判別分析における予測誤差推定法のロバストネス

小 西 貞 則

0. はじめに

Fisherの線形判別関数は，二群の平均値間の差を最も良く識別できるように構成された線形関数である．この線形判別関数を用いたときの予測される誤差をどのように評価するかが問題となる．基本的には二つのアプローチの仕方があり，一つは主として正規性の仮定のもとで数式的に求める方法，もう一つはCross-Validationなどに代表されるノンパラメトリックな方法である．現実のデータに対して，正規性の仮定を設定することが難しいことから，予測誤差の各種推定法の非正規モデルのもとでの様相を調べる必要がある．ここでは，正規性の仮定のもとで導かれた推定法の頑健性およびノンパラメトリックな手法の種々のモデルのもとでの有効性を検討することを目的とする．

1. モデル

母集団確率分布モデルとして，正規性からの様々なタイプの逸脱を表現することのできる混合多変量正規分布を仮定する．パラメータ既知のこのモデルのもとで，予測誤差を数式的に評価し，これを理論値として推定法比較の基準とした．さらに，混合多変量正規モデルのもとでMardia (1970, *Biometrika*) による多変量歪度・尖度を合わせて導き，正規分布からの剝離の度合を見るための目安とした．

2. 推定法

(1) 期待予測誤差の漸近評価 (NM法)：正規性の仮定のもとで，期待予測誤差を標本数に関して漸近的に評価したもので，式中のパラメータをその推定値で置き換えたもの (Okamoto (1963, *AMS*), McLachlan (1974, *Biometrics*))．(2) 見かけ上の誤判別率の偏り修正 (NA法)：偏りを正規性の仮定のもとで漸近的に評価し修正を加えたもの (McLachlan (1976, *Biometrika*))．(3) Bootstrap法 (BS法)：見かけ上の誤判別率の偏りを観測されたデータからのresamplingによって推定し修正したもの．(4) Cross-Validation (CV法)．

3. 数値比較

正規モデルから著しく逸脱していなければ，NM法は理論値，平均二乗誤差，標準誤差の比較に置いて，最も安定した推定法といえる．しかし正規分布より特にすその重い分布に対しては，過大な推定値を与える傾向がある．NA法，BS法は様々なタイプの非正規モデルのもとで，極めて理論値に近い値を示す．見かけ上の誤判別率の偏りそのものの推定値の標準誤差をみると，NA法のそれはBS法の1/3～1/2であった．また，BS法のresamplingの回数は，この場合約200で十分と思われる．マハラノビス距離の小さい場合のCV法の利用は注意を要する．

オーダー k の二項分布に関する或る推定問題

安 芸 重 雄

成功の確率が p であるような独立試行の系列を考える．k と n を正整数とし，固定しておく．このとき，n 回目の試行までに〝k 回続いた成功〟という事象が起こった回数の分布をパラメータ (p, n) を持つオーダー k の二項分布と言い $B_k(n, p)$ と書く．$k=1$ のときは通常の二項分布であるが，それ以外の場合には確率関数の表現は簡単ではない (Hirano[2] 参照)．

$X_1,\cdots,X_m$ が独立に $B_k(n, p)$ に従うとき，$X_1,\cdots,X_m$ から p を推定する問題を考える．$B_k(n, p)$ の平

均が

$$\sum_{j=1}^{[n/k]}\{(n-jk+1)p^{jk}-(n-jk)p^{jk+1}\}$$

と書ける（[1] 参照）のでモーメント法が簡単に実行できる．或る種の工夫をすれば最尤法も使えるが，このモデルではモーメント法による推定量の漸近効率が非常に良いので実用上はモーメント法を用いれば十分であろう．

さて，sample size が1の場合を考えてみる．この場合も $n\to\infty$ のときには一致推定量を作ることができる．$\xi_1, \xi_2, \cdots$ をオーダー k の幾何分布（$G_k(p)$）に従う独立な確率変数列とする．確率変数 R_n を

$$R_n=\begin{cases}0, & \text{if} \quad \xi_1>n\\ r, & \text{if} \quad \xi_1+\cdots+\xi_r\leqq n \quad \text{and} \quad \xi_1+\cdots+\xi_{r+1}>n\end{cases}$$

によって定義すると，$R_n\sim B_k(n,p)$ であり，$R_n\to\infty$ in prob. as $n\to\infty$ であることはすぐわかる．さて，

(1)　$$n/R_n\to E(\xi_1) \quad \text{in prob. as} \quad n\to\infty$$

を示す．R_n の定義より，

$$S(R_n)/R_n\leqq n/R_n<S(R_n)/R_n+\xi_{R_n+1}/R_n$$

となる．ここで $S(m)=\xi_1+\cdots+\xi_m$．また，$\xi_{R_n+1}/R_n\to 0$ in prob. も簡単にわかるので，結局，

(2)　$$S(R_n)/R_n\to E(\xi_1) \quad \text{in prob.}$$

が証明できれば(1)式が示される．ところが(2)式は Révész [3] の Theorem 10.1 により保証される．

$E(\xi_1)=(1-p^k)/((1-p)p^k)$ であるから $n/X_1=(1-p^k)/((1-p)p^k)$ の解 $\hat{p}$ は p に確率収束する．

参考文献

[1] Aki, S. and Hirano, K. (1986). *Res. Memo.*, No. **316**, The Inst. of Statist. Math.
[2] Hirano, K. (1986). *Fibonacci Numbers and Their Appl.*, Reidel, 43-53.
[3] Révész, P. (1968). *The Laws of Large Numbers*, Academic Press.

線形結合統計量の一様正規近似

松　縄　　規

要　旨

$\{X_i\}$ $(i=1,\cdots,n)$ を pdf. $f_i(x)$, cdf. $F_i(x)$ に従う連続分布からの独立な n 個の確率変数とする．この時 (I) $\{X_i\}$ $(i=1,\cdots,n)$ の線形結合，(II) $\{X_i\}$ に基づく順序統計量の中の k 個の統計量の線形結合，の二つについて，密度関数表現による一様漸近正規性の一般論を与える．

1. 必要補題

$\{X_s\}, \{Y_s\}, \{Z_s\}$ $(s=1, 2,\cdots)$ を可測空間 $(R_{(m)}, \boldsymbol{B}_{(m)})$ 上の確率変数，$R_{(m)}$ を m 次元ユークリッド空間，$\boldsymbol{B}_{(m)}$ をその部分集合からのボレル集合体，$\nu_{(m)}$ をこの空間上のルベック測度とする．

補題 I.

(I-1)　$X_s=Y_s+Z_s$ の関係があるものとする．この時，(I-2) $\{X_s\}, \{Y_s\}$ の一様絶対連続性 $C(\boldsymbol{B})$, (I-3) $\{Y_s\}$ の一様有界性 $B(\boldsymbol{S})$, (I-4) $Z_s\to 0$ in p $(s\to\infty)$ が同時に満されるならば $X_s\sim Y_s$ $(\boldsymbol{B})_d$, $(s\to\infty)$.