

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 $1/3 \sim 1/2$ 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, \dots, X_m が独立に $B_k(n, p)$ に従うとき、 X_1, \dots, X_m から p を推定する問題を考える。 $B_k(n, p)$ の平