Analyse discriminante sur données binaires lorsque les populations d'apprentissage et de test sont différentes

Julien Jacques* et Christophe Biernacki**

*Laboratoire de Statistiques et Analyse des Données,
Université Pierre Mendès France,
38040 Grenoble Cedex 9, France.
julien.jacques@iut2.upmf-grenoble.fr,
http://www.julien.jacques2.free.fr

** Laboratoire Paul Painlevé UMR CNRS 8524,
Université Lille I,
59655 Villeneuve d'Ascq Cedex, France.
christophe.biernacki@math.univ-lille1.fr

Résumé. L'analyse discriminante généralisée suppose que l'échantillon d'apprentissage et l'échantillon test, qui contient les individus à classer, sont issus d'une même population. Lorsque ces échantillons proviennent de populations pour lesquelles les paramètres des variables descriptives sont différents, l'analyse discriminante généralisée consiste à adapter la règle de classification issue de la population d'apprentissage à la population test, en estimant un lien entre ces deux populations. Ce papier étend les travaux existant dans un cadre gaussien au cas des variables binaires. Afin de relever le principal défi de ce travail, qui consiste à déterminer un lien entre deux populations binaires, nous supposons que les variables binaires sont issues de la discrétisation de variables gaussiennes latentes. Une méthode d'estimation et des tests sur simulations sont présentés, puis des applications dans des contextes biologique et d'assurance illustrent ce travail

1 Introduction

L'analyse discriminante classique suppose que l'échantillon d'apprentissage et l'échantillon test, qui contient les individus à classer, sont issus d'une même population. Depuis les travaux de Fisher (1936), qui introduit une règle de discrimination linéaire entre deux groupes, de nombreuses évolutions ont été proposées (cf. McLachlan (1992) pour une revue). Toutes ces évolutions concernent la nature de la règle de discrimination : paramètrique, semi-paramètrique ou encore non paramétrique.
Une évolution alternative, introduite par Van Franeker et Ter Brack (1993) puis développée par Biernacki *et al.* (2002), considère le cas où l'échantillon d'apprentissage et l'échantillon test ne sont pas nécessairement issus d'une même population. Biernacki et al. définissent plusieurs