UPDATING A LOGISTIC DISCRIMINATION RULE
Comparing Some Logistic Submodels in Credit-scoring

Farid Beninel† and Christophe Biernacki‡
†*CREST-ENSAI & UMR 6086, Campus de Ker Lann, rue Blaise Pascal, 35170 Bruz, France*

‡*Université Lille1, UFR de Mathématiques & UMR 6524, 59655 Villeneuve d'Ascq, France*

Keywords: Credit scoring, Discriminant rule, Error rate, Learning sample, Logistic model, Misclassification rate, Generalized discrimination, Updating a discriminant rule, Subpopulations mixture, Supervised classification.

Abstract: Often a discriminant rule to predict individuals from a certain subpopulation is given, but the individuals to predict belong to another subpopulation. Two distinct approaches are usually implemented. The first approach is to apply the same discriminant rule for the two subpopulations. The second approach is to estimate a new rule for the second subpopulation. The first classical approach does not take into account differences between subpopulations. The second approach is not reliable in cases of few available individuals from the second subpopulation. In this paper we develop an intermediate approach: we get a rule to predict in the second population combining the experienced rule of the first population and the available learning sample from the second. Different models combining the first rule and the labeled sample from the second population are estimated and tested.

1 INTRODUCTION

Given a categorical target variable and a set of covariates, we deal with the issue of predictive discrimination in the context of a mixture of two subpopulations. More precisely, we have to construct a rule that assigns individuals from one subpopulation, to one of prespecified set of classes based on a vector of measurements (or covariates) taken on those individuals. The available data consist in small learning sample from the subpopulation to predict and a discriminant rule on the second one.

Such a problem arises in various fields of application. The particular problem which motivates this work concerns the prediction of some particular borrowers behavior in credit-scoring. Here, the concerned particular borrowers are not customer of the bank where the loan is demanded. Hence, in the first subpopulation borrowers are customers and in the second subpopulation borrowers are not customers. The behavior is given by the target variable with *creditworthy* and *not-creditworthy* as the prespecified classes.

This work extends a realized work (Biernacki et al., 2002) related to prediction of gender of birds given their morphometric measures and generalizing the Gaussian predictive discrimination method. In such an application individuals are seabirds from *Calanectris Diomedea* species and the mixture of the two subpopulations results from subspecies *Borealis* and *Diomedea* distinguished by their geographical distribution (Thibault et al., 1997), (Bretagnolle et al., 1998).

As it is well known that the geographical location affects on measures of size (Zink and Remsen, 1986), the use of classical predictive discrimination methods for predicting gender of birds from different locations is unreliable. Here, by classical methods, we mean methods based on pure models (*i.e.*, models more adapted for an *homogeneous population*): Gaussian discriminant analysis, logistic discriminant analysis, neural networks, classification and regression trees...

It is therefore necessary to have a good discriminant method which takes into account the geographical location. So, in (Biernacki et al., 2002), we introduced a discriminant rule based on a Gaussian mixture model associated with the design

Beninel F. and Biernacki C. (2009).
UPDATING A LOGISTIC DISCRIMINATION RULE - Comparing Some Logistic Submodels in Credit-scoring.
In *Proceedings of the International Conference on Agents and Artificial Intelligence*, pages 267-274
DOI: 10.5220/0001662302670274

vector of morphometric characteristics.

In our problem of *credit scoring*, the bank has to predict the behavior of borrowers to pay back loan, on the basis of variables description. For this second example, the subpopulations result from *differences elsewhere*: customers and not-customers. These differences could influence (in addition to covariates) the target variable. It is obvious that informations related to customers are more reliable than those related to *not-customers*. For example, the debt ratio and expenditures may be underestimated among the *not-customers* when requesting the loan.

An other example in credit scoring is when subpopulations result from *changes over time*.
In this case, a first discriminant rule predicting borrowers behaviour classes is built. Such a rule is derived from the observation of borrowers over a time interval $[T, T+1]$ (as from a population Ω). In addition, when these individuals are observed again over a new interval $[T+\tau, T+\tau+1]$ of the same length (as from population Ω^*), another allocation rule is often necessary.
Obviously, changes in the economic and social environments could induce significant changes in the population of borrowers and could affect the *risk credit*.

As pointed out in (Tuffery, 2007), implementation of an allocation rule devoted to the prediction of the risk classes requires stability in the studied population and in the distribution of the available covariates. In the issue that we study, the two subpopulations are not exchangeable *i.e.*, there is an experienced rule defined on a first subpopulation and a learning sample of small size from some different second one.

Here, by allocation rule we mean a decision function $\Psi_\theta = (\psi_{\theta 1}, \ldots, \psi_{\theta g})$ $(\mathbb{R}^d \rightarrow \mathbb{R}^g)$ such that $x \in \mathbb{R}^d$ is allocated to class with label $k_0 = arg\max_{k=1,\ldots,g} \psi_{\theta k}(x)$ where θ is the associated parameter.
Usually, $\psi_{\theta k}(x)$ is a posterior probability to belong in the class k or more generally, a corresponding score (as the Anderson score, for example).
Hence, given a decision function or a classifier ψ_θ, one could consider that the experienced discriminant rule on Ω is given unless we have the estimate $\widehat{\theta}$. Then, the only remaining problem is to estimate the parameter θ^* corresponding to the discriminant rule on Ω^*.

Usually, two classical approaches are used to obtain an estimation of θ^*: The first approach consists in taking the same estimate than in Ω *i.e.*, $\widehat{\theta^*} = \widehat{\theta}$ and a the second approach in determining $\widehat{\theta^*}$ using only the learning sample $S^* \subset \Omega^*$.
If we denote by ν the number of components of θ^*, one could present the first approach as leading to the estimate $\widehat{\theta^*} = g_1(\widehat{\theta})$ where $g_1 = Id_\nu(\mathbb{R}^\nu \mapsto \mathbb{R}^\nu)$ and the second one as leading to the estimate $\widehat{\theta^*} = g_2(S^*)$ with $g_2(\mathbb{R}^{Card(S^*)\times\nu} \mapsto \mathbb{R}^\nu)$.

The first approach does not take into account the difference between the two subpopulations. The second one needs a learning sample of a sufficient size and here we deal with the problem of a small one. This raises the problem of accuracy of the estimate $\widehat{\theta^*} = g_2(S^*)$.
Thus, the problem here, is to take account of the characteristics of the available sample as recommended rightly by David Hand (Hand, 2005). He noted that the advantage of an advanced method of modelling relatively to a simple one (linear, for example) is often in a better modeling of the study sample.

To circumvent the problem of specific data, we exploit the idea that information related to one of the two subpopulations contains some information related to the other. Thus, we search an acceptable relationship between the two available distributions (*i.e.*, the distribution of covariates on Ω and this one on Ω^*).
The relationship between distributions of covariates on Ω and Ω^* induces a parametric relationship $\theta^* = \Phi_\gamma(\theta)$ between parameters.
The estimation method to derive θ^* is a *plug in* one *i.e.*, given the link function Φ_γ and considering $\theta = \widehat{\theta}$, we use the learning sample S^* to estimate γ. The estimate depends now on S^* and θ *i.e.*,

$$\widehat{\theta^*} = \Phi_{\widehat{\gamma}(S^*)}(\widehat{\theta}) = g(\widehat{\theta}, S^*). \quad (1)$$

The problem of the smallness of the sample S^* arises again when estimating γ. However, the number of components of γ should be much lower than those of θ^*. Hence, this could be well appropriate.

In the case of the Gaussian mixture model, this *plug in* approach appears very promising. In (Biernacki et al., 2002) we introduced a somewhat similar *plug in* method to build a generalized discriminant rule devoted to prediction on a Gaussian subpopulation (*i.e.*, the restriction of the covariates vector is a Gaussian per class), learning on another one.
In this work, we extend this idea to the logistic discriminant model *i.e.*, for each of the two subpopulations the response variable depends on covariates

according to a logistic model. θ and θ^* are respectively the vectors of covariates (including intercept) effect. The link between the two subpopulations consists in a direct relationship between the parameters *i.e.*, $\theta^* = \Phi_\gamma(\theta)$.
Given the function Φ_γ, each system of constraints on γ generates a logistic submodel. We focus on the estimation of each logistic submodel and the comparison of some of these submodels from the error-cost point of view.

2 GENERALIZED LOGISTIC DISCRIMINATION

2.1 The Logistic Model

Let $x \in \mathbb{R}^d$ be a vector of covariates and an associated response variable $Y \sim \mathcal{M}_g(1, \pi_1, \ldots, \pi_g)$ where $g \geq 2$. Let us set $t_k(x,\theta) = \mathbb{P}(Y = k|x;\theta)$, with $\theta = \{(\beta_{0k}||\beta_k') \in \mathbb{R}^{d+1}, k = 1, \ldots, g\}$. Here, $(\beta_{0k}||\beta_k')$ is the concatenation of the k^{th} intercept and the k^{th} vector of covariates effect.

The multinomial logistic model is defined by the *generalized logit* given by the following equation

$$\log\left(\frac{t_k(x,\theta)}{t_g(x,\theta)}\right) = \beta_{0k} + \beta_k' x. \tag{2}$$

Equivalently, the model is defined by the probability distribution of $Y_{|x}$, given by

$$t_k(x,\theta) = \frac{\exp(\beta_{0k} + \beta_k' x)}{1 + \sum_{j=1}^{g-1} \exp(\beta_{0j} + \beta_j' x)}, \quad k = 1, \ldots, g \tag{3}$$

and where $(\beta_{0g}||\beta_g')$ is the null vector of $\mathbb{R}^{d+1}$.

The discriminant rule based on this probabilistic model, in the case of uniform errors cost, consists in assigning the observation $x \in \mathbb{R}^d$ to the group $k_0 = arg\max_{k=1,\ldots,g} t_k(x,\theta)$.
For the general case, including non uniform errors cost, $k_0 = arg\min_{l=1,\ldots,g}\{\sum_{k=1}^{g} C(k|l)t_k(x,\theta)\}$, where $C(k|l)$ is the misallocation cost value, when assigning an observation from class $\{Y = l\}$ to class $\{Y = k\}$.

The aim of this communication is the study and comparison of some logistic submodels (or constrained logistic models) resulting from situations where one has an experienced rule to predict on a first subpopulation, a small learning sample from the second which contains the individuals to predict.

2.2 The Logistic Mixture Model

Let us denote

- Ω, Ω^* two subpopulations from a same population and p, p^* the associated prior probabilities,
- $\widetilde{X} \in \mathbb{R}^d$ the covariates vector observed over the disjoint union $\Omega \sqcup \Omega^*$,
- $\widetilde{Y}$ a categorical target variable.

We set $(\widetilde{X},\widetilde{Y})_{|\Omega} = (X,Y)$ and $(\widetilde{X},\widetilde{Y})_{|\Omega^*} = (X^*,Y^*)$ and denote by $(x,y),(x^*,y^*)$ their respective values.
Here, we consider the logistic model, over Ω, as given by

$$\begin{cases} Y \sim \mathcal{M}_g(1, \pi_1, \ldots, \pi_g), \\ t_k(x,\theta) = \frac{\exp(\beta_{0k} + \beta_k' x)}{1 + \sum_{j=1}^{g-1} \exp(\beta_{0j} + \beta_j' x)}, \end{cases} \tag{4}$$

and over Ω^*, by

$$\begin{cases} Y^* \sim \mathcal{M}_g(1, \pi_1^*, \ldots, \pi_g^*), \\ t_k^*(x^*,\theta^*) = \frac{\exp(\beta_{0k}^* + \beta_k^{*\prime} x^*)}{1 + \sum_{j=1}^{g-1} \exp(\beta_{0j}^* + \beta_j^{*\prime} x^*)}, \end{cases} \tag{5}$$

where $t_k^*(x^*,\theta^*) = \mathbb{P}(Y^* = k|x^*;\theta^*)$.

Here we define the logistic mixture model as follows:

$$\text{for } \widetilde{x} \in \mathbb{R}^d, \quad \mathbb{P}(\widetilde{Y} = k|\widetilde{x}) = p t_k(\widetilde{x},\theta) + p^* t_k^*(\widetilde{x},\theta^*). \tag{6}$$

When the subpopulation of an observation ω such that $\widetilde{X}(\omega) = \widetilde{x}$ is unknown, its allocation (to the appropriate class) requires parameters p, p^*, θ, θ^*.

In the problem we solve in this paper, we have to predict individuals from Ω^* and so $\mathbb{P}(\widetilde{Y} = k|\widetilde{x}) = t_k^*(\widetilde{x},\theta^*)$. Consequently, we have to use an allocation rule which requires the only parameter θ^*.

2.3 Generalized Logistic Discrimination

Different problems of discrimination under the logistic mixture model, could be studied. The resolution of these problems depends on the relevance of the available data. Particularly, we identify two problems:

A first problem is the simultaneous estimation of θ and θ^*. A second problem is to estimate θ^* in situations of a given θ.
The resolution of the first problem requires two learning samples of sufficient size (one sample from each subpopulation). While the second problem requires only one of these samples, and this is the problem we have to study.

Specifically, we have already an allocation rule on Ω (*i.e.*, we have θ or more usually, a given estimate $\widehat{\theta}$ of θ) and we want to get a new rule to predict on Ω^* (*i.e.*, to estimate θ^*) with as available data to estimate a sample $S^* = \{(x_i^*, z_i^*) : \ i = 1, \ldots, n^*\}$.

The practice has resulted in the two following cases:

case1. we have a unique population (*i.e.*, we consider $\Omega = \Omega^*$ and therefore $\theta = \theta^*$),

case2. we detect the mixture *i.e.*, we consider $\Omega \neq \Omega^*$ and we estimate θ^* using only S^*.

Finally, in these usual practices, it is believed to know everything (case 1) or nothing on Ω^* (case 2). In real problems links between subpopulations could exist and consequently, informations on Ω could provide some information on Ω^*.

3 LINKS BETWEEN SUBPOPULATIONS

3.1 Linear Links Models

In this work, we limit the study to the models defined by a linear relationship between parameters θ^* and θ *i.e.*, for all $k = 1, \ldots, g-1$,

$$\beta_{0k}^* = \alpha_k + \beta_{0k}, \qquad \beta_k^* = \Lambda_k \beta_k \quad , \tag{7}$$

where $\alpha_k \in \mathbb{R}$ and Λ_k is a $d \times d$ diagonal matrix (or a d dimensional vector).

Replacing β_{0k}^* and β_k^* in Equation (5) by their values given by the Equation (7), we obtain the new parameterisation

$$t_k^*(x^*, \theta, \gamma) = \frac{\exp(\beta_{0k} + \alpha_k + \beta_k' \Lambda_k x^*)}{1 + \sum_{j=1}^{g-1} \exp(\beta_{0j} + \alpha_j + \beta_j' \Lambda_j x^*)}, \tag{8}$$

where $\gamma = \{(\alpha_k || \Lambda_k) \in \mathbb{R}^{d+1} : k = 1, \ldots, g-1\}$.

As it will be seen in subsection 3.2, linear link models defined by Equations (7) are those obtained when the random vectors $X_{|Y=k}$, $k = 1, \ldots, g$ (*resp.* $X_{|Y^*=k}^*$, $k = 1, \ldots, g$) are Gaussian homoscedastic.

The constrained situation where for all k, $\alpha_k = 0$ and $\Lambda_k = \mathbf{I}_d$ (the d-dimensional identity matrix), returns **case1** of the classical approach.
The situation where α_k and Λ_k are unconstrained, returns **case2**.

We will compare these two classical situations to intermediate parsimonious models. Thus, the purpose of this communication is the estimation and the comparison of the models listed below:

(**M1** $\equiv$ **case1**) $\alpha_k = 0$ and $\Lambda_k = \mathbf{I}_d$ for all $1 \leq k \leq g-1$. The score functions are invariable.

(**M2**) $\alpha_k = 0$ and $\Lambda_k = \lambda_k \mathbf{I}_d$ with $\lambda_k \in \mathbb{R}$. Each score function (corresponding to a fixed class) changes *w.r.t.* λ_k. The ranks corresponding to individual scores are invariant.

(M3) $\alpha_k \in \mathbb{R}$ and $\Lambda_k = \mathbf{I}_d$. The score functions differ only *w.r.t.* the intercept and thus, changes the threshold for assignment to classes. The differences between scores and the corresponding ranks are invariable.

(**M4**) $\alpha_k \in \mathbb{R}$ and $\Lambda_k = \lambda_k \mathbf{I}_d$. Here the ranking of the scores is invariable.

(**M5**) $\alpha_k = 0$ and $\Lambda_k \in \mathbb{R}^d$; the threshold is invariable but covariates coefficients could change.

(**M6** $\equiv$ **case2**) $\alpha_k \in \mathbb{R}$ and $\Lambda_k \in \mathbb{R}^d$. All parameters are free.

If we denote by $\prec$ the symbol of nesting between models, we establish the partial ranking **M1** $\prec$ **M2** $\prec$ **M5** $\prec$ **M6** and **M1** $\prec$ **M3** $\prec$ **M4** $\prec$ **M6**.
These relations are used to compare models with information criteria as the Schwarz criterion (BIC) or the Akaike one (AIC) (Lebarbier and Mary-Huard, 2006).

3.2 Results from Homoscedastic Gaussian Model

For each subpopulation the design vector is a mixture of homoscedastic Gaussian distributions *i.e.*,

$$\forall k = 1, \ldots, g \quad X_{|k} \sim N_d(\mu_k, \Sigma) \text{ and } X_{|k}^* \sim N_d(\mu_k^*, \Sigma^*).$$

The link between Ω and Ω^* is given by

$$X_{|k}^* \stackrel{d}{=} D_k X_{|k} + b_k, \quad k = 1, \ldots, g, \tag{9}$$

where D_k is a diagonal real matrix. It's known from (De Meyer et al., 2000) that the link using a linear function is the only link $\phi_k = (\phi_{k1}, \ldots, \phi_{kd})$ such that $X_{|k}^* \stackrel{d}{=} \phi_k(X_{|k})$ and which verifies the assumptions **A1** and **A2** that follow.

A1 : ϕ_k is a *component to component* link *i.e.*, function $\phi_{kj}(\mathbb{R}^d \mapsto \mathbb{R})$ transforms the only j^{th} component. Hence, we consider ϕ_{kj} as an $(\mathbb{R} \mapsto \mathbb{R})$ function.

A2 : ϕ_{kj} is a $\mathcal{C}^1$ function.

We derive from the Equation (9) the following relations between parameters of the two Gaussian distributions.

$$\mu_k^* = D_k\mu_k + b_k, \quad \Sigma_k^* = D_k\Sigma D_k. \tag{10}$$

The matrices D_k, allowing equal variances (see. Equation (10)), are such that $D_k = A_kD$ with D a diagonal matrix and A_k another diagonal matrix with diagonal components in $\{-1,+1\}$.
We consider a link model as given by a set $\{D,A_1,\ldots,A_g,b_1,\ldots,b_g\}$.

It is well known that there exists a particular link between parameters of a generating Gaussian mixture model and those of a the corresponding logistic one (Anderson, 1982): for $k = 1,\ldots,g$, note f_k the density function of the Gaussian distribution $N_d(\mu_k,\Sigma)$, we have the Bayes formulae

$$\mathbb{P}(Y=k|x) = \frac{\pi_k f_k(x)}{\sum_{j=1}^g \pi_j f_j(x)}. \tag{11}$$

We derive the *generalized logit* (where g is the reference category)

$$\begin{aligned}\log(\tfrac{\mathbb{P}(Y=k|x)}{\mathbb{P}(Y=g|x)}) &= (\mu_k-\mu_g)'\Sigma^{-1}x + \log(\tfrac{\pi_k}{\pi_g}) \\ &+\tfrac{1}{2}(\|\mu_g\|^2_{\Sigma^{-1}} - \|\mu_k\|^2_{\Sigma^{-1}}).\end{aligned} \tag{12}$$

Consequently, the parameters given by the following equations

$$\begin{cases} \beta_{0k} = \log(\tfrac{\pi_k}{\pi_g}) + \tfrac{1}{2}(\|\mu_g\|^2_{\Sigma^{-1}} - \|\mu_k\|^2_{\Sigma^{-1}}), \\ \beta_k = \Sigma^{-1}(\mu_k-\mu_g). \end{cases} \tag{13}$$

are logistic parameters (corresponding to the intercept and covariates effect).

In an analogous manner, the parameters of the logistic model derived from the Gaussian subpopulation Ω^*, are

$$\begin{cases} \beta^*_{0k} = \log(\tfrac{\pi^*_k}{\pi^*_g}) + \tfrac{1}{2}(\|\mu^*_g\|^2_{\Sigma^{*-1}} - \|\mu^*_k\|^2_{\Sigma^{*-1}}), \\ \beta^*_k = \Sigma^{*-1}(\mu^*_k-\mu^*_g). \end{cases} \tag{14}$$

Using Equations (10) and setting $D_k = A_kD$ and $b_k = b$, we establish for the model (of link) $(D,A_1,\ldots,A_g,b_1,\ldots,b_g)$, the equations returning the link between parameters of the logistic models corresponding to the two subpopulations. More precisely

$$\begin{cases} \beta^*_{0k} = \beta_{0k} + \alpha_k, \\ \beta^*_k = A_kD\beta_k, \end{cases} \tag{15}$$

where $\alpha_k = \alpha(\mu_k,\mu_g,\Sigma,b,\pi^*_k,\pi^*_g) = \log(\frac{\pi^*_k}{\pi^*_g}) + < \mu_g, D_g^{-1}b >_{\Sigma^{-1}} - < \mu_k, D_k^{-1}b >_{\Sigma^{-1}}$.

4 PARAMETERS ESTIMATION

4.1 The Maximum Likelihood Method

The problem now is to estimate the parameters $\gamma = (\alpha_1,\ldots,\alpha_{g-1},\Lambda_1,\ldots,\Lambda_{g-1})$ involved in Equation (8) giving, for an individual from Ω^*, the corresponding probabilities of belonging to classes. The estimation is based on sample $S^* = \{(Y_i^*,x_i^*) : i = 1,\ldots,n^*\}$.

We use the maximum likelihood estimator. The conditional maximized likelihood is

$$L_{Y^*|x^*}(\gamma) = \prod_{i=1}^{n^*}\prod_{k=1}^{g} t_k^*(x_i^*,\theta,\gamma)^{Z_{ik}}, \tag{16}$$

where $Z_{ik} = 1$ if $Y_i^* = k$ and 0 elsewhere.
That is to maximize the log-likelihood expressed by the equation

$$\begin{aligned}\mathcal{L}_{Y^*|x^*}(\gamma) &= \textstyle\sum_{i:Z_{ig}=1} \log\left(\frac{1}{1+\sum_{j=1}^{g-1}\exp(h_j(x_i^*))}\right) \\ &+\textstyle\sum_{k=1}^{g-1}\sum_{i:Z_{ik}=1}\log\left(\frac{\exp(h_k(x_i^*))}{1+\sum_{j=1}^{g-1}\exp(h_j(x_i^*))}\right),\end{aligned} \tag{17}$$

where h_k is the k^{th} Anderson score *i.e.*,

$$h_k(x^*) = \beta^*_{0k} + \beta^{*'}_k x^* = \beta_{0k} + \alpha_k + \beta'_k\Lambda_k x^*.$$

According to the constraints imposed on γ, it leads to a non-linear equations system. Table 1 gives for each model (or each corresponding equations system) the number of unknown parameters to estimate.

Table 1: Here, ν is the dimension of the estimated parameter γ.

model	**M2**	**M3**	**M4**	**M5**	**M6**
ν	$g-1$	$g-1$	$2g-2$	$dg-d$	dg

In (Beninel and Biernacki, 2007), we give for the case $g = 2$, the system of likelihood equations, the corresponding Hessian and condition of the uniqueness of the solution. Here we treat the more complex case $g > 2$, leading to a more complex non-linear equations system, but without more difficulties from the mathematical point of view.

4.2 Using an Avalaible Logistic Procedure

The estimation method could be reduced to the use of an existing logistic procedure as the proc LOGISTIC in SAS system. We present such a technique in the context of a dichotomic response variable.

The unique Anderson score is

$$h(x^*) = \beta_0 + \alpha + \sum_{j=1}^{d} \lambda_j(\beta_j x^{*j}), \qquad (18)$$

where λ_j is the j^{th} diagonal component of matrix Λ, β_j and x^{*j} respectively the j^{th} component of β and of the design vector $\mathbf{x}^*$.

Let us set $\beta_0^* = \beta_0 + \alpha$ and $\widetilde{x}^* = \beta * x^*$ the vector obtained using a *component to component* product (or, Hadamard product) and $\widetilde{x}^{*j}$ its j^{th} component. $\widetilde{X}^*$ is the new design vector and we can view $\widetilde{x^*}^j$ as the j^{th} weighted covariate. The score function given by equation (18) is now written as

$$h(x^*) = \beta_0^* + \sum_{j=1}^{d} \lambda_j \widetilde{x^*}^j, \qquad (19)$$

Consequently, for each model among **M2**,...,**M6** we have to estimate (using an available logistic procedure) $\gamma = (\beta_0^*, \lambda_1, \ldots, \lambda_d) \in \Gamma \subset \mathbb{R}^{d+1}$ on the basis of the transformed learning sample $\widetilde{S}^* = \{(Y_i^*, \widetilde{x}_i^*) : i = 1, \ldots, n^*\}$.

Depending on the model, the dimension of the parameters space Γ is variable. We explicit for each model the numerical computation *via* the LOGISTIC procedure.
We set $z = \sum_{j=1}^{d} \widetilde{\mathbf{x}}^{*j}$, corresponding to the Anderson score related to the logistic model on Ω. We give in the following the Anderson score on Ω^*, depending on the transformed data.

M2 : We have to estimate the parameter $\lambda \in \mathbb{R}$ such that

$$h(x^*) = \beta_0 + \lambda z, \qquad (20)$$

Here, the intercept is fixed as equal to β_0.

M3 : We estimate the intecept $\beta_0^* \in \mathbb{R}$ *i.e.*,

$$h(x^*) = \beta_0^* + z. \qquad (21)$$

Here the effect of the covariate Z is constrained to be equal to one.

M4 : We estimate $(\beta_0^*, \lambda) \in \mathbb{R}^2$ such that

$$h(x^*) = \beta_0^* + \lambda z. \qquad (22)$$

We have to use the available logistic procedure without constraints.

M5 : We have to estimate $\Lambda^* \in \mathbb{R}^d$ such that

$$h(x^*) = \beta_0 + \Lambda^* \widetilde{x}^*. \qquad (23)$$

The intercept is constrained to be equal to β_0.

Here we consider model (**M1**) as the simplest model and model (**M6**) as the more complex model.
When the response variable is polytomic ($g > 2$), the number of possible constrained models is much larger. For example for $g = 3$, we identify 15 submodels to compare.

5 NUMERICAL EXPERIMENTS

5.1 Data Description

The data are from a German bank and cover a sample of 1000 consumer's credits. Each of these consumer is described by a binary response variable `Kredit` ($\{$`Kredit` $= 1\}$ for credit-worthy or $\{$`Kredit` $= 0\}$ for not credit-worthy). In addition, 20 covariates of different types (continuous, nominal, ordinal) assumed to influence creditability are recorded. Examples of these covariates are:

`Hoehe`: the amount of credit in "Deutche Mark" [metrical],

`Laufzeit`: duration of credits in months [metrical],

`Laufkont`: account balance [categorical],

`Moral`: behaviour repayment of other loans [categorical]...

For a complete access to these data we refer to the book (Fahrmeir and Hamerle, 1984) or the current website `http://www.stat.uni-muenchen.de/service/datenarchiv/Kredit`.

These data are also described in the book of (Fahrmeir and Tutz, 1994) (see. pages 31–34). The prior probabilities corresponding to the categories of the response variable are structurally unbalanced.Thus, for a consistent estimation of the *logit* model the given sample is stratified (300 consumers such that $\{$`Kredit` $= 0\}$ and the remaining 700 consumers such that $\{$`Kredit` $= 1\}$.

These data are frequently used by specialists of *credit scoring* when testing, calibrating and comparing methods.

5.2 Covariates Selection

In order to evaluate the data quality and the influence of the covariates (on creditability) a primary data processing is realized including univariate and bivariate statistics. Bivariate statistics measuring the dependency between the selected covariates and the target variable are computed.
This primary data processing highlights categories of covariates with null frequencies of responses 0 or 1 of the target variable. Such a characteristic in data creates the separability which implies divergence of the likelihood maximization algorithm, when estimating the parameters of the logistic model. In such a situation these categories are combined with close categories of the same covariate.
In addition, the primary data allows to determine covariates which influence the target variable. Thus, logistic regression is tested with different combinations of covariates among those appearing as jointly influencing variable `kredit`. Like other authors, who worked on these data, the more influencing covariates are those introduced in section 5.1 and variables `Beszeit` (present employement since) and `Sparkont` (savings account).
Apart from `Sparkont` , the other covariates are unchanged. In effect, to avoid the separability situation, categories 4 and 5 of `Sparkont` are grouped together.

5.3 Subpopulations Definition

We use the variable `Laufkont` to carry out the separation in two subpopulations. The non-customers of the bank(`Laufkont` = 1) constitute a subpopulation and the customers (`Laufkont` > 1) constituting a second subpopulation. Although `laufkont` is a covariate, we use it to define the two subpopulations and avoid bias of the difference in the amount and reliability of data related to the two subpopulations.

5.4 Experiments Description

We implement in SAS the program hat manages data and estimate the models. The implementation of these models extending logistic regression is as follows:
Step 1: We apply the logistic regression with as *primary data* the design matrix related to customers (or the learning sample $S_L \subset \Omega$). The obtained estimate θ is used to compute the new covariates subject to step2: The continuous covariates are multiplied by corresponding component of θ and binary covariates (or categories of qualitative variables) are multiplied by the corresponding parameter.
Step 2: The second step consists in the estimation of parameters related to models **M2** to **M6**. Such an estimation is based on a learning sample $S^* \in \Omega^*$. Sample S^* is derived (from *non-customers*) using the `surveyselect` procedure to obtain a stratified random sample. Percentages of responses $\{\texttt{kredit} = 1\}$ and $\{\texttt{kredit} = 0\}$ are close to that ones of Ω^*.
Given the sample size, the simulations are to draw B samples S^* from *non-customers* to estimate the 5 models and for each of S^* corresponds a test sample S_T^* of the remaining *non-customers*.
Let $C(l|k)$ denote the cost of misallocation of a borrower from $\{Y_i^* = k\}$ into $\{Y_i^* = l\}$, $k,l = 0,1$. Let us set $\rho(1,0) = \frac{C(1|0)}{C(0|1)}$, corresponding to error costs ratio. Under acceptable assumptions related to the prior probabilities p_0, p_1 and for each fixed pair (S^*, S_T^*) we estimate the exact error-rate of assignation to classes given by

$$C(0|1)\mathbb{P}(\widehat{Y^*} = 0|Y^* = 1) + C(1|0)\mathbb{P}(\widehat{Y^*} = 1|Y^* = 0)$$

(or without loss, $\mathbb{P}(\widehat{Y^*} = 0|Y^* = 1) + \rho(0,1)\mathbb{P}(\widehat{Y^*} = 1|Y^* = 0)$).
For fixed $\rho(1,0), p1, p0$, we get at $B = 30$ iterations a stratified random sample of frequencies (n_0, n_1) (n_0 the number of responses $\{\texttt{kredit} = 0\}$ in the learning sample and n_1 the number of responses $\{\texttt{kredit} = 1\}$.

From Figure A.1, it appears that the ranks of BIC values do not depends on $\rho(1,0)$ values. Models **M3**, **M4** seem the best models (to generate data) among **M1**, ..., **M6**.
Figure A.2, below gives the mean risk according to the costs ratio. The sample size here is set at $n_0 = n_1 = 20$.
For all values of $\rho(1,0)$, models **M2** and **M5** appear the best ones, from the risk point of view. The most practiced model **M1**, is very bad.

6 CONCLUSIONS

The simulations confirm the difference between subpopulations. It appears that the best models from the cost point of view (**M2** and **M5**), are not the generating best models. Indeed, models **M3** and **M4** minimize the BIC criterion.
The models **M2** and **M5** are those who move the intercept. This coincides with the fact that for stratified samples (as here), instead of the estimation of the intercept β_0^*, one estimates $\beta_0^* \pm \log(\pi n_0/(1-\pi)n_1)$.
In this work, the use of estimated logistic discriminant rule is simple from the programming point of view as we adapt an existing SAS procedure.

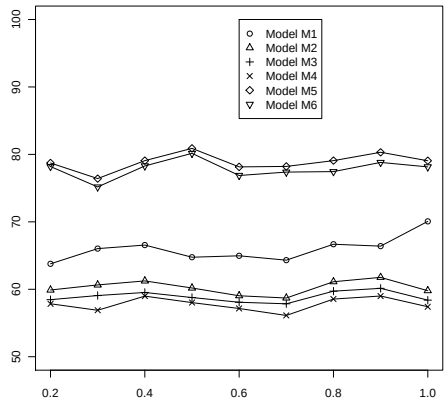

Figure 1: Mean BIC value depending on $\rho(1,0)$ value.

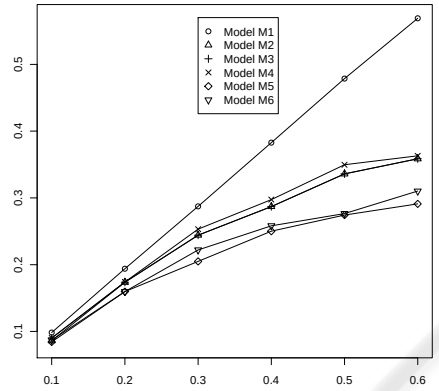

Figure 2: Mean risk value depending on $\rho(1,0)$ value.

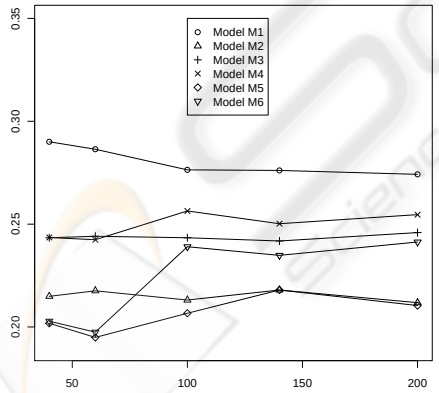

Figure 3: Mean Risk value depending on the learning sample size value.

REFERENCES

Anderson, J. A. (1982). Logistic discrimination. In *Handbook of Statistics (Vol. 2), P.R. Krishnaiah and L. Kanal (Eds.). Amsterdam: North Holland*, pages 169–191.

Beninel, F. and Biernacki, C. (2007). Relaxations de la régression logistique: modèles pour l'apprentissage sur une sous-population et la prédiction sur une autre. *RNTI*, A1:207–218.

Biernacki, C., Beninel, F., and Bretagnolle, V. (2002). A generalized discriminant rule when training population and test population differ on their descriptive parameters. *Biometrics*, 58:387–397.

Bretagnolle, V., Genevois, F., and Mougeot, F. (1998). Intra and intersexual function in the call of a non passerine bird. *Behaviour*, 135:1161–1202.

De Meyer, B., Roynette, B., Vallois, P., and Yor, M. (2000). On independent times and positions for brownian motion. *Institut Elie Cartan*, 1.

Fahrmeir, L. and Hamerle, A. (1984). *Multivariate statistiche Verfahren*. De Gruyter, Berlin.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Series in Statistics. Springer-Verlag, New York.

Hand, D. J. (2005). Classifier technology and the illusion of progress. *Technical Report, Imperial college, London*.

Lebarbier, E. and Mary-Huard, T. (2006). Une introduction au critère bic : fondements théoriques et interprétation. *JSFDS*, 147(1):39–57.

Thibault, J. C., Bretagnolle, V., and Rabouam, C. (1997). Cory's shearwater calonectris diomedia. *Birds of Western Paleartic Update*, 1:75–98.

Tuffery, S. (2007). Améliorer les performances d'un modèle prédictif: perspectives et réalité. *RNTI*, A(1):42–74.

Zink, R. and Remsen, J. (1986). Evolutionary processes and patterns of geographic variation in birds. *Current Ornithology*, 4:1–69.