastie et al. (1994) は，Θ の正規化の条件として，$\Theta'Z'Z\Theta/n=I_q$ という制約を課し，(2.3) 式の第二項 $\mathrm{tr}\{\Theta'Z'P_XZ\Theta\}$ を最大化する Θ を求めることに問題を帰着させた．この問題の解 $\hat{\Theta}$ は，正規化の条件のもとで $Z'P_XZ$ を固有値分解したときの，大きい固有値 $\alpha_1>\alpha_2>\cdots>\alpha_q$ に対応する固有ベクトルで構成される．

正準相関分析の理論から，回帰分析の枠組みで求めた推定量 $\hat{W}=(X'X)^{-1}X'Z\hat{\Theta}$ は，(2.1) 式を解くことで構成した Fisher 線形判別関数の推定量 $\hat{W}_{\mathrm{Fisher}}$ と以下の関係がある（Mardia et al. (1979), Hastie et al. (1994, 1995)）．

$$\hat{W}_{\mathrm{Fisher}}=\hat{W}D\,. \tag{2.4}$$

ここで，D は q 次元対角行列で，その第 (k,k) 成分は $d_{k,k}=1/(\alpha_k(1-\alpha_k^2)^{1/2})$ である．この関係を用いることで，Fisher の線形判別関数は構成され，将来のデータは各群の平均との距離が最小となる群に判別される．

ここで注意すべきことは，前節の Fisher の線形判別分析と照らし合わせて分かるように，中心行列 Θ は実質的にパラメータの構造を持たず，回帰分析の枠組みで求めた推定量 $\hat{W}$ を用いて，$D\hat{W}'\boldsymbol{m}_k=\hat{\boldsymbol{\theta}}(k)$, $k=1,\ldots,L$ という式で表されることである．Hastie et al. (1994)は，中心行列 Θ をパラメータと捉えず，射影子行列 P_X を与えたとき，自動的に決定される量であるという点に着目し，GCV に基づいてモデルの予測精度，つまり，データに潜む未知の構造に対する推定量 $\hat{W}$ の適合度を評価している．

以上のように，回帰分析の枠組みで判別関数を構成したが，変数間に線形性を仮定する線形回帰モデルでは，その性質上，複雑なデータの分析に対して有効に機能せず，より柔軟なモデルが必要とされる．次節では，非線形回帰分析とカーネル法を融合した KFDA を提案する．

3. 非線形回帰分析を利用した判別関数の構成

3.1 Kernel Flexible Discriminant Analysis

前節で紹介した線形判別分析においては，観測データを線形変換する判別関数を作成していた．ここでは線形回帰分析に基づく Fisher の線形判別分析を非線形モデルへ拡張するために，観測データ $\boldsymbol{x}=(x_1,\ldots,x_p)'$ にカーネル関数に基づく非線形変換

$$\boldsymbol{\phi}:=\boldsymbol{x}\to\boldsymbol{\phi}(\boldsymbol{x})=(\phi_1(\boldsymbol{x}),\ldots,\phi_s(\boldsymbol{x}))',\ (R^p\to R^s) \tag{3.1}$$

を施して空間 R^s に射影し，この空間からの線形変換 $W'\boldsymbol{\phi}(\boldsymbol{x})=(\boldsymbol{w}_1'\boldsymbol{\phi}(\boldsymbol{x}),\ldots,\boldsymbol{w}_q'\boldsymbol{\phi}(\boldsymbol{x}))'$ による q 次元部分空間への射影を構成して判別を行う．部分空間の次元 q は，最大限 $\min\{s,L-1\}$

次元までとることができ，特に群の個数 L に比べて特徴変数 $\boldsymbol{x}$ の次元 p が比較的小さいとき，Fisher の線形判別法は，p 次元部分空間までしか判別関数に情報を取り入れることが出来なかったが，KFDA では，非線形変換における次元 s を大きく取ることで，常に $L-1$ 次元部分空間で判別が可能であり，非線形構造を持つデータに対してより柔軟な分析をおこなえるという利点がある．

本稿では，(3.1)式のカーネル関数 $\phi_j(\boldsymbol{x})$, $j=1,\ldots,s$ として，次のガウスカーネルを用いる．

$$\phi_j(\boldsymbol{\mu}_j;\boldsymbol{x})=\exp\left(-\sum_{k=1}^{p}\frac{(x_k-\mu_{jk})^2}{2\sigma_k^2}\right),\quad j=1,2,\ldots,s\,. \tag{3.2}$$

ただし，$\boldsymbol{\mu}_j=(\mu_{j1},\ldots,\mu_{jp})'$ はカーネルの位置を定める p 次元中心ベクトル，$\boldsymbol{\sigma}=(\sigma_1^2,\ldots,\sigma_p^2)'$ はカーネルの分散を制御する形状パラメータである．ガウスカーネルの有用性・利便性は，非線形回帰・非線形判別問題などの様々なデータ解析を通じて報告されている(安道 他(2001, 2002)，Ando et al. (2002))．

(3.2)式のガウスカーネルに含まれる中心ベクトル $\boldsymbol{\mu}_j$ は，観測データの密度関数を大まかに捉えるように決定する．ここでは，n 個の p 次元観測データ $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ を k-means 法によって，s 個のクラスター $A_1,\ldots,A_s$ に分割し，各クラスター A_j に含まれる n_j 個のデータに基づいて $\boldsymbol{\mu}_j$ を次のように決定する．

$$\boldsymbol{\mu}_j=\frac{1}{n_j}\sum_{\boldsymbol{x}_\alpha\in A_j}\boldsymbol{x}_\alpha,\qquad j=1,\ldots,s\,. \tag{3.3}$$

従って，形状パラメータ $\boldsymbol{\sigma}$ を与えると s 個の既知のガウスカーネルが構成されることになる．形状パラメータ $\boldsymbol{\sigma}$ の与え方については，後節で話題にする．以下，一般性を失うことなく，s 次元空間での平均ベクトルは $\bar{\boldsymbol{\phi}}=\sum_{\alpha=1}^{n}\boldsymbol{\phi}(\boldsymbol{x}_\alpha)/n=\boldsymbol{0}$ となるように平行移動したものとして，モデルのパラメータ W の推定法について述べる．

2.2 節で説明した線形回帰の枠組みと同様，n 個の各観測データの中心ベクトル $\boldsymbol{\theta}(g_\alpha)$ からの距離を最小とするように W は推定されるが，非線形モデルの柔軟さ故にデータに強く依存したモデルが構成されてしまい，その結果，汎化能力が充分でないことがしばしばある．そこで，モデルの信頼性，汎化能力を向上させるため，パラメータ W の推定には平滑化法を適用する．

$$ASR(\Theta,W)=\sum_{\alpha=1}^{n}||\boldsymbol{\theta}(g_\alpha)-W'\boldsymbol{\phi}(\boldsymbol{x}_\alpha)||^2+\frac{n\lambda}{2}\sum_{k=1}^{q}\boldsymbol{w}_k'\boldsymbol{w}_k\,. \tag{3.4}$$

ここで，λ は平滑化パラメータと呼ばれ，データへの適合度とモデルの複雑さを調整する働きをもつとともに，モデル推定の安定性に寄与する機能をもつ．また，ベイズ理論の観点から考察すると，パラメータ $\boldsymbol{w}_k$ の事前分布として平均 $\boldsymbol{0}$, 分散 $I_s/(n\lambda)$ の正規分布を仮定していることになる．

このとき，(3.4)式の最小化に基づく推定量 $\hat{W}$ は，

$$\hat{W}=(\Phi'\Phi+n\lambda I_s)^{-1}\Phi' Z\Theta \tag{3.5}$$

で与えられる．ただし，$\Phi=(\boldsymbol{\phi}(\boldsymbol{x}_1),\ldots,\boldsymbol{\phi}(\boldsymbol{x}_n))'$，$I_s$ は s 次元単位行列，Z は前節で導入したデータが属する群を表す行列とする．

いま $P_\Phi=\Phi(\Phi'\Phi+n\lambda I_s)^{-1}\Phi'$ を射影子行列とすると，各群の中心 Θ は，正規化の条件 $\Theta' Z'Z\Theta/n=I_q$ のもとで $Z'P_\Phi Z$ を固有値分解したときの，大きい固有値 $\alpha_1>\alpha_2>\cdots>\alpha_q$ に対応する固有ベクトルで構成される．この $\hat{\Theta}$ を代入して得られる $\hat{W}=(\Phi'\Phi+n\lambda I_s)^{-1}\Phi' Z\hat{\Theta}$

を用いて，非線形判別関数 $D\hat{W}'\boldsymbol{\phi}(\boldsymbol{x})$ が構築され，将来の観測データ $\boldsymbol{x}$ は $D\hat{W}'\boldsymbol{\phi}(\boldsymbol{x})$ と各群の中心 $D\hat{W}'\bar{\boldsymbol{\phi}}_k, k=1,\ldots,L$ との距離が最小になる群に判別される．ただし，$\bar{\boldsymbol{\phi}}_k = \frac{1}{n_k}\sum_{g_\alpha=k}\boldsymbol{\phi}(\boldsymbol{x}_\alpha)$ は，第 k 群の平均ベクトル，q 次元対角行列 D の第 (k,k) 成分は $d_{k,k}=1/(\alpha_k(1-\alpha_k^2)^{1/2})$ とする．

3.2 KFDA の数理的背景

本節では，KFDA の背後にある数理的構造を理論的に明確にしていく．いま，超高次元空間(特徴空間)への写像 $\boldsymbol{\psi}: R^p \to \mathcal{F}, \boldsymbol{x} \to \boldsymbol{\psi}(\boldsymbol{x})$ によって，p 次元観測データ $\boldsymbol{x}$ を特徴空間 $\mathcal{F}$ へ写像し，その写像した空間における線形判別関数 $\Gamma'\boldsymbol{\psi}(\boldsymbol{x}) = (\boldsymbol{\gamma}_1'\boldsymbol{\psi}(\boldsymbol{x}),\ldots,\boldsymbol{\gamma}_q'\boldsymbol{\psi}(\boldsymbol{x}))', \Gamma=(\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_q)$, $\boldsymbol{\gamma}_j=(\gamma_{j1},\gamma_{j2},\ldots)'$ に基づき特徴ベクトル $\boldsymbol{\psi}(\boldsymbol{x})$ の判別を考える．一般に，特徴空間 $\mathcal{F}$ は極めて高次元，時には無限次元である場合もある(Schölkopf et al. (1998))．KFDA は，$\mathcal{F}$ において観測データの特徴抽出をおこない，この特徴の判別を q 次元空間でおこなっていることを以下に示していく．

まず，特徴空間 $\mathcal{F}$ での群間分散 $S_B^{\mathcal{F}}$ と群内分散 $S_W^{\mathcal{F}}$, 第 k 群の中心 $\bar{\boldsymbol{\psi}}_k$ は以下で与えられる．(ここでは，特徴空間 $\mathcal{F}$ での平均ベクトル $\bar{\boldsymbol{\psi}}=\sum_{\alpha=1}^n\boldsymbol{\psi}(\boldsymbol{x}_\alpha)/n$ は $\bar{\boldsymbol{\psi}}=\boldsymbol{0}$ と基準化されているものとする．)

$$
\begin{aligned}
S_B^{\mathcal{F}} &= \sum_{k=1}^{L}\frac{n_k}{n}\bar{\boldsymbol{\psi}}_k\bar{\boldsymbol{\psi}}_k', \\
S_W^{\mathcal{F}} &= \frac{1}{n}\sum_{k=1}^{L}\sum_{g_\alpha=k}(\boldsymbol{\psi}(\boldsymbol{x}_\alpha)-\bar{\boldsymbol{\psi}}_k)(\boldsymbol{\psi}(\boldsymbol{x}_\alpha)-\bar{\boldsymbol{\psi}}_k)', \\
\bar{\boldsymbol{\psi}}_k &= \frac{1}{n_k}\sum_{g_\alpha=k}\boldsymbol{\psi}(\boldsymbol{x}_\alpha), \quad k=1,2,\ldots,L.
\end{aligned}
$$

このとき，$S_B^{\mathcal{F}}$ と $S_W^{\mathcal{F}}$ の比を最大化する推定量 $\hat{\Gamma}$ は，Fisher の線形判別の枠組みと同様に，条件 $\Gamma' S_W^{\mathcal{F}}\Gamma = I_q$ の下で，$\mathrm{tr}\,\Gamma' S_B^{\mathcal{F}}\Gamma$ を最大化することで得られる．この推定量を用いて，ラベルが未知のデータ $\boldsymbol{x}$ は，$\hat{\Gamma}'\boldsymbol{\psi}(\boldsymbol{x})$ が各群の中心 $\hat{\Gamma}'\bar{\boldsymbol{\psi}}_k$ に最も近い群へ判別される．

一般に，特徴空間 $\mathcal{F}$ は極めて高次元であるため，推定量 $\hat{\Gamma}$ を得るのに必要な計算量が膨大になってしまうことが容易に予想される．しかしながら，特徴ベクトルの内積がカーネル関数の値 $\phi(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{\psi}(\boldsymbol{x})'\boldsymbol{\psi}(\boldsymbol{y})$ で定義される特徴変換 $\boldsymbol{\psi}(\boldsymbol{x})$ を利用することで計算量を大幅に減少させることができる．この手法は，カーネルトリック(Schölkopf et al. (1998)，Vapnik (1998))と呼ばれ，その有用性は極めて高い．最もポピュラーなカーネルは，(3.2)式のガウスカーネルであり，その他にも多項式カーネル，シグモイドカーネルなどが挙げられる．

パラメータ $\boldsymbol{\gamma}_j$ の構造は，$w_{j\cdot}$ を係数とした $\boldsymbol{\gamma}_j=\sum_{\alpha=1}^n w_{j\alpha}\boldsymbol{\psi}(\boldsymbol{x}_\alpha), j=1,\ldots,q$ で与えられることが一般に知られているが(Roth and Steinhage (1999))，ここでは，データ数とパラメータ数が接近することを避けるために，次の構造 $\boldsymbol{\gamma}_j=\sum_{i=1}^s w_{ji}\boldsymbol{\psi}(\boldsymbol{\mu}_i)$ を提案する．ただし $\boldsymbol{\mu}_i$ は，観測データを k-means 法で s 個のクラスターに分割したときにおける各クラスターの中心(3.3)である．

このとき，簡単な式変形から

$$
\begin{aligned}
\boldsymbol{\gamma}_j' S_B^{\mathcal{F}}\boldsymbol{\gamma}_j &= \left(\sum_{i=1}^{s}w_{ji}\boldsymbol{\psi}(\boldsymbol{\mu}_i)\right)'\left(\sum_{k=1}^{L}\frac{n_k}{n}\bar{\boldsymbol{\psi}}_k\bar{\boldsymbol{\psi}}_k'\right)\left(\sum_{i'=1}^{s}w_{ji}\boldsymbol{\psi}(\boldsymbol{\mu}_{i'})\right) \\
&= \sum_{i=1}^{s}\sum_{i'=1}^{s}\sum_{k=1}^{L}w_{ji}\frac{n_k}{n}\left(\frac{1}{n_k}\sum_{g_\alpha=k}\boldsymbol{\psi}(\boldsymbol{\mu}_i)'\boldsymbol{\psi}(\boldsymbol{x}_\alpha)\right)\left(\frac{1}{n_k}\sum_{g_\alpha=k}\boldsymbol{\psi}(\boldsymbol{\mu}_{i'})'\boldsymbol{\psi}(\boldsymbol{x}_\alpha)\right)' w_{ji'}
\end{aligned}
$$

$$= \sum_{i=1}^{s}\sum_{i'=1}^{s}\sum_{k=1}^{L} w_{ji}\frac{n_k}{n}\left(\frac{1}{n_k}\sum_{g_\alpha=k}\phi(\boldsymbol{\mu}_i,\boldsymbol{x}_\alpha)\right)\left(\frac{1}{n_k}\sum_{g_\alpha=k}\phi(\boldsymbol{\mu}_{i'},\boldsymbol{x}_\alpha)\right)w_{ji'}$$
$$= \boldsymbol{w}_j' S_B^\phi \boldsymbol{w}_j, \qquad (j=1,\ldots,q)$$

が成立し，同様に $\boldsymbol{\gamma}_j' S_W^{\mathcal{F}} \boldsymbol{\gamma}_j = \boldsymbol{w}_j' S_W^\phi \boldsymbol{w}_j$, $(j=1,\ldots,q)$, $\boldsymbol{\gamma}_j'\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{w}_j'\boldsymbol{\phi}(\boldsymbol{x})$ の関係が導かれる．ただし，$\boldsymbol{\phi}(\boldsymbol{x})$ は，(3.1) 式による非線形変換，S_B^ϕ, S_W^ϕ は，写像 $\boldsymbol{\phi}(\boldsymbol{x})$ による変換後のデータ $\boldsymbol{\phi}(\boldsymbol{x}_1),\ldots,\boldsymbol{\phi}(\boldsymbol{x}_n)$ の群間分散 $S_B^\phi = \sum_{k=1}^{L}\frac{n_k}{n}\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k'$, 群内分散 $S_W^\phi = \frac{1}{n}\sum_{k=1}^{L}\sum_{g_\alpha=k}(\boldsymbol{\phi}(\boldsymbol{x}_\alpha) - \bar{\boldsymbol{\phi}}_k)(\boldsymbol{\phi}(\boldsymbol{x}_\alpha)-\bar{\boldsymbol{\phi}}_k)'$ とし，$\bar{\boldsymbol{\phi}}_k$ は第 k 群の平均ベクトル $\bar{\boldsymbol{\phi}}_k = \frac{1}{n_k}\sum_{g_\alpha=k}\boldsymbol{\phi}(\boldsymbol{x}_\alpha)$ である．

すなわち，制約条件 $\Gamma' S_W^{\mathcal{F}}\Gamma = I_q$ の下で，$\mathrm{tr}\,\Gamma' S_B^{\mathcal{F}}\Gamma$ を最大にする推定量 $\hat{\Gamma}$ に基づく判別関数 $\hat{\Gamma}'\boldsymbol{\psi}(\boldsymbol{x})$ [A] は，$W' S_W^\phi W = I_q$ の下で，$\mathrm{tr}\,W' S_B^\phi W$ を最大にする推定量 $\hat{W} = (\hat{\boldsymbol{w}}_1,\ldots,\hat{\boldsymbol{w}}_q)'$, つまり，$S_W^{\phi\,-1} S_B^\phi$ の大きいほうから q 個の固有値に対応する固有ベクトル $\hat{\boldsymbol{w}}_1,\ldots,\hat{\boldsymbol{w}}_q$ を用いる判別関数 $\hat{W}'\boldsymbol{\phi}(\boldsymbol{x})$ [B] と等価であることを示唆している．さらに，KFDA の枠組み [C] においては，(3.5)式で $\lambda = 0$ とし，(2.4)式の関係を用いると，[A]・[B] の判別関数が構成され，このときは，これら 3 つの判別方式 [A] ∼ [C] は一致する．

以上をまとめると，KFDA の枠組みにおいて，(3.1)式のカーネル関数に基づく非線形変換 $\boldsymbol{\phi}(\boldsymbol{x})$; $R^p \to R^s$ は，R^s へ射影した空間を経由して判別関数を構成しているが，そこに内在する理論においては，KFDA は超高次元空間の特徴空間 $\mathcal{F}$ へ観測データを写像して特徴の抽出をおこない，その特徴ベクトルを利用して判別を実行しているのである．次節では，KFDA の性質について説明していく．

注意． ここでは，特徴空間 $\mathcal{F}$ での平均ベクトル $\bar{\boldsymbol{\psi}}$ を $\bar{\boldsymbol{\psi}} = \mathbf{0}$ と基準化しているが，基準化されていない場合でも，いま展開した理論と同様な議論が可能である(Schölkopf et al. (1998))．

3.3　KFDA の性質について

本節では，人工データである Synthetic Data (Ripley (1994, 1996))の解析を通して，平滑化パラメータ λ と(3.2)式のガウスカーネルの形状パラメータ $\boldsymbol{\sigma}$ の設定が，どのような影響を判別境界の非線形性・複雑度，あるいは安定性に与えるかを検証する．

Synthetic Data の特徴変数は $p=2$ 次元で，2 群からなり，250 個の学習データと 1000 個の予測データが与えられている．図 1 は，学習データ(○: $g_\alpha = 1$, △: $g_\alpha = 2$)，および真の境界領域(- - -)を示している．

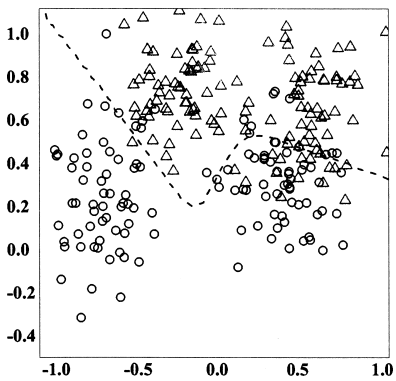


図 1.　Synthetic Data (○: $g_\alpha = 1$, △: $g_\alpha = 2$) と真の境界領域 (- - -).

まず,(3.2)式のガウスカーネルに含まれる形状パラメータ $\boldsymbol{\sigma}$ の設定が, どのような影響を判別関数に与えるかを説明する．KFDA の非線形変換の次元を $s = 20$, 平滑化パラメータ $\lambda = 10^{-2}$ と固定したもとで, 学習データを用いて, 形状パラメータの値を $\boldsymbol{\sigma} = (0.01, 0.005), (0.1, 0.05), (1, 0.5)$ として非線形判別関数 $\hat{W}'\boldsymbol{\phi}(\boldsymbol{x})$ を構成した．ただし，$\hat{W}$ は(3.5)式で与えられる推定量であり，2 群判別 $(L = 2)$ を考えているので，部分空間の次元は $q = \min\{s, L-1\} = 1$ となる．

図 2 は，推定結果を表しているが，形状パラメータの値を小さくするにつれて，判別境界線が複雑になっていくことがわかる．これは，p 次元観測データに(3.1)式のカーネル関数に基づく非線形変換を施して s 次元空間へと数式の上では常に写像しているが，3.2 節で紹介した特徴空間 $\mathcal{F}$ で考えると，形状パラメータの値を小さくすることで，$\mathcal{F}$ の次元が超高次元になることを示している．つまり，ガウスカーネルの形状パラメータを利用することで，特徴空間の次元を調節することができるのである．また,(3.1)式の非線形変換を施したときの s 次元空間については，s が大きくなるに比例して最終的な判別境界が複雑になることは自明である．

平滑化パラメータは，判別境界の非線形性・複雑度に影響を与える他に，パラメータ推定における安定性に寄与している．安定性への影響を見るために，$\log_{10}(\lambda) = -2, -3, -5$ と与えて，250 個の学習データから 10 組のブートストラップ標本(標本サイズ; $n = 250$)を発生させ，各ブートストラップ標本に基づいて判別境界曲線を推定した．ただし，ブートストラップ標本は，データの組を復元抽出することで構成し，形状パラメータの値を $\boldsymbol{\sigma} = (0.1, 0.05)$, 非線形変換の次元を $s = 20$ と固定している．

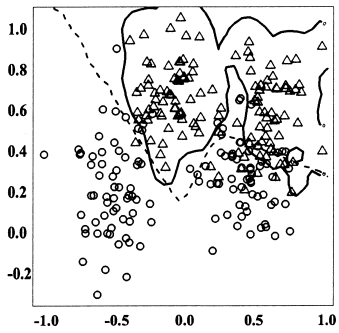
図 3 は，推定された判別境界曲線を示しているが，学習データへの適合度を良くするために，平滑化パラメータの値を小さくとり過ぎると，判別境界曲線が暴れてしまうことがわかる．ベイズの観点から考えると,(3.4)式の最小化に基づく推定方法は，パラメータ W に分散 $I_s/(n\lambda)$ の正規分布を事前分布として仮定していることになるので，平滑化パラメータ λ を小さくすると，パラメータの分散(つまり，判別関数の分散)が大きくなるので，これは自然な結果である．また，$\lambda = 0$ (つまり，最小二乗法)として，パラメータを推定したが，多くの場合(3.5)式の逆行列が退化してしまい，推定量 $\hat{W}$ は得られなかった．つまり，罰則項を付与することで，このような計算上の問題も避けることが可能なのである．このように，平滑化パラメータの選択が安定性の意味で重要であることは，図 3 から一目瞭然であるし，数値的にも認識される．

以上のように，真の構造を捉えるためには，非線形空間の次元 s, 平滑化パラメータ λ，及びガウスカーネルに含まれる形状パラメータ $\boldsymbol{\sigma}$ の値を適切に選択する必要がある．次節では，モデル評価規準を導出し，s, λ, $\boldsymbol{\sigma}$ を選択する手法について述べる．

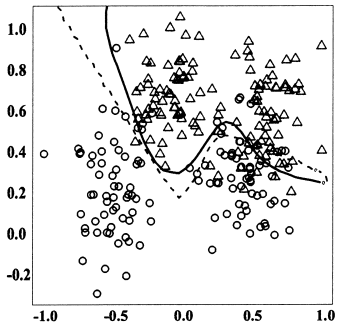
4. モデル評価規準

近年，コンピュータの急速な発展は，多様な統計的分析手法の開発を支援してきたが，それにともない，各々の統計モデルを評価するための規準が必要となってきた．統計的モデル選択問題は，変数選択，ニューラルネットの素子数決定，カーネル関数の最適バンド幅決定，決定木のノード数選択，変化点の検出など，非常に幅広い応用範囲を持ち，その重要性から，現在，理論的・応用的側面の両面で広く研究されている．本節では，KFDA と尤度概念を結びつけ，情報量及びベイズ理論の観点からモデル評価規準を導出する．

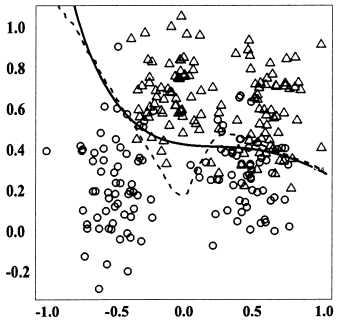
前節では，非線形回帰分析の枠組みによって，モデルに含まれるパラメータ W を (3.4) 式の最小化に基づいて推定した.(3.4)式の第一項目はデータに対する適合度であり，二乗誤差和であることから，q 変量正規分布の密度関数を仮定していると考える．



(a) $\boldsymbol{\sigma} = (0.01, 0.005)$

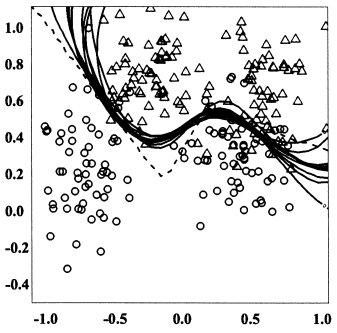


(b) $\boldsymbol{\sigma} = (0.1, 0.05)$

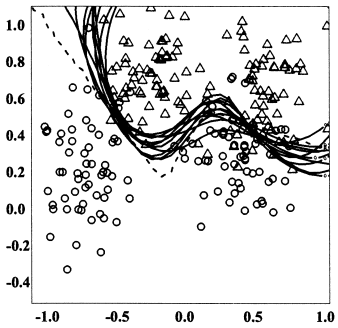


(c) $\boldsymbol{\sigma} = (1, 0.5)$

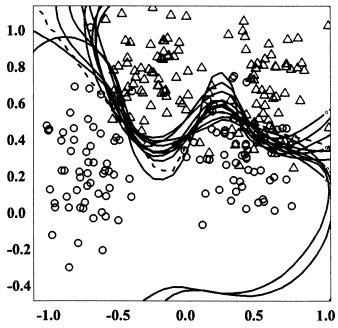
図 2. 推定した判別境界曲線.



(a) $\log_{10}(\lambda) = -2$



(b) $\log_{10}(\lambda) = -3$



(c) $\log_{10}(\lambda) = -5$

図 3. 推定した判別境界曲線.

$$f(\boldsymbol{\theta}(g_\alpha)|\boldsymbol{x}_\alpha; W, \Sigma)$$
$$= \frac{1}{(2\pi)^{q/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\theta}(g_\alpha) - W'\boldsymbol{\phi}(\boldsymbol{x}_\alpha))'\Sigma^{-1}(\boldsymbol{\theta}(g_\alpha) - W'\boldsymbol{\phi}(\boldsymbol{x}_\alpha))\right].$$

ただし，$\Sigma = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_q^2\}$ である．すなわち，(3.4)式の最小化により非線形判別関数を構

成することは，次の罰則付き対数尤度関数の最大化により統計モデル $f(\boldsymbol{\theta}(g_\alpha)|\boldsymbol{x}_\alpha; W, \Sigma)$ を推定することに対応する．

$$l_\lambda(W,\Sigma) = \sum_{\alpha=1}^{n} \log f(\boldsymbol{\theta}(g_\alpha)|\boldsymbol{x}_\alpha; W, \Sigma) - \frac{n\lambda}{2} \sum_{k=1}^{q} \boldsymbol{w}_k' \boldsymbol{w}_k \,. \tag{4.1}$$

推定量 $\hat{W}$ と群の中心 $\boldsymbol{\theta}(g_\alpha) = (\theta_1(g_\alpha), \ldots, \theta_q(g_\alpha))$ は，3.1 節の方法で求められ，誤差分散 σ_k^2 は以下のように推定される．

$$\hat{\sigma}_k^2 = \frac{1}{n} \sum_{\alpha=1}^{n} \left\{ \theta_k(g_\alpha) - \hat{\boldsymbol{w}}_k' \boldsymbol{\phi}(\boldsymbol{x}_\alpha) \right\}^2 , \qquad k = 1, \ldots, q \,.$$

以上のように，統計モデル $f(\boldsymbol{\theta}(g)|\boldsymbol{x}; \hat{W}, \hat{\Sigma})$ は構成され，KFDA に尤度概念を与えた．この統計モデル $f(\boldsymbol{\theta}(g)|\boldsymbol{x}; \hat{W}, \hat{\Sigma})$ に含まれる未知の量，非線形空間の次元 s, 平滑化パラメータ λ, 及びガウスカーネルに含まれる形状パラメータ $\boldsymbol{\sigma}$ の値を適切に選択する方法を次節から提案する．

4.1 情報量規準

情報量規準(Akaike (1974)，Konishi and Kitagawa (1996))とは，想定したモデルとデータを生成した真のモデルとの Kullback-Leibler 距離(Kullback and Leibler (1951))を最小とするモデル，つまり期待対数尤度が最大となるモデルを選択し，予測の意味で優れたモデルを構築しようという考えに基づき導出されたモデル評価規準である．期待対数尤度は真の分布(未知)に依存することから，Akaike (1974)は対数尤度を利用して期待対数尤度を推定したときのバイアスを補正し，(期待対数尤度) ≈ (対数尤度) − (バイアス) と近似することを考えた．

その結果，想定したモデルが真のモデルに十分近いという仮定の下で，最尤法によって推定したモデルの期待対数尤度と対数尤度のバイアスは，モデルの自由パラメータ数で近似できるとし，情報量規準 AIC $= -2$ (対数尤度) $+ 2$ (パラメータ数) を提案した．しかし，AIC は最尤法に基づき構成したモデルを評価する規準であり，本稿では最尤法の枠を外した罰則付き最尤法に基づきモデルを構成していることから，単純に AIC を適用することには問題がある．

安道 他(2001)，Ando et al. (2002)は，目的変数が一変量の場合に，罰則付き最尤法に基づいて推定された正規回帰モデルに関する情報量規準 GIC (Konishi and Kitagawa (1996))を提案した．この結果を目的変数が多変量の正規回帰モデルに拡張すると，$l_\lambda(W,\Sigma)$ (4.1)式を最大化することで与えられる推定値 $\hat{W}$, $\hat{\Sigma}$ に基づく統計モデル $f(\boldsymbol{\theta}(g)|\boldsymbol{x}; \hat{W}, \hat{\Sigma})$ を評価する情報量規準 GIC は次で与えられる．

$$\mathrm{GIC}(s, \lambda, \boldsymbol{\sigma}) = -2 \sum_{\alpha=1}^{n} \log f(\boldsymbol{\theta}(g_\alpha)|\boldsymbol{x}_\alpha; \hat{W}, \hat{\Sigma}) + 2\mathrm{tr}\left\{ J^{-1} I \right\} . \tag{4.2}$$

ただし, I と J は $q(s+1)$ 次正方行列で, その $((k-1)(s+1)+1) \sim k(s+1)$ 列, $((l-1)(s+1)+1) \sim l(s+1)$ 行の成分は，次の $(s+1)$ 次正方行列で与えられる．

$$I_{kl} = \frac{1}{n} \begin{pmatrix} \Phi' \Lambda_k - \lambda \hat{\boldsymbol{w}}_k \mathbf{1}_n' \\ \boldsymbol{p}_k' \end{pmatrix} (\Lambda_l \Phi, \boldsymbol{p}_l) \,,$$

$$J_{kl} = \frac{1}{n\hat{\sigma}_k^2} \begin{pmatrix} \Phi'\Phi + n\hat{\sigma}_k^2 \lambda I_s & \Phi' \Lambda_k \mathbf{1}_n \\ \mathbf{1}_n' \Lambda_k \Phi & n/(2\hat{\sigma}_k^2) \end{pmatrix}, (k = l), \quad J_{kl} = O \,, \quad (k \neq l) \,.$$

ここで，$\Lambda_k = \mathrm{diag}\left[\theta_k(g_1) - \hat{\boldsymbol{w}}_k' \boldsymbol{\phi}(\boldsymbol{x}_1), \ldots, \theta_k(g_n) - \hat{\boldsymbol{w}}_k' \boldsymbol{\phi}(\boldsymbol{x}_n)\right] / \hat{\sigma}_k^2$, O は全ての成分が 0 の行

列，$\mathbf{1}_n = (1, 1, \ldots, 1)'$, n 次元ベクトル $\boldsymbol{p}_k$ の第 α 成分は

$$p_{k\alpha} = \left(\theta_k(g_\alpha) - \hat{\boldsymbol{w}}_k'\boldsymbol{\phi}(\boldsymbol{x}_\alpha)\right)^2 / 2\hat{\sigma}_k^4 - 1/2\hat{\sigma}_k^2$$

で与えられる．(4.2)式で与えた GIC を最小にするモデルを最適なモデルとして選択する．

Hastie and Tibshirani (1990)は射影子行列 $P_\Phi = \Phi(\Phi'\Phi + n\lambda I_s)^{-1}\Phi'$ を平滑化行列と呼び，モデルの自由度を平滑化行列の対角和 $\mathrm{tr}\{P_\Phi\}$ として定義した．この結果を用いて，平滑化行列の対角和をバイアス補正項として近似した次のモデル評価規準 AIC_M も考えられる．

$$\mathrm{AIC}_M(s, \lambda, \boldsymbol{\sigma}) = -2\sum_{\alpha=1}^{n} f(\boldsymbol{\theta}(g_\alpha)|\boldsymbol{x}_\alpha; \hat{W}, \hat{\Sigma}) + 2q\mathrm{tr}(\{P_\Phi\} + 1)\,. \tag{4.3}$$

つまり，最尤法に基づいて推定されたモデルの評価規準である AIC のバイアス補正項(モデルのパラメータ数)を平滑化行列の対角和で置き換えたものに相当する．AIC_M の計算は容易であるが，本来，AIC は最尤法に基づき構成したモデルを評価する規準である．本稿では，罰則付き最尤法に基づいてパラメータ推定がなされていることから，その妥当性について，またパラメータ数を平滑化行列の対角和で置き換えることの妥当性に関しても研究の余地がある．

4.2 ベイズ型情報量規準

AIC と同様に最尤法に基づき構成したモデルの評価規準として，Schwarz (1978)のベイズ型情報量規準(Bayesian Information Criterion; BIC)がある．ベイズ理論の観点から考えると，罰則付き対数尤度関数(4.1)式の最大化によりモデルを推定することはパラメータ W の事前分布として，次の s 次元正規分布の積を仮定することと対応している．

$$\pi(W|\lambda) = \prod_{k=1}^{q}\left(\frac{n\lambda}{2\pi}\right)^{s/2}\exp\left\{-\frac{n\lambda}{2}\boldsymbol{w}_k'\boldsymbol{w}_k\right\}. \tag{4.4}$$

モデルの事前確率を等しいと仮定し，モデルの事後確率に積分のラプラス近似を適用することでベイズの理論から導かれるベイズ型モデル評価規準 は次式で与えられる．

$$\begin{aligned}\mathrm{BIC}(s, \lambda, \boldsymbol{\sigma}) = {} & -2\sum_{\alpha=1}^{n}\log f(\boldsymbol{\theta}(g_\alpha)|\boldsymbol{x}_\alpha; \hat{W}, \hat{\Sigma}) + n\lambda\sum_{k=1}^{q}\hat{\boldsymbol{w}}_k'\hat{\boldsymbol{w}}_k + q\log n \\ & + \log|J| - q\log 2\pi - qs\log\lambda\,.\end{aligned} \tag{4.5}$$

ただし，J は(4.2)式で与えられるものとし，BIC を最小とするモデルを最適なモデルとして選択する．

また，AIC と同様に，Schwarz の BIC のパラメータ数を平滑化行列の対角和に置き換えた次のモデル評価規準 BIC_M も考えられる．

$$\mathrm{BIC}_M(s, \lambda, \boldsymbol{\sigma}) = -2\sum_{\alpha=1}^{n} f(\boldsymbol{\theta}(g_\alpha)|\boldsymbol{x}_\alpha; \hat{W}, \hat{\Sigma}) + \log(n)q(\mathrm{tr}\{P_\Phi\} + 1)\,. \tag{4.6}$$

情報量規準，ベイズ型情報量規準などモデル選択規準の研究については，Konishi (1999)，小西(2000)，Konishi et al. (2004)の優れた文献があるので参照されたい．

5. 数値例

本節では，提案する手法を 4.1 節: 人工データ(Wave Form Data)，4.2 節: 音声認識，4.3 節: 文字認識に適用し，その解析を通して提案する手法の有効性を検証する．

5.1 Wave Form data

本節では Wave Form Data (Hastie et al. (1994, 2001)) の解析を通して，3.3 節で述べたモデル評価規準と予測誤差の関係，及び提案する方法の有効性・汎化能力について検討する．Wave Form Data の特徴変数は 21 次元 $\boldsymbol{x} = (x_1, \ldots, x_{21})'$ で，3 群からなり，次のモデルに従い発生する．

$$x_k = \begin{cases} uH_1(k) + (1-u)H_2(k) + \varepsilon_k \text{ if } g = 1 \\ uH_1(k) + (1-u)H_3(k) + \varepsilon_k \text{ if } g = 2 \\ uH_2(k) + (1-u)H_3(k) + \varepsilon_k \text{ if } g = 3 \end{cases} \quad k = 1, \ldots, 21.$$

ここで，u は $[0,1]$ の一様乱数，誤差項 ε_k は標準正規分布に従い発生する．また，関数 $H_1(k), H_2(k), H_3(k)$ はそれぞれ $H(k) = \max\{6 - |k-11|, 0\}$, $H_2(k) = H_1(k-4)$, $H_3(k) = H_1(k+4)$ で定義される．

解析においては，それぞれの群の事前確率を同確率として，学習データ・予測データを計 300 個，500 個それぞれ発生させた．例えば，学習データについてはそれぞれの群から約 100 個ずつのデータを発生させたことになる．

まず，モデル評価規準と予測誤差の関係について議論する．そもそも，判別分析における最終的な目標の一つは，将来観測されるデータを正しく判別する(予測誤差を小さくする)ことである．交差検証法や，ブートストラップ法(Efron and Tibshirani (1993))は予測の観点から提案されており，直感的に分かりやすい理由からよく用いられる方法である．その本質は，学習データの誤判別率のバイアス補正をおこなって，直接，予測誤差を推定する点にある．

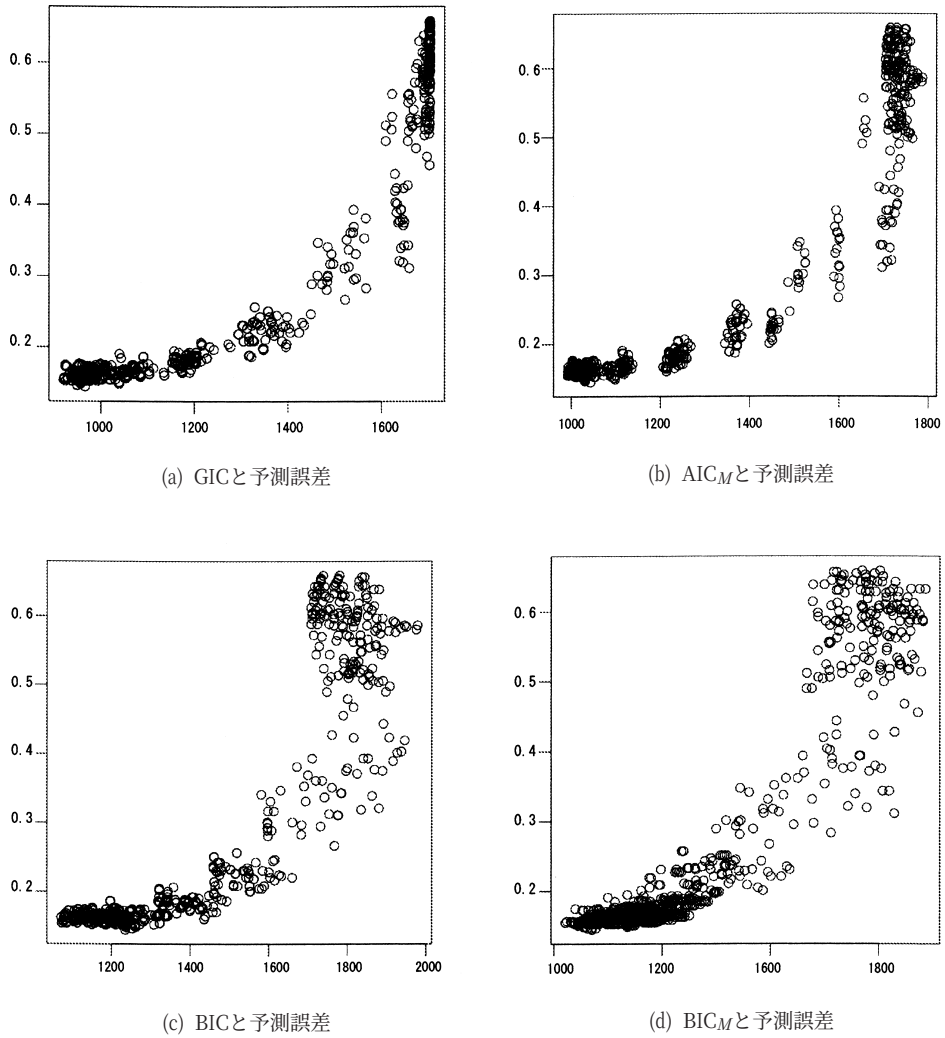
3.3 節で述べたモデル評価規準は，期待対数尤度やベイズの事後確率を最大にしようと意図したもので，直接的には予測誤差を推定しないが，ある意味では予測誤差に対する指標である．つまり，“モデル評価規準の値が小さい $\approx$ 予測誤差が小さい”という関係が成り立っていれば，評価規準の値を小さくするモデルを予測精度が高いモデルとして自動的に選択する根拠が与えられる．そこで，モデル選択規準の値と予測誤差の関係を図 4 に示した．図から見てとれるように，評価規準の値が小さければ，予測誤差も小さいという傾向があった．

次に，今までに提案されている様々な統計的判別手法との比較をおこなった．表 1 は，学習誤差と予測誤差の 10 回の試行の平均値であり，KFDA と GIC (4.2)式の組み合わせが最も小さい予測誤差を与えている．また，他の規準を用いたときも，他の判別手法と比較して，KFDA の予測能力が目立っている．このことから，KFDA の汎化能力，つまり将来のデータの判別能力が優れていることが結論づけられる．

5.2 音声認識

本節では音声認識データ(Hastie et al. (1994))の分析を通して，提案する手法を様々な手法と比較し，その実用性を検証する．音声認識データは，11 個の英語の母音(表 2)に関する特徴を表す 10 次元の変数から構成される 11 群判別データで，15 人の被験者(男性 8 人，女性 7 人)が表 2 にある 11 単語を発音して得られた 990 個のデータから構成されている．解析に際しては，Hastie らと同様に，8 人(男女各 4 人)から得られた 528 個のデータをモデルの推定に用い，残り(男性 4 人，女性 3 人)の 462 個のデータを予測データとして用いた．

モデル選択に際しては，本稿で提案した GIC (4.2)式，BIC (4.5)式，または取り上げた AIC_M (4.3)式，BIC_M (4.6)式の 4 規準を用い，この規準の比較もおこなった．表 3 は，Hastie et al. (1994)らが提案したスプライン加法モデルに基づく FDA など，これまでに提唱された様々な判別手法による結果と，KFDA による予測結果との比較をまとめたものである．表から，KFDA と BIC の組み合わせが最も小さい予測誤差 40%を与えており，線形判別法などの他の手法と比較しても予測誤差の意味で高い汎化能力を示しており，将来の観測データに対する判別能力



(a) GICと予測誤差　(b) AIC_Mと予測誤差

(c) BICと予測誤差　(d) BIC_Mと予測誤差

図 4.　各モデル評価規準の値(横軸)と予測誤差(縦軸)の関係.

が優れていることがわかる.

実際の音声認識問題に適用するときには音声識別率の向上が望まれるが，学習データ数が少ないことに加え，ここでの学習・予測データはそれぞれ違う人の発音が用いられているので，学習データの人数と数を増やすことでモデルの識別率をさらに改善できると考えられる.

5.3　文字認識

本節では，文字認識データ(ZIP Code Data (Le Cun et al. (1990)))の解析を通じて，KFDA の有効性を検証する．識別対象は数字の 0 ～ 9 であり，図 5 はデータベースの一部を図示したものである．それぞれの画像サイズは 16×16 ピクセルのグレイスケールデータで，256 $(= 16 \times 16)$ 次元特徴変数を用いた 10 群識別問題を考えることになる．解析に際しては，1000 個の学習

表 1. 学習・予測誤差(%)の比較(10 回の平均値). ただし, KFDA 以外の結果は Hastie et al. (2001)から引用した.

判別手法	学習誤差	予測誤差
Linear discriminantation	12.1	19.1
Quadratic discriminantation	3.9	20.5
Classification tree	7.2	28.9
Flexible Discriminant Analysis (MARS (degree1))	10.0	19.1
Flexible Discriminant Analysis (MARS (degree2))	6.8	21.5
Kernel Flexible discriminant analysis with GIC	**9.6**	**15.3**
Kernel Flexible discriminant analysis with AIC_M	**11.0**	**15.4**
Kernel Flexible discriminant analysis with BIC	**9.3**	**15.5**
Kernel Flexible discriminant analysis with BIC_M	**12.0**	**16.1**

表 2. 発音する単語から得られる母音.

母音	単語	母音	単語
i	heed	I	hid
E	head	A	had
a:	hard	Y	hud
O	hod	Ↄ:	hoard
U	hood	u:	who'd
ɜ:	heard		

表 3. 様々な判別手法との比較(%). ただし, KFDA 以外の結果は Hastie et al. (2001)から引用した.

判別手法	学習誤差	予測誤差
Linear discrimination	32	56
Quadratic discrimination	1	53
Classification tree	5	56
Flexible discriminant analysis (BRUTO)	6	44
Flexible discriminant analysis (MARS (degree1))	9	45
Flexible discriminant analysis (MARS (degree2))	–	42
Single layer perceptron	–	67
Multi layer perceptron (88 hidden units)	–	49
Gaussian node network (528 hidden units)	–	45
Nearest-neighbor	–	44
Kernel Flexible discriminant analysis with GIC	**6**	**42**
Kernel Flexible discriminant analysis with AIC_M	**6**	**42**
Kernel Flexible discriminant analysis with BIC	**6**	**40**
Kernel Flexible discriminant analysis with BIC_M	**7**	**44**

データを用いてモデルのパラメータを推定し, 1500 個の予測データで予測誤差を推定する. 線形判別をおこなった結果, 学習誤差, 予測誤差はそれぞれ 1.8%, 10.4% となり, 高次元データの解析ゆえに, 2 次判別においては, 分散共分散行列を対角行列と制限(Dudoit et al. (2002))しても, 分散共分散行列が退化してしまい実行できなかった. しかし, KFDA は推定の安定性

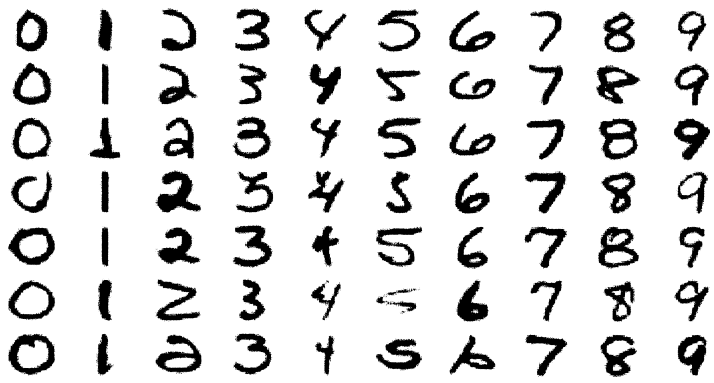


図 5. サンプルデータ.

表 4. 様々な判別手法との比較(%). ただし，NA はモデル推定ができなかったことを示す.

判別手法	学習誤差	予測誤差
Linear discrimination	1.8	10.4
Quadratic discrimination	NA	NA
Multi layer perceptron ($m = 5, \log_{10}(\lambda) = -5$)	1.1	11.8
Multi layer perceptron ($m = 5, \log_{10}(\lambda) = -8$)	1.0	11.2
Multi layer perceptron ($m = 10, \log_{10}(\lambda) = -5$)	1.0	5.8
Multi layer perceptron ($m = 10, \log_{10}(\lambda) = -8$)	1.0	5.6
Multi layer perceptron ($m = 15, \log_{10}(\lambda) = -5$)	0.2	4.8
Multi layer perceptron ($m = 15, \log_{10}(\lambda) = -8$)	0.2	4.7
Kernel Flexible discriminant analysis with GIC	**1.6**	**2.4**
Kernel Flexible discriminant analysis with AIC$_M$	**2.0**	**4.0**
Kernel Flexible discriminant analysis with BIC	**1.9**	**3.1**
Kernel Flexible discriminant analysis with BIC$_M$	**2.0**	**4.1**

を考慮しているので，2 次判別関数よりも柔軟な構造を備えているにもかかわらず，非線形判別関数の構築が可能となり，KFDA の頑健性が有効に機能していることが認識される.

KFDA のモデリングにおいて本質的な非線形変換の次元 s，平滑化パラメータ λ，形状パラメータ $\boldsymbol{\sigma}$ の選択には，GIC，BIC, AIC$_M$, BIC$_M$ の 4 規準を利用した．表 4 は様々な手法との比較である．ただし，Multilayer perceptron (Ripley (1996))はソフトマックス変換と正則化法(正則化項には Weight decay を用いた)を組み合わせ，正則化対数尤度の最大化により推定されている．ここでは，中間層の素子数 m と正則化パラメータ λ の組み合わせ (m, λ) を 6 組用意して解析をおこなった．この表から，本稿で提案する判別方法は極めて有効な手法であることがわかる.

6. おわりに

本稿では Fisher の線形判別手法の非線形への拡張問題について検討し，新しい非線形多群判別手法 “Kernel Flexible Discriminant Analysis” を提案し，その数理的構造をカーネルトリックを利用して明確にした．さらに，尤度概念と巧みに結びつけ，モデルの評価規準を情報量，及

びベイズ理論の観点から提案した．

Hastie et al. (1994)が提案した，スプライン加法モデルに基づく FDA は，Fisher の線形判別手法の非線形化へのブレイクスルーであったが，十分な数のデータが観測・測定できない場合や，極めて高次元のデータの分析を必要とする場合，たとえ十分大きい平滑化パラメータを用いたとしても，モデルの自由度が標本数を上回り，統計的問題が生じるとともに，計算機の上でも困難に遭遇する場合があった．Hastie et al. (1994)は，計算量を考慮した GCV をモデル選択に利用しているが，提案されている GCV には未知の量が含まれ，GCV 規準そのものを最初に最適化する必要があった．

そこで，カーネル関数を利用した KFDA を提案し，尤度概念を与えることで，統計的モデル選択問題を情報量・及びベイズ理論の観点から考察した．これによって，モデルの推定と評価を融合した非線形モデリング手法を構成することができた．KFDA は，モデルの汎化能力を表す予測誤差が，Hastie et al. (1994)が提案した FDA を含む従来の様々な判別手法と比較して小さく，将来のデータの判別能力が極めて優れていることが結論づけられた．

今後，アレイデータの解析(Dudoit et al. (2002))，リモートセンシングデータの分析，債券格付けや倒産などの信用リスク計測などに応用し，その有用性を検証することを研究課題とする．

謝　辞

査読者の方々には，大変貴重なご指摘をいただきました．ここに記して厚く御礼申し上げます．

参 考 文 献

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automat. Control*, **AC-19**, 716–723.

Andersen, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed., Wiley, New York.

安道知寛，井元清哉，小西貞則(2001). 動径基底関数ネットワークに基づく非線形回帰モデルとその推定，応用統計学，**30** (1), 19–35.

安道知寛，島内順一郎，小西貞則(2002). 動径基底関数ネットワークモデルに基づく非線形判別とその応用，応用統計学，**31** (2), 123–139.

Ando, T., Konishi, S. and Imoto, S. (2002). Nonlinear regression modeling via regularized radial basis function networks, Research Memo., No. 845, The Institute of Statistical Mathematics, Tokyo.

Buja, A., Hastie, T. and Tibshirani, R. J. (1989). Linear smoothers and additive models (with discussion), *Ann. Statist.*, **17**, 453–555.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *J. Amer. Statist. Assoc.*, **97**, 77–87.

Efron, B. and Tibshirani, R. (1993). *An Introduction to Bootstrap*, Chapman & Hall, London.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall, London.

Hastie, T., Tibshirani, R. and Buja, A. (1994). Flexible discriminant analysis by optimal scoring, *J. Amer. Statist. Assoc.*, **89**, 1255–1270.

Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis, *Ann. Statist.*, **23**, 73–102.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer, New York.

石井健一郎，上田修功，前田栄作，村瀬洋(1998)．『わかりやすいパターン認識』，オーム社，東京．

Konishi, S. (1999). Statistical model evaluation and information criteria, *Multivariate Analysis, Design of Experiments, and Survey Sampling* (ed. S. Ghosh), 369–399, Marcel Dekker, New York.

小西貞則(2000)．統計的モデリングと情報量規準構成の理論──汎関数に基づくアプローチ──，数学，**52**, 128–141.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection, *Biometrika*, **83**, 875–890.

Konishi, S., Ando, T. and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks, *Biometrika*, **91** (in press).

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *Ann. Math. Statist.*, **22**, 79–86.

Le Cun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R., Hubbard, W. and Jackel, L. (1990). Handwritten digit recognition with a back-propagation network, *Advances in Neural Information Processing System*, Vol. 2 (ed. D. Touretzky), Morgan Kaufman, Denver, Colorado.

Mardia, K., Kent, J. and Bibby, J. (1979). *Multivariate Analysis*, Academic Press, New York.

Ripley, B. D. (1994). Neural networks and related methods for classification, *J. Roy. Statist. Soc. Ser. B.*, **56**, 409–456.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press, New York.

Roth, V. and Steinhage, V. (1999). Nonlinear discriminant analysis using kernel functions, Tech. Report, IAI-TR-99-7, University Bonn.

Schölkopf, B., Smola, A. and Muller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, **10**, 1299–1319.

Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Statist.*, **6**, 461–464.

Stone, C. J. (1974). Cross-validatory choise and assessment of statistical predictions (with discussion), *J. Roy. Statist. Soc. Ser. B*, **36**, 111–147.

Vapnik, V. N. (1998). *Statistical Learning Theory*, Wiley, New York.

Wahba, G. (1990). *Spline Functions for Observational Data*, CBMS-NSF Regional Conference Series, SIAM, Philadelphia, Pennsylvania.

柳井晴夫，高根芳雄(1985)．『新版多変量解析法』，朝倉書店，東京．

 Proceedings of the Institute of Statistical Mathematics Vol. 51, No. 2, 389–406 (2003)

Kernel Flexible Discriminant Analysis for Classifying High-dimensional Data with Nonlinear Structure and Its Applications

Tomohiro Ando

(Graduate School of Mathematics, Kyushu University)

Discriminant analysis aims at the classification of an object into one of given classes based on information from a set of characteristics. Among the many available methods, Fisher linear discriminant analysis, the most popular approach, has so far contributed to development of science and a social system. With the advent of powerful computers and the information age, however, the issue of discriminant analysis has exploded both in sample size and data complexity. Researchers have since begun to tackle nonlinear discriminant analysis problems in a more realistic fashion.

It is well known that Fisher's linear discriminant analysis is equivalent to multi-response linear regression using optimal scoring. We propose nonlinear versions of Fisher's discriminant analysis, "Kernel Flexible Discriminant Analysis (KFDA)", by replacing the linear regression function with a nonlinear kernel function.

Observing that the least square approach tends to yield poor results, we use a smoothing approach in consideration of the predictive performance of the discriminant function. To determine the "best model" among the candidates, we investigate the likelihood of KFDA models and propose model selection criteria from information-theoretic and Bayesian points of view. Real data analysis and Monte Carlo experiments indicate that the proposed KFDA approach performs well in practical situations.

Key words: Fisher's linear discriminant analysis, kernel method, optimal scaling, smoothing, model selection.