A-021

説明変数に対する属性別パラメータを考慮した判別モデル

Discrimination Analysis Considering Attribute Score Parameters of Explanatory Variables

山下 遥 †* 後藤 正幸 †
Haruka Yamashita Masayuki Goto

1 研究背景・目的

判別問題では，説明変数やサンプルにある種の構造を仮定できる場合がある．例えば，野球選手の打撃成績データの場合,「本塁打数」「打点」「盗塁数」などの説明変数は「長打力」や「走力」など名前でグルーピングすることができる．この場合，判別モデルを全変数を用いて構築するよりも,「長打力」や「走力」を用いて判別（e.g., オールスターゲームに出場できるか，できないか）モデルを構築した方が有益な分析ができることが考えられる．本研究では，このような「長打力」や「走力」に当たる変数グループの名称を「指標」と呼び，各指標が複数の変数を持つようなデータ構造に着目する．

さらに上記のようなデータの中には，複数の指標が複数の変数によって構成されると同時に，サンプル属性（e.g., 日本人選手か外国人選手か）が分かっていると同時に，サンプル属性によって異なる判別モデルが得られる場合を考える（e.g., 日本人は外国人に比べて「走力」が重視される）．この場合，まずサンプル属性別の変数ごとに主成分分析を行い，複数の変数をサンプル属性別の指標に合成してから，線形判別分析を行う方法が提案されている [1]．しかしながら，こうしたアプローチには，サンプル属性別の指標それぞれに対して行われる主成分分析が，サンプル属性を用いた判別分析とは独立に行われるという問題点，主成分分析よって求められた係数パラメータがサンプル属性別に異なるため，合成された指標の意味合いがサンプル属性によって異なるという問題点が存在する．

そこで本研究では，サンプル属性別の変数の値を指標の値と合成する際に主成分分析でなく，各指標の線形判別係数を固定したもとでの線形判別分析を行う方法を提案する．さらに，本研究の提案モデルの適用例としてプロ野球の打撃成績データから，日本人，外国人別のオールスターゲームへの出場に関する判別分析を行う．

2 本研究の問題設定

本研究では以下のような問題設定のもとで，線形判別法を基礎とした新たな説明変数に対するサンプル属性別パラメータを考慮した線形判別モデルを提案する．

まず，表 1 のように I 個の指標 $i(i = 1, ..., I)$（e.g., “走力”）が J_i 個の変数 $j(j = 1, ..., J_i)$（e.g., “走力”を構成するための “盗塁数” と “3 塁打数”）によって構成され，かつサンプル属性 k の中の s 番目のサンプルが群 C_1 または C_2 のいずれかに属することが分かっている時，説明変数 $x_{s(k)ij}$ によって，各サンプルが群 C_1 に属するか C_2 に属するのかを判別する問題を考えることにする．

*早稲田大学

表 1: 本研究で想定するデータ構造

		1			…	I			群
k	$s(k)$	1	…	J_1	…	1	…	J_I	
1	1(1)	$x_{1(1)11}$	…	$x_{1(1)1J1}$	…	$x_{1(1)I1}$	…	$x_{1(1)IJI}$	C_1
	⋮	⋮		⋮	⋮	⋮		⋮	⋮
	$S(1)$	$x_{S(1)11}$	…	$x_{S(1)1J1}$	…	$x_{S(1)I1}$	…	$x_{S(1)IJI}$	C_2
2	1(2)	$x_{1(2)11}$	…	$x_{1(2)1J1}$	…	$x_{1(2)I1}$	…	$x_{1(2)IJI}$	C_2
	⋮	⋮		⋮	⋮	⋮		⋮	⋮
	$S(2)$	$x_{S(2)11}$	…	$x_{S(2)1J1}$	…	$x_{S(2)I1}$	…	$x_{S(2)IJI}$	C_1

その際，各指標 i がサンプル属性 k 別にどれだけ群 C_1 または C_2 への判別に寄与しているのかを捉えるためのアプローチとして，まず，サンプル属性 k 別かつ指標 i 別に主成分分析を行い，説明変数 $x_{s(k)ij} \in R$ を指標値 $z_{s(k)i}$ へと合成し，その $z_{s(k)i}$ を説明変数として群 C_1, C_2 の判別分析を行う手順が提案されている [1]．

しかしながら，こうしたアプローチにおいて，以下の 2 つの問題点を指摘することができる．

(i) サンプル属性 k 別の指標 i それぞれに対して行われる主成分分析は群への判別とは独立に行われる．よって，説明変数 $x_{s(k)ij}$ を指標値 $z_{s(k)i}$ へと合成する際に各群への寄与が考慮されていない．

(ii) 主成分分析よって求められた係数パラメータがサンプル属性別に異なるため，合成された指標値 $z_{s(k)i}$ の意味合いがサンプル属性によって異なる．その結果，推定されたサンプル属性の判別係数を，サンプル属性の間で単純には比較することができなくなってしまう．

3 説明変数に対する属性別パラメータを考慮した判別モデル

3.1 属性別パラメータを考慮した判別関数の提案

本研究では，前節で示した問題設定において，複数の説明変数 $x_{s(k)ij}$ を合成して指標値 $z_{s(k)i}$ を構成する際にまず，各指標および各サンプル属性の判別係数 a_{ki} を固定したもとで変数 j に関する線形判別分析を行う (i.e., 判別係数 b_{ij} を求める)．さらに得られた各サンプルの判別得点を合成された指標値 $z_{s(k)i}$ と表し，以下のような線形判別関数 $F_{s(k)}$ を提案する．

$$F_{s(k)} = \sum_{i=1}^{I} a_{ki} \sum_{j=1}^{J_i} b_{ij} x_{s(k)ij} \quad (1)$$

サンプル $s(k)$ の判別値 $F_{s(k)}$ の値が 0 より大きい場合は群 C_1 へ，0 より小さい場合は群 C_2 へと判別する．ここに，$z_{s(k)i} = \sum_{j=1}^{J_i} b_{ij} x_{s(k)ij}$ は複数の説明変数 $x_{s(k)ij}$ の指標値 $z_{s(k)i}$ への合成に相当し，判別係数 a_{ki} が指標値に対する判別係数を表す．また，$x_{s(k)ij}$ を要素とするベクトル $\boldsymbol{x}_{s(k)} = (x_{s(k)ij})$ に正規性を仮定する．

このモデルは，複数の変数 j を合成して $z_{s(k)i}$ を構成する際に，主成分分析でなく (1) 式の構造の下でパラメータ a_{ki}，b_{ij} を最適化することで判別を考慮した指数値を算出することができる．さらに，b_{ij} を，サンプル属性別でなく (i.e., b_{kij}でなく)，これらに関して共通のパラメータとすることにより，推定された (1) 式で表される判別係数の a_{ki} を，サンプル属性 k の間で比較することができる．

3.2 判別関数のパラメータ推定

(1) 式のパラメータ a_{ki}, b_{ij} は解析的に数式解を得ることができない．そこで，a_{ki},b_{ij} をそれぞれ交互に推定し，収束させるアルゴリズムを提案する．提案するアルゴリズムは以下のとおりである．

STEP0 a_{ki} を全て 1 に固定する．

STEP1 a_{ki} を定数とみなし, (1) 式の $a_{ki}x_{(k)ij}$ を説明変数とする．このとき群 C_1 および群 C_2 における母平均の推定値をそれぞれ $\bar{\boldsymbol{x}}^{[1]}$,$\bar{\boldsymbol{x}}^{[2]}$，2 つの群全体での母分散共分散行列の推定値を $\boldsymbol{S}$ とすると，ベクトル $\boldsymbol{b} = (b_{ij})$ は，以下のように推定される．

$$\boldsymbol{b} = \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}^{[1]} - \bar{\boldsymbol{x}}^{[2]}) \quad (2)$$

さらに，$\boldsymbol{b}$ の要素の二乗和が 1 になるように基準化する．

STEP2 STEP1 で推定した b_{ij} を定数とみなし，$b_{ij}x_{(k)ij}$ を説明変数とする．このときの，群 C_1 および群 C_2 における母平均の推定値をそれぞれ $\bar{\boldsymbol{x}}'^{[1]}$, $\bar{\boldsymbol{x}}'^{[2]}$，2 つの群全体でのサンプルの母分散共分散行列を $\boldsymbol{S}'$ とすると，ベクトル $\boldsymbol{a} = (a_{ki})$ は，以下のように推定される．

$$\boldsymbol{a} = \boldsymbol{S}'^{-1}(\bar{\boldsymbol{x}}'^{[1]} - \bar{\boldsymbol{x}}'^{[2]}) \quad (3)$$

STEP3 $\boldsymbol{a}$ および $\boldsymbol{b}$ の値が変化しなくなるまで STEP1, 2 を繰り返す．

4 提案モデルを用いたプロ野球データの解析

本研究の提案モデルの適用例としてプロ野球データを分析する．2015 年度で規定打席に達していた 58 名のプロ野球選手の本塁打数，長打率，打点を変数として長打力の指標，打率，安打数，出塁率を変数として打撃確実性の指標，そして盗塁数，三塁打数を変数として走力の指標を構成し，日本人選手と外国人選手をサンプル属性と位置づけ，2015 年度のオールスターゲームに出場したか否かに対する長打力，打撃確実性，そして走力の 3 つの指標を用いた判別モデルを構築するために，(i)Zhao らの従来法 [1] ,(ii) 提案モデルによりデータを解析した．

判別精度を表 2 に, 従来法によって得られたモデルのパラメータを表 3 に，そして提案法によって得られたモデルのパラメータを表 4 示す．

表 2: 判別率の比較

従来法	提案法
63.50%	76.90%

表 3: 従来法によって得られたパラメータ

日本人選手							
長打力			打撃確実性			走力	
0.36			0.24			0.23	
本塁打数	長打率	打点	打率	安打数	出塁率	盗塁数	三塁打数
0.58	0.58	0.57	0.61	0.59	0.54	0.67	0.74
外国人選手							
長打力			打撃確実性			走力	
0.36			0.24			0.54	
本塁打数	長打率	打点	打率	安打数	出塁率	盗塁数	三塁打数
0.67	0.49	0.55	0.53	0.26	0.80	0.32	0.95

表 4: 提案法によって得られたパラメータ

日本人選手							
長打力			打撃確実性			走力	
0.30			0.66			0.14	
外国人選手							
長打力			打撃確実性			走力	
0.43			0.20			0.49	
本塁打数	長打率	打点	打率	安打数	出塁率	盗塁数	三塁打数
−0.54	0.53	0.30	−0.31	0.40	−0.05	0.30	−0.07

以上の結果より，提案法によって求めた判別関数の方が当てはまりが良いことがわかる．また，従来法では，変数を合成する際の重みがサンプル属性（日本人か外国人か）によって異なっているので，得られた指標に対する判別係数の値の比較が難しいが，提案法では重みがサンプル属性間において共通となっているため，得られた指標に対する解釈が容易となっている．

例えば，提案モデルによって得られたパラメータに着目すると，日本人選手と外国人選手に望まれる長打力，打撃確実性，走力の違いが明らかになる．特に，打撃確実性では，その違いが大きく出ており，提案法の有効性を示唆することができる．

5 結論

本研究では，サンプル属性別の指標をへと合成する際に各指標の線形判別係数を固定したもとでの判別分析を行う新たな判別モデルを提案した．さらに，本研究の提案モデルの適用例としてプロ野球データを分析した．

参考文献

[1] Wenyi Zhao, Arvindh Krishnaswamy, Rama Chellappa, Daniel L. Swets, John Weng,“Discriminant Analysis of Principal Components for Face Recognition,” *Face Recognition*, Vol.163, pp.73–85, 1998.