Algumas questões em aberto na análise discriminante para três grupos

Carla M. Santos*
Ana M. Pires**
**Departamento de Matemática, Univ. Portucalense Infante D. Henrique*
***Departamento de Matemática e Centro de Matemática Aplicada, Instituto Superior Técnico*

Sumário: No desenvolvimento da maior parte dos métodos de estimação de regras discriminantes (com excepção dos critérios genéricos de optimalidade) são, em geral, considerados apenas dois grupos, pressupondo-se que a generalização a mais de dois grupos é simples e directa. Se isso é verdade para, por exemplo, a regra discriminante linear de Fisher, já não o é para muitos outros tipos de metodologias, em que se incluem alguns métodos não paramétricos e a maior parte dos métodos robustos.

Neste trabalho ilustram-se algumas dificuldades que podem surgir nas aplicações a três grupos e apontam-se possíveis soluções. Pelo facto de ser mais fácil a visualização das soluções adoptadas bem como das respectivas consequências, considera-se apenas o caso bivariado.

Palavras-chave: Análise Discriminante; Projection Pursuit.

Abstract: In the development of most of the discriminant rules estimation methods (with the exception of the general optimal criteria) only the two group case is studied, assuming beforehand that the generalisation to more than two groups is straightforward. Although that is so for Fisher´s linear discriminant function, it is not for other kinds of methodologies, including some nonparametric methods and most of the robust methods.

This work illustrates some difficulties that can emerge in the applications to three groups and indicates possible solutions. Only the bivariate situation is considered because the adopted solutions and their consequences are easier to visualise.

Keywords: Discriminant analysis; Projection Pursuit.

1 Introdução

Como é do conhecimento geral, o critério de Fisher, que permite a separação em dois grupos, é ainda hoje um dos mais aplicados pela simplicidade dos resultados a que conduz, associado à linearidade imposta e ao facto de não assumir à priori nenhuma forma distribucional (ver, e.g., Hand, 1997, Johnson e Wichern, 1992 ou McLachlan, 1992). Quando se procede à sua generalização a três grupos, aplicando-o a dois grupos de cada vez, as linhas de separação são três rectas que se intersectam no mesmo ponto (ver Figura 1).

No entanto, tal como é referido também em Johnson e Wichern (1992), embora a "robustez" associada à classificação linear para dois grupos possa ser avaliada de uma forma relativamente simples, quando este número aumenta esta abordagem não leva a conclusões gerais, já que as propriedades dependem da configuração da distribuição em estudo. Nomeadamente, quando há desvios significativos da normalidade

surge a necessidade de criar metodologias que abranjam a utilização de estimadores robustos.

Note-se que o critério de Fisher, apesar de não assumir uma forma distribucional para as observações, só garante a obtenção de regras com propriedades óptimas no caso de populações gaussianas com matrizes de covariâncias idênticas, pelo que fazem todo o sentido considerações de robustez.

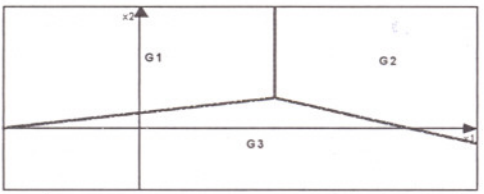


Figura 1: Rectas de separação para o critério de Fisher

Encontrando-se na literatura diversos métodos robustos para dois grupos, opta-se aqui por considerar o chamado método PP (de "Projection-Pursuit") proposto por Pires (1995) na sua variante que se baseia nos estimadores-M de Huber, o qual mostrou possuir boas propriedades. No entanto, a sua generalização a mais de dois grupos, por aplicação a dois de cada vez, pode dar origem à situação ilustrada na Figura 2. As linhas de separação, embora continuem a ser lineares, em geral não se intersectam, i.e., não conduzem a uma partição. Como consequência surgem pontos sem possibilidade de serem classificados.

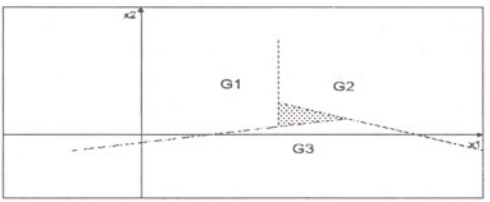


Figura 2: Rectas de separação para o método PP

2 Soluções propostas

→ Fixar duas das rectas discriminantes e fazer passar a terceira pelo ponto de intersecção, deslocando-as paralelamente ⇒ três possibilidades (PP1/PP2/PP3).

→ Fazer passar as três rectas pelo baricentro do triângulo correspondente à região de indefinição, deslocando-as paralelamente (PP4).

2.1 Avaliação da regra discriminante

Uma vez que o objectivo é classificar, a melhor solução deverá ser avaliada através da taxa de erro actual (ou seja da probabilidade de erro na classificação de

novas observações). A estimativa mais directa daquela taxa é chamada taxa de erro aparente, calculada pela reclassificação da amostra depois de obtida a regra discriminante e procedendo-se à contagem dos indivíduos mal classificados no i-ésimo grupo (m_i) relativamente ao número total de indivíduos em cada grupo (n_i).

$$e_{ap} = \sum_{i=1}^{3} m_i \Big/ \sum_{i=1}^{3} n_i \ .$$

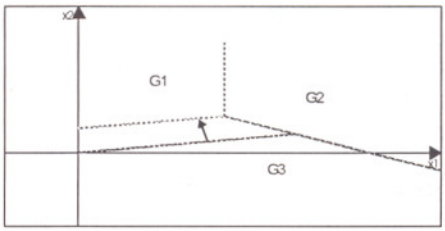


Figura 3: Solução PP1

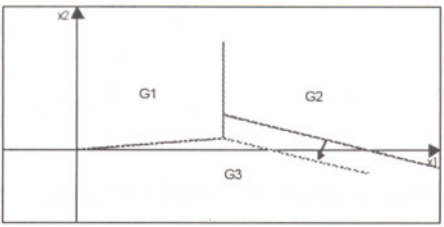


Figura 4: Solução PP2

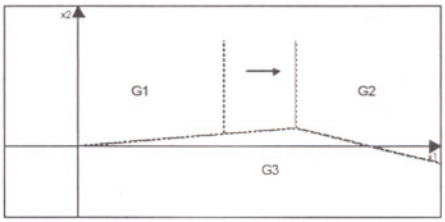


Figura 5: Solução PP3

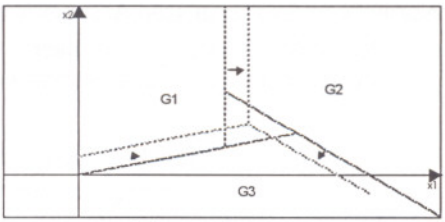


Figura 6: Solução PP4

No entanto, os resultados destas estimativas são geralmente enviesados no sentido optimista (têm tendência a ser menores que o verdadeiro valor) já que são classificados os mesmos indivíduos que contribuíram para a construção da regra. Um processo para eliminar o enviesamento consiste em utilizar uma fórmula semelhante mas relativamente a um conjunto de observações, chamado conjunto teste, independente do conjunto com base no qual foi construída a regra. Designa-se esta estimativa por ($\hat{e}_{act}$) (veja-se Pires, 1995).

3 Aplicação

Descreve-se em seguida o conjunto das situações onde foram testadas as soluções, bem como as taxas de erro para estas quatro soluções e para o critério de Fisher. Para que se pudessem comparar os resultados e no sentido de obter alguma indicação útil para trabalhos futuros, nomeadamente em situações reais onde é habitual a existência de observações discordantes em relação a um dado modelo ou à "maioria" (*outliers*), procedeu-se a uma aplicação a três conjuntos de dados simulados.

Em qualquer uma das três situações tomaram-se amostras de 50 observações de cada grupo: Grupo 1 (G1), Grupo 2 (G2) e Grupo 3 (G3).

→ Normal (**N**)

$$G_1 \sim \mathrm{N}\left(\mu_1 = \begin{pmatrix} -1 \\ 3 \end{pmatrix}, \Sigma\right)$$

$$G_2 \sim \mathrm{N}\left(\mu_2 = \begin{pmatrix} 5 \\ 3 \end{pmatrix}, \Sigma\right) \qquad \textit{onde} \qquad \Sigma = \begin{pmatrix} 1{,}3^2 & 0 \\ 0 & 1{,}3^2 \end{pmatrix}$$

$$G_3 \sim \mathrm{N}\left(\mu_3 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \Sigma\right)$$

→ Normal Simetricamente Contaminada (**NSC**)

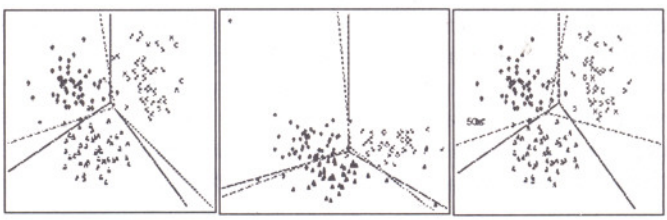
$$G_i \sim 0{,}9\,\mathrm{N}_2(\mu_i, \Sigma) + 0{,}1\,\mathrm{N}_2(\mu_i, 9\Sigma),\ i = 1,2,3$$

→ Normal com Contaminação Assimétrica num dos grupos (**NCA**)

Situação idêntica ao primeiro conjunto de dados mas em que 5 observações de G_2 foram substituídas pelo ponto (-4,1).

Tabela 1: Taxa de erro aparente (em %)

	N	NSC	NCA
Fisher (População)	3,3333	6,6667	6,6667
Fisher (Amostra)	3,3333	8,0000	8,6667



Como nas situações tratadas é conhecida a população de onde são provenientes as observações, foi gerado um conjunto de dados independente de:

- 3000 observações em cada grupo para a primeira e terceira situação.
- 6000 observações em cada grupo para a segunda situação.

As dimensões das amostras teste são consequência de se ter exigido uma precisão de 0.005 com um grau de confiança de 99%, baseado nas respectivas estimativas resultantes da taxa de erro aparente.

Nas Tabelas 1 e 2 apresentam-se as taxas de erro aparente e actual (método do conjunto teste), em percentagem.

Tabela 2: Estimativas da taxa de erro actual - método do conjunto teste (em %)

	N	NSC	NCA
Fisher (População)	4,2556	7,5444	4,2556
Fisher (Amostra)	4,8778	7,8222	5,8222
PP1	5,4778	8,0556	5,4667
PP2	5,2333	7,9222	5,0444
PP3	5,4000	8,0833	5,1000
PP4	5,3333	7,8944	5,1111

Para que se possa ter uma ideia mais clara das estimativas da taxa de erro actual, em termos do aumento do número de observações mal classificadas para algumas das situações implementadas, quando confrontadas com o critério de Fisher, observe-se a Tabela 3.

Tabela 3: Estimativas do aumento do número de observações mal classificadas relativamente ao critério de Fisher, com os parâmetros da população (em %)

	N	NSC	NCA
Fisher (Amostra)	14,6214	3,6819	36,8146
Melhor solução PP	22,9765	4,6392	18,5379
Pior solução PP	28,7206	7,1429	28,4595

Em relação à terceira situação (NCA) e para a amostra gerada para a estimação da taxa de erro, julga-se que a Tabela 4 permite evidenciar a robustez do método PP, em termos das diferenças na ordem de grandeza das observações mal classificadas, relativamente ao critério de Fisher; repare-se, e.g., no número de observações do grupo 2 que foram mal classificadas na região 3.

Tabela 4: Número de observações mal classificadas (NCA)

	T			A			PP			PP1			PP2			PP3			PP4		
G \ R	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1		27	116		78	174		44	202		44	287		44	202		31	202		36	225
2	30		44	18		119	18		32	18		35	18		26	37		32	26		29
3	123	43		82	53		109	46		62	46		109	55		114	43		94	50	
Total	383			524			461			492			454			488			480		

↓

\+ 5 Observações na região de indefinição.

4 Conclusões

A vantagem da utilização de um estimador robusto ficou patente em particular na situação mais contaminada (**NCA**).

Não se conseguiu estabelecer a superioridade de nenhuma das soluções PP, o que não surpreende, devido ao "pequeno" tamanho das regiões de indefinição. No entanto, esta questão não é de importância fundamental pois a escolha da melhor solução pode ser sempre feita caso a caso, desde que se utilize um bom estimador das taxas de erro.

Pensa-se que a realização de um estudo de simulação com mais amostras, bem como a aplicação a dados reais, permitirá ter uma ideia mais precisa das vantagens dos métodos robustos relativamente aos métodos clássicos. Por outro lado terá também interesse explorar situações com mais variáveis e mais grupos.

Referências

[1] Hand, D.J. (1997). *Construction and Assessment of Classification Rules*. John Wiley & Sons, New York.

[2] Johnson, R.A. e Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*. Prentice-Hall.

[3] McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, New York.

[4] Pires, A.M. (1995). *Análise Discriminante. Novos Métodos Robustos de Estimação*. Tese de Doutoramento, Instituto Superior Técnico, U.T.L., Lisboa.