# Relaxations de la régression logistique : modèles pour l'apprentissage sur une sous-population et la prédiction sur une autre

Farid Beninel*, Christophe Biernacki**

*CREST ENSAI
rue Blaise Pascal, Campus de Ker Lann
35170 Bruz, France
Farid.Beninel@ensai.fr

**Université Lille1, UFR de mathématiques, UMR 6524
59655 Villeneuve d'Ascq, France
Christophe.Biernacki@math.univ-lille1.fr

**Résumé.** Habituellement en analyse discriminante on a à prédire le groupe d'appartenance à partir des variables de description ou covariables. La règle de prédiction est élaborée en utilisant un échantillon d'apprentissage soumis aux mêmes conditions externes que les individus à prédire. Dans ce travail, on s'intéresse à la prédiction d'individus d'une certaine sous-population utilisant un échantillon d'apprentissage d'une autre sous-population. En assurance-finance, le problème apparaît quand il faut inférer le groupe d'appartenance de *sociétaires-clients* soumis à certaines conditions externes et que la règle est élaborée à partir d'individus soumis à d'autres. On propose différents modèles étendant la discrimination logistique classique. Ces modèles se fondent sur des relations acceptables entre les fonctions scores que l'on associerait à chacune des sous-populations en présence.

## 1 Introduction

Traditionnellement, l'analyse discriminante procède de la façon suivante (McLachlan 1992) : un échantillon provient d'une population et une partition en plusieurs groupes de cet échantillon est connue. À partir des variables disponibles, une règle de classement est alors établie dans le but de classer tout nouvel élément non étiqueté. Néanmoins, une hypothèse sous jacente à cette procédure est que ces nouveaux individus et les individus constituant l'échantillon d'apprentissage proviennent de la même population. L'analyse discriminante généralisée consiste à étendre le problème de l'analyse discriminante classique lorsque cette hypothèse fondamentale est relaxée.
Dans Biernacki *et al.* (2002) on considère cette extension dans le cas de la discrimination gaussienne multivariée. À partir d'hypothèses simples et raisonnables sur la nature du lien stochastique entre les deux sous-populations d'où proviennent respectivement l'échantillon d'ap-