Extension of model-based classification for binary data when training and test populations differ

Julien Jacques, Christophe Biernacki

Extension of model-based classification for binary data when training and test populations differ

J. Jacques[†1*] and C. Biernacki[†]
[†]*Laboratoire Paul Painlevé UMR CNRS 8524, Université Lille I*

Abstract

Standard discriminant analysis supposes that both the training sample and the test sample are derived from the same population. When these samples arise from populations differing from their descriptive parameters, a generalization of discriminant analysis consists in adapting the classification rule related to the training population to another rule related to the test population, by estimating a link map between both populations. This paper extends an existing work in the multinormal context to the case of binary data. In order to solve the problem of defining a link map between the two binary populations, it is assumed that the binary data result from the discretization of latent Gaussian data. An estimation method and a robustness study are presented, and two applications in a biological context illustrate this work.

1 Introduction

Consider a first (training) sample of individuals described by explanatory variables, for which a partition in groups or classes is known. Consider also a new sample of individuals (test sample), drawn from the same population of the training sample. Discriminant analysis consists in estimating an allocation rule from the training sample in order to classify individuals of the test sample (see McLachlan [1992] for a survey).
Since works of Fisher [1936], who introduced a linear discriminant rule between two groups, numerous evolutions have been proposed. All of them concern the nature of the discriminant rule: Parametric quadratic rule (see for example Tomassone et al. [1988]), semi parametric rule (as logistic discrimination, Anderson [1972]) or non parametric rule (Fix and Hodges [1951], Friedman and Stuetzle [1981], Hand [1982], Silverman [1986]).
An alternative approach, introduced by Van Franeker and Ter Brack [1993] and developed further by Biernacki et al. [2002], considers the case in which the training sample does not necessarily arise from the same population as the one of the test sample. Biernacki et al. define several models of *generalized discriminant analysis* in a multinormal context, and conduct experiments for biological data consisting of birds from the same species, but with different geographical origins. In many

[*1]Corresponding author. Email: julien.jacques@polytech-lille.fr

domains (insurance, medicine, biology, *etc*.) a large number of applications deals with binary data as well. The goal of the present paper is to extend the generalized discriminant analysis of Biernacki et al. [2002] to the case of binary data.
The paper is organized as follows. The next section presents the data and the latent class model for both training and test populations. Section 3 makes the assumption that these binary data are discretized latent continuous variables in which the order information is lost. This hypothesis is the key to establish a general stochastic link map between the two populations, from which many pertinent parsimonious sub-models can be obtained. In Section 4, estimation is performed based on the maximum likelihood principle using the EM algorithm. Then, a robustness study for the Gaussian assumption is carried out in Section 5 involving both theoretical and experimental (simulated data) arguments. In Section 6, two applications in a biological context illustrate realistic situations where the proposed generalized discriminant analysis outperforms standard discriminant analysis and clustering. Finally, the last section concludes this paper by discussing possible extensions of the present work.

2 The data and the latent class model

The data consist in two samples: The first sample S, labelled and drawn from the training population P, and the second sample $\tilde{S}$, unlabelled and drawn from the test population $\tilde{P}$. A fundamental assumption of the present work is that populations P and $\tilde{P}$ may be different.
The training sample S is composed of n pairs $(\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n)$, where $\mathbf{x}_i$ is the binary explanatory vector for the ith object with $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^T \in \{0,1\}^d$ and where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})^T$ is the group membership with z_{ik} being equal to 1 if the ith object belongs to the kth group and being equal to 0 otherwise ($i = 1, \ldots, n$, $k = 1, \ldots, K$). The number of binary explanatory variables and the number of groups are respectively denoted by d and K. Each pair $(\mathbf{x}_i, \mathbf{z}_i)$ is assumed to be an independent realization of the random vector $(\mathbf{X}_1, \mathbf{Z}_1)$ with distribution:

$$X_{1j_{|Z_{1k}=1}} \sim \mathcal{B}(\alpha_{kj}) \quad \text{for all } j = 1, \ldots, d \quad \text{and} \quad \mathbf{Z}_1 \sim \mathcal{M}(1, p_1, \ldots, p_K), \tag{1}$$

where $\mathcal{B}(\alpha_{kj})$ is the Bernoulli distribution of parameter α_{kj} ($0 < \alpha_{kj} < 1$), and $\mathcal{M}(1, p_1, \ldots, p_K)$ defines the one order multinomial distribution of parameters $p_1, \ldots, p_K$ ($0 < p_k < 1$, $\sum_{k=1}^{K} p_k = 1$). Moreover, using the latent class model assumption of conditional independence of the explanatory variables (Celeux and Govaert [1991], Everitt [1984]), the probability function of $\mathbf{X}_1$ conditionally to the group membership is:

$$f_k(x_{11}, \ldots, x_{1d}) = \prod_{j=1}^{d} {\alpha_{kj}}^{x_{1j}} (1 - \alpha_{kj})^{1 - x_{1j}}. \tag{2}$$

This is one of the most popular generative method for discriminating categorical data since it is straightforward to implement and it is often efficient (see for instance experimental comparisons with alternative methods, not necessarily generative, in **?**, **?** Chap. 9 or also **?**).
Similarly, the test sample $\tilde{S}$ is composed of $\tilde{n}$ pairs $(\tilde{\mathbf{x}}_1, \tilde{\mathbf{z}}_1), \ldots, (\tilde{\mathbf{x}}_{\tilde{n}}, \tilde{\mathbf{z}}_{\tilde{n}})$, where the d explanatory variables are the same as in the training sample, but where the $\tilde{\mathbf{z}}_i$ are unknown. These pairs are

assumed to be independent realizations of $(\tilde{\mathbf{X}}_1, \tilde{\mathbf{Z}}_1)$ with distribution:

$$\tilde{X}_{1j_{|\tilde{Z}_{1k}=1}} \sim \mathcal{B}(\tilde{\alpha}_{kj}) \quad \text{for all } j = 1, \ldots, d \quad \text{and} \quad \tilde{\mathbf{Z}}_1 \sim \mathcal{M}(1, \tilde{p}_1, \ldots, \tilde{p}_K). \tag{3}$$

The explanatory variables $\tilde{X}_{1j}$, for $j = 1, \ldots, d$, are also assumed to be conditionally independent. Basically, the distribution of $(\tilde{\mathbf{X}}_1, \tilde{\mathbf{Z}}_1)$ differs from this one of $(\mathbf{X}_1, \mathbf{Z}_1)$ only by the values of the parameters α_{kj} and p_k.
Our goal is to estimate the unknown labels $\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_{\tilde{n}}$ by using information from both training and test samples. The challenge resides in finding a link map between the populations P and $\tilde{P}$.
Remarks

- In fact, since both labelled and unlabelled data are used together in the inference process, our problem is related to the so-called semi-supervised learning purpose. Obviously, the originality of our work is that the data sets do not necessarily arise from the same population.
- The use of the terminology "test" for the sample $\tilde{S}$ (and the population $\tilde{P}$) is abusive because this sample is used to determine the discrimination rule. Nevertheless, this terminology is adopted in order to facilitate the link with the standard discrimination methods. Moreover, it appears to be a usual notation in the semi-supervised classification community (see for instance **?**, Chap. 1).

3 Relationship between test and training populations

3.1 Formalizing the link between populations

In a multinormal context, a linear stochastic relationship between P and $\tilde{P}$ is not only justified (under very few assumptions that will be recalled later) but also intuitive (Biernacki et al. [2002]). In the binary context, since such an intuitive relationship seems more difficult to exhibit, an additional assumption is stated: The binary variables are supposed to result from the discretization of some latent Gaussian variables. For instance, if a binary variable is a product purchased by a customer, it is assumed that the customer gives a score to the product, and buys it only if this score is greater than a given threshold. This assumption is not new in statistics: See for example Thurstone [1927], who used this idea in his comparative judgment model to choose between two stimuli. **?** also modelled multivariate ordered categorical variables as a latent multinormal distribution involving a possibly full correlation matrix (see **?** for a more recent reference). Moreover, Everitt [1988] proposed a classification algorithm for binary, categorical and continuous data.
Thus, the explanatory variables $X_{1j_{|Z_{1k}=1}}$ having a Bernoulli distribution $\mathcal{B}(\alpha_{kj})$ are assumed to arise from the discretization of latent continuous variables $Y_{1j_{|Z_{1k}=1}}$ in the following manner:

$$X_{1j_{|Z_{1k}=1}} = \begin{cases} 0 & \text{if } \lambda_j Y_{1j_{|Z_{1k}=1}} < \lambda_j s_j \\ 1 & \text{if } \lambda_j Y_{1j_{|Z_{1k}=1}} \geq \lambda_j s_j \end{cases} \qquad \text{for } j = 1, \ldots, d, \tag{4}$$

where $s_j \in \mathbb{R}$ is the discretization threshold, and $\lambda_j \in \{-1, 1\}$ is introduced to avoid choosing to which value of $X_{1j_{|Z_{1k}=1}}$, 0 or 1, corresponds a positive value of $Y_{1j_{|Z_{1k}=1}}$, and then to avoid binary

variables to inherit from the natural order induced by continuous variables.
Moreover, the joint distribution of $\mathbf{Y}_{1_{|Z_{1k}=1}} = (Y_{11_{|Z_{1k}=1}}, \ldots, Y_{1d_{|Z_{1k}=1}})^T$ is assumed to be multivariate normal, with marginal distributions $\mathcal{N}(\mu_{kj}, \sigma_{kj}^2)$, such the obtained discretized variables $X_{1j_{|Z_{1k}=1}}$ ($j = 1, \ldots, d$) are independent in order to retrieve the conditional independence assumption on the binary variables. It should be noted that conditional independence of the latent Gaussian variables $Y_{1j_{|Z_{1k}=1}}$ ($j = 1, \ldots, d$) is a sufficient assumption for conditional independence of the binary variables. However, this assumption may not be necessary.
From (1) and (4) the following relationship between α_{kj}, λ_j, μ_{kj} and σ_{kj} can be derived:

$$\alpha_{kj} = p(X_{1j_{|Z_{1k}=1}} = 1) = 1 - \Phi\big(\lambda_j \frac{s_j - \mu_{kj}}{\sigma_{kj}}\big) \tag{5}$$

where Φ is the $\mathcal{N}(0, 1)$ cumulative density function.
As for the variable X_{1j}, the binary variable $\tilde{X}_{1j}$ is also assumed to arise from the discretization of a latent Gaussian variable $\tilde{Y}_{1j}$ with distribution $\mathcal{N}(\tilde{\mu}_{kj}, \tilde{\sigma}_{kj}^2)$. The equations are the same as (4) and (5), by changing α_{kj} into $\tilde{\alpha}_{kj}$, μ_{kj} into $\tilde{\mu}_{kj}$ and σ_{kj} into $\tilde{\sigma}_{kj}$. The thresholds $\tilde{s}_j$ are naturally supposed to be equal to s_j ($\tilde{s}_j = s_j$) since, with the previous example, it is equivalent to assume that customers of both populations buy the product if the score is greater than the same threshold. In the same spirit $\tilde{\lambda}_j$ is supposed to be equal to λ_j, so the rule of purchase – lower or higher than the threshold – is the same for both populations.
In a Gaussian setting, Biernacki et al. [2002] showed that the only possible link map between the latent continuous variable $\mathbf{Y}_{1_{|Z_{1k}=1}}$ of P and $\tilde{\mathbf{Y}}_{1_{|\tilde{Z}_{1k}=1}}$ of $\tilde{P}$ is linear when the two following reasonable hypotheses are satisfied: (i) The transformation between P and $\tilde{P}$ is $\mathcal{C}^1$ and (ii) the jth component $\tilde{Y}_{1j_{|\tilde{Z}_{1k}=1}}$ of $\tilde{\mathbf{Y}}_{1_{|\tilde{Z}_{1k}=1}}$ is only related to the jth component $Y_{1j_{|Z_{1k}=1}}$ of $\mathbf{Y}_{1_{|Z_{1k}=1}}$. More precisely, this relationship is expressed by

$$\tilde{\mathbf{Y}}_{1_{|\tilde{Z}_{1k}=1}} \sim A_k \mathbf{Y}_{1_{|Z_{1k}=1}} + \mathbf{b}_k, \tag{6}$$

where A_k is a diagonal matrix of $\mathbb{R}^{d\times d}$ containing the elements a_{kj} and $\mathbf{b}_k$ is a vector of $\mathbb{R}^d$ containing the elements b_{kj} ($1 \leq k \leq K$, $1 \leq j \leq d$).
By using (6) and (5) the following relationship between the parameters $\tilde{\alpha}_{kj}$ and α_{kj} can be obtained (see details in Appendix A):

$$\tilde{\alpha}_{kj} = \Phi\Big(\delta_{kj}\ \Phi^{-1}(\alpha_{kj}) + \lambda_j \gamma_{kj}\Big), \tag{7}$$

where $\delta_{kj} \in \mathbb{R}^+ \setminus \{0\}$, $\lambda_j \in \{-1, 1\}$ and $\gamma_{kj} \in \mathbb{R}$. Note that this relationship corresponds to a linear link between the *probit* functions of both α_{kj} and $\tilde{\alpha}_{kj}$. Conditionally to the fact that α_{kj} are known (they will be estimated in practice), estimation of the Kd continuous parameters $\tilde{\alpha}_{kj}$ is thus obtained from the estimated parameters of the link between P and $\tilde{P}$ (plug-in method): δ_{kj}, γ_{kj} and λ_j. Note that the choice of the discretization thresholds s_j is not important. However, estimating the number of parameters for the link map is $2Kd$ and one thus obtain that the model is overparameterized. This fact should not be surprising since the underlying Gaussian model is by far more complex (in terms of the number of parameters) than the Bernoulli model. Hence we need to reduce the number of free continuous parameters in (7), and one way to do this is to propose sub-models defined *via* imposing natural additional constraints on the transformation between both populations P and $\tilde{P}$.

3.2 Models of constraints on the stochastic link map

The parameters δ_{kj} ($1 \leq k \leq K$ and $1 \leq j \leq d$) will be successively constrained to be equal to 1 (denoted by 1), to be class- and dimension-independent (δ), to be only class-dependent (δ_k) or only dimension-dependent (δ_j). In the same way, γ_{kj} can be constrained to be equal to 0, γ (constant w.r.t. k and j), γ_k (constant w.r.t. j) or γ_j (constant w.r.t. k). Thus, 16 models can be defined and indexed using the following *ad-hoc* notation: $[1\,0]$ means $\delta_{kj} = 1$ and $\gamma_{kj} = 0$ (it corresponds to the usual discriminant analysis model), $[\delta_k\,\gamma_j]$ means $\delta_{kj} = \delta_k$ and $\gamma_{kj} = \gamma_j$ ($1 \leq k \leq K$ and $1 \leq j \leq d$), *etc*.
For these 16 models, an additional assumption on the group proportions is taken into account: Either the proportions of $\tilde{P}$ are constrained to be equal to those of P, or they have to be estimated. In the following, $[p_k\,1\,0]$ denotes the model $[1\,0]$ with equal proportions whereas $[\tilde{p}_k\,1\,0]$ denotes this model with free proportions. The number of constrained models is thus growing to 32. Table 1 gives the number of (continuous) parameters to be estimated for these models. If the mixing proportions are different from P to $\tilde{P}$, $K-1$ must be added to these values.

Table 1: Number of continuous parameters (param.) to be estimated for the constrained models.

model	$[p_k\,1\,0]$	$[p_k\,1\,\gamma]$	$[p_k\,1\,\gamma_k]$	$[p_k\,1\,\gamma_j]$	$[p_k\,\delta\,0]$	$[p_k\,\delta\,\gamma]$	$[p_k\,\delta\,\gamma_k]$	$[p_k\,\delta\,\gamma_j]$
param.	0	1	K	d	1	2	$K+1$	$d+1$
model	$[p_k\,\delta_k\,0]$	$[p_k\,\delta_k\,\gamma]$	$[p_k\,\delta_k\,\gamma_k]$	$[p_k\,\delta_k\,\gamma_j]$	$[p_k\,\delta_j\,0]$	$[p_k\,\delta_j\,\gamma]$	$[p_k\,\delta_j\,\gamma_k]$	$[p_k\,\delta_j\,\gamma_j]$
param.	K	$K+1$	$2K$	$K+d$	d	$d+1$	$K+d$	$2d$

Finally, we chose to use the BIC criterion (*Bayesian Information Criterion*, Schwarz [1978]) for automatic selection among the 32 generalized discriminant models. However, other criteria such AIC (*An Information Criterion*, **?**) could be used as well. BIC is defined by:

$$\text{BIC} = -2l(\hat{\boldsymbol{\theta}}) + \nu \log(\tilde{n}),$$

where $\boldsymbol{\theta} = (\tilde{p}_k, \delta_{kj}, \lambda_j, \gamma_{kj})$ for $1 \leq k \leq K$ and $1 \leq j \leq d$, $l(\hat{\boldsymbol{\theta}})$ is the maximum log-likelihood corresponding to the estimation $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, and ν is the number of free continuous parameters associated to the given model. The model leading to the smallest BIC value is then selected.
Before estimating the parameter $\boldsymbol{\theta}$ by the maximum likelihood method, we need to discuss identifiability of each parametrization.

3.3 Model identifiability

Some of the constrained models previously defined can be non-identifiable. It is necessary to clarify these identifiability problems, which arise at two levels: Identifiability of the model parameters when α_{kj} is transformed into $\tilde{\alpha}_{kj}$, and identifiability of the transformation which ensures that α_{kj} can be only transformed into $\tilde{\alpha}_{kj}$ and not into $\tilde{\alpha}_{k'j}$ (with $k' \neq k$). We call respectively them *intra-group* and *inter-group* identifiability.
The reader can find theoretical and experimental discussion about these two kinds of identifiability in Appendix B. The conclusion of this discussion is identifiability will occur in practical situations.

4 Parameter estimation

In this section, only the situation where proportions are unknown is presented, otherwise it is straightforward.

4.1 The three estimation steps

Generalized discriminant analysis needs three estimation steps. The first step consists in estimating parameters p_k and α_{kj} ($1 \leq k \leq K$ and $1 \leq j \leq d$) from population P based on training sample S. Since S is a labelled sample, the maximum likelihood estimate is simply given by (Everitt [1984], Celeux and Govaert [1991]):

$$\hat{p}_k = \frac{1}{n}\sum_{i=1}^{n} z_{ik} \qquad \text{and} \qquad \hat{\alpha}_{kj} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} z_{ik}.$$

The second step consists in estimating parameters $\tilde{p}_k$ and $\tilde{\alpha}_{kj}$ ($1 \leq k \leq K$ and $1 \leq j \leq d$) of population $\tilde{P}$ by using $\hat{p}_k$, $\hat{\alpha}_{kj}$ ($1 \leq k \leq K$ and $1 \leq j \leq d$) and $\hat{\boldsymbol{\theta}}$. Thus, for estimating $\tilde{\alpha}_{kj}$, the parameters δ_{kj}, γ_{kj} and λ_j of the link between P and $\tilde{P}$ have to be estimated, and then an estimate of $\tilde{\alpha}_{kj}$ is deduced by plug-in inside Equation (7). This step is described below.
Finally, the third step consists in estimating group membership of individuals from the test sample $\tilde{S}$, by *maximum a posteriori*.

4.2 Estimation of the link parameters

For the second step above, maximum likelihood estimation can be efficiently based on the EM algorithm (Dempster et al. [1977]). The likelihood is given by:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{\tilde{n}} \sum_{k=1}^{K} \tilde{p}_k \prod_{j=1}^{d} \tilde{\alpha}_{kj}^{\tilde{x}_{ij}} (1-\tilde{\alpha}_{kj})^{1-\tilde{x}_{ij}},$$

and the completed log-likelihood is:

$$l_c(\boldsymbol{\theta}; \tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_{\tilde{n}}) = \sum_{i=1}^{\tilde{n}} \sum_{k=1}^{K} \tilde{z}_{ik} \log\left(\tilde{p}_k \prod_{j=1}^{d} \tilde{\alpha}_{kj}^{\tilde{x}_{ij}} (1-\tilde{\alpha}_{kj})^{(1-\tilde{x}_{ij})}\right).$$

The E step From a current value $\boldsymbol{\theta}^{(q)}$ of the parameter $\boldsymbol{\theta}$, the E step of the EM algorithm consists in computing the conditional expectation of the completed log-likelihood:

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)}) &= E_{\boldsymbol{\theta}^{(q)}}[l_c(\boldsymbol{\theta}; \tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_{\tilde{n}}) | \tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{\tilde{n}}] \\
&= \sum_{i=1}^{\tilde{n}} \sum_{k=1}^{K} t_{ik}^{(q)} \Big\{ \log(\tilde{p}_k) + \sum_{j=1}^{d} \log\left(\tilde{\alpha}_{kj}{}^{\tilde{x}_{ij}} (1-\tilde{\alpha}_{kj})^{1-\tilde{x}_{ij}}\right) \Big\}
\end{aligned}$$

where

$$t_{ik}^{(q)} = p(\tilde{Z}_{ik} = 1 | \tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{\tilde{n}}; \boldsymbol{\theta}^{(q)}) = \frac{\tilde{p}_k^{(q)} \prod_{j=1}^{d} (\tilde{\alpha}_{kj}^{(q)})^{\tilde{x}_{ij}} (1 - \tilde{\alpha}_{kj}^{(q)})^{(1-\tilde{x}_{ij})}}{\sum_{\kappa=1}^{K} \tilde{p}_\kappa^{(q)} \prod_{j=1}^{d} (\tilde{\alpha}_{\kappa j}^{(q)})^{\tilde{x}_{ij}} (1 - \tilde{\alpha}_{\kappa j}^{(q)})^{(1-\tilde{x}_{ij})}}$$

is the conditional probability for the individual i to belong to the group k.

The M step The M step of the EM algorithm consists in choosing the value $\boldsymbol{\theta}^{(q+1)}$ which maximizes the conditional expectation $\mathcal{Q}$ computed at the E step:

$$\boldsymbol{\theta}^{(q+1)} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}}\, \mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)}) \tag{8}$$

where Θ is a parameter space depending on the model at hand. The M step is now described for each component of $\boldsymbol{\theta} = \{\tilde{p}_k, \delta_{kj}, \lambda_j, \gamma_{kj}\}$.
For proportions, the maximum is:

$$\tilde{p}_k^{(q+1)} = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} t_{ik}^{(q)}.$$

The parameters δ_{kj} and γ_{kj} are never considered because the full generalized discriminant analysis model is overparameterized. Thus, only the constrained models are to be estimated. In this context, it is proved in Appendix C that $\mathcal{Q}$ is a strictly concave function of δ_k, δ_j, δ, γ_k, γ_j and γ. Thus, the maximisation of $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)})$ is computed by an alternated iterative algorithm which consists in a succession, componentwise, of simplex algorithms if the optimization is unconstrained ($\gamma_k, \gamma_j \in \mathbb{R}$). If the optimization is constrained ($\delta_k, \delta_j > 0$), the same algorithm is used but if the optimization leads to a negative result, the value 0 is retained (in this case the EM algorithm is used on its generalized form: GEM, Dempster et al. [1977]). The starting point of the alternating algorithm is $\boldsymbol{\theta}^{(q)}$, and this one of the EM algorithm, $\boldsymbol{\theta}^{(0)}$, is the point which corresponds to $P = \tilde{P}$.
For the discrete parameters λ_j, if the dimension d is sufficiently low, the maximization is carried out by computing $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)})$ for all 2^d possible values for these discrete parameters. Since computational limits are quickly reached, a relaxation method can be used, which consists in assuming that λ_j is not a binary parameter in $\{-1, 1\}$ but a continuous one in $[-1, 1]$, named λ_j^* (see Wolsey [1998] for instance). Optimization is thus performed on this continuous parameter, with the previous alternated algorithm since $\mathcal{Q}$ is a strictly concave function of λ_j (Appendix C), and the solution $\lambda_j^{*(q+1)}$ is then discretized to obtain a binary solution $\lambda_j^{(q+1)}$ as follows: $\lambda_j^{(q+1)} = \operatorname{sgn}(\lambda_j^{*(q+1)})$, where sgn denotes the sign function. This relaxed approach is not used in the experiments of this paper but see Jacques [2005] for some examples of use.
Remark: Here, the estimation of (p_k, α_{kj}) and θ is performed in a sequential fashion. This procedure enjoys the advantage to be algorithmically straightforward unlike the procedure involving the full likelihood function of all parameters. Moreover, experiments below indicate good behaviour of this strategy. However, since full likelihood estimates are expected to have less bias, this new way could be explored in future works (see Section 7).

5 Robustness study to the Gaussian assumption

In this section, we first prove that the Gaussian hypothesis can be weakened into a new assumption and, then, some experiments illustrate and evaluate the robustness of the methodology when this weakened assumption is itself violated.

5.1 Theoretical robustness study

Under the hypothesis that the link between the latent variables of both populations P and $\tilde{P}$ is linear, the assumption of normality of the latent variables $Y_{1j_{|Z_{1k}=1}}$ and $\tilde{Y}_{1j_{|\tilde{Z}_{1k}=1}}$ is in fact not necessary, and sufficient conditions are the following:

$$F_{Y_{1j_{|Z_{1k}=1}}}(y) = \Phi(\frac{y-\mu_{kj}}{\sigma_{kj}}) \qquad \text{for } y = s_j \text{ and } y = \frac{s_j - b_{kj}}{a_{kj}}, \tag{9}$$

where $F_{Y_{1j_{|Z_{1k}=1}}}$ denotes the conditional cumulative density function of variable $Y_{1j_{|Z_{1k}=1}}$ and μ_{kj}, σ_{kj} its mean and standard deviation respectively. Note the difference between initial assumption in Subsection 3.1: This equality was stated for all y values in $\mathbb{R}$, whereas it is now stated for only two specific values of y.
However it is straightforward to verify that Equations (5) and (11) remain valid in this context. Since they are the key for establishing Relationship (7), this latter is still right.

5.2 Practical robustness study

To illustrate the robustness against Condition (9) itself described in the previous subsection, consider the following example: For all $1 \leq j \leq 5$ and $1 \leq k \leq 2$, $X_{1j_{|Z_{1k}=1}}$ (respectively $\tilde{X}_{1j_{|\tilde{Z}_{1k}=1}}$) is the binary discretization (with the threshold $s_j = s$) of $Y_{1j_{|Z_{1k}=1}}$ (resp. $\tilde{Y}_{1j_{|\tilde{Z}_{1k}=1}}$), whose distribution is a mixture of two Gaussians.
The mixture density distribution of $Y_{1j_{|Z_{1k}=1}}$ and $\tilde{Y}_{1j_{|\tilde{Z}_{1k}=1}}$ (for fixed j and k), and the Gaussian corresponding density (with the same moments as $Y_{1j_{|Z_{1k}=1}}$ and $\tilde{Y}_{1j_{|\tilde{Z}_{1k}=1}}$) are given in (Figure 1).

The values of the discretization threshold, of the transformation parameters between P and $\tilde{P}$ (a_{kj} and b_{kj}) and of the first and second moments of $Y_{1j_{|Z_{1k}=1}}$ (μ_{kj}, σ^2_{kj}) are arbitrary chosen: $s = 13$; $a_{kj} = 1.2$ and $b_{kj} = 1$ for $1 \leq k \leq 2$ and $1 \leq j \leq 5$; $(\mu_{1j}, \mu_{2j}) = (10, 13)$ and $(\sigma_{1j}, \sigma_{2j}) = (1.8, 1.7)$ for $1 \leq j \leq 5$.
Now, let choose (by numerical optimization) the parameters of the mixture density of $Y_{1j_{|Z^k=1}}$ with respect to the following constraints, for $0 \leq \epsilon \leq 1$:

$$\begin{cases} F_{Y_{1j_{|Z_{1k}=1}}}(s_j) = \Phi(\frac{s_j-\mu_{kj}}{\sigma_{kj}}) + (1 - \Phi(\frac{s_j-\mu_{kj}}{\sigma_{kj}})) \times \epsilon \\ F_{Y_{1j_{|Z_{1k}=1}}}(\frac{s_j-b_{kj}}{a_{kj}}) = \Phi(\frac{s_j-a_{kj}\mu_{kj}-b_{kj}}{a_{kj}\sigma_{kj}}) + (1 - \Phi(\frac{s_j-a_{kj}\mu_{kj}-b_{kj}}{a_{kj}\sigma_{kj}})) \times \epsilon. \end{cases} \tag{10}$$

When $\epsilon = 0$, these constraints correspond to (9) and then satisfy the assumptions of Generalized Discriminant Analysis (GDA), and the greater is ϵ, the less (9) is respected. (Figure 2) illustrates the cumulative density function of the latent variables for different values of ϵ.

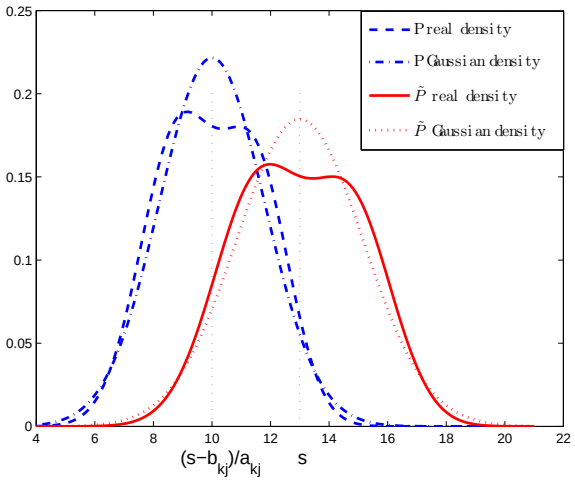
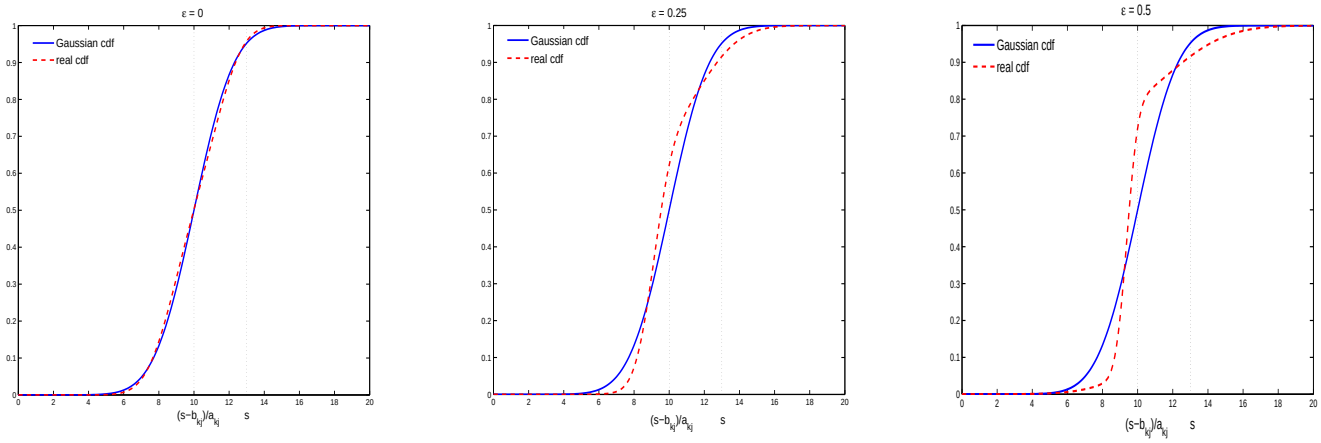


Figure 1: Density distribution of $Y_{1j_{|Z_{1k}=1}}$ and $\tilde{Y}_{1j_{|\tilde{Z}_{1k}=1}}$ for the robustness study.



$\epsilon = 0$ $\epsilon = 0.25$ $\epsilon = 0.5$

Figure 2: Cumulative density function of the latent variables for different values of ϵ.

The sample size is set to 50, the experiments are repeated 20 times and the mean error rate, estimated on "out-of-sample" data, is presented on (Figure 3) for different strategies: GDA with model $[\tilde{p}_k\ \delta_k\ \gamma_k]$, standard discriminant analysis and clustering. Moreover, the optimal error rate is also given. GDA outperforms usual methods when ϵ is moderate ($\epsilon < 0.22$), and for higher ϵ, the assumed model of GDA is too incorrect and then clustering becomes better.

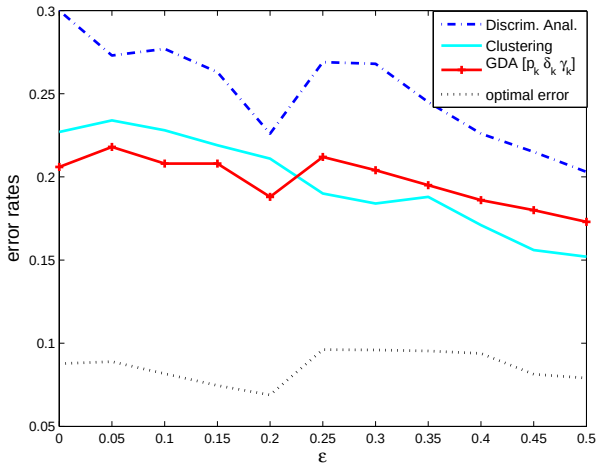


Figure 3: Classification error rate for usual discriminant analysis, generalized discriminant analysis (GDA) and clustering compared to the optimal error rate for different values of ϵ.

6 Comparison of methods on biological data

6.1 Discretized continuous data

The first motivations for which GDA was developed are biological applications (Biernacki et al. [2002], Van Franeker and Ter Brack [1993]), in which the aim was to predict sex of birds from biometrical variables. Very powerful results have been obtained with multinormal assumptions.
The species of birds considered in the present application is Cory's Shearwater *Calanectris diomedea* (Thibault et al. [1997]). Two subspecies can be identified: *borealis* which lives in the Atlantic islands (the Azores, Canaries, *etc*.) and *diomedea* which lives in the Mediterranean islands (Balearics, Corsica, *etc*.).
A sample of *Borealis* ($n = 206$, 45% females) was measured using skins in several National Museums. Five morphological variables were measured: Culmen (bill length), tarsus, wings and tail lengths, and culmen depth. Similarly, a sample of *diomedea* ($n = 38$, 58% females) was measured using the same set of variables. In this example, two groups are present, males and females, and all the birds are of known sex (from dissection). (Figure 4) illustrates differences between the two subspecies *borealis* and *diomedea*, for two biometrical variables.

To provide an application of the present work, the continuous biometrical variables are discretized into binary data. As it can be shown on (Figure 4), discretization must be carried out carefully, especially concerning the choice of the discretization threshold. Indeed, if this threshold equals the mean of the biometrical variables for one subspecies for instance, then all the values for the other subspecies will be on the same side of this discretization threshold. Consequently, conditionally to each biometrical variables, the threshold is chosen such that there is (roughly) a maximum of individuals of each population on each side of the discretization threshold.
The subspecies *borealis* is selected as the training population and the subspecies *diomedea* as the

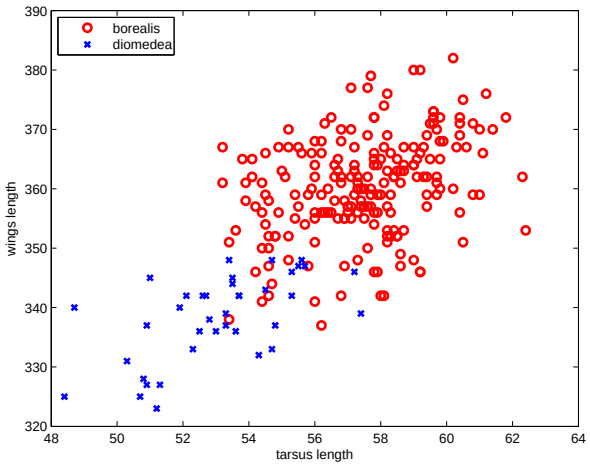


Figure 4: Wings and tarsus lengths for *diomedea* and *borealis*.

test population. The 32 GDA models, including standard discriminant analysis $[p_k\, 1\, 0]$, and clustering are applied on these data. The classification error rate and the value of the BIC criterion are given in Table 2.

Table 2: Classification error rates (%) and value of the BIC criterion for test population *diomedea* with training population *borealis*.

model	$[p_k\, 1\, 0]$	$[p_k\, 1\, \gamma]$	$[p_k\, 1\, \gamma_k]$	$[p_k\, 1\, \gamma_j]$	$[p_k\, \delta\, 0]$	$[p_k\, \delta\, \gamma]$	$[p_k\, \delta\, \gamma_k]$	$[p_k\, \delta\, \gamma_j]$
error	42.1	23.68	15.78	18.42	57.89	23.68	15.78	18.42
BIC	648	216	218	225	263	214	218	214
model	$[p_k\, \delta_k\, 0]$	$[p_k\, \delta_k\, \gamma]$	$[p_k\, \delta_k\, \gamma_k]$	$[p_k\, \delta_k\, \gamma_j]$	$[p_k\, \delta_j\, 0]$	$[p_k\, \delta_j\, \gamma]$	$[p_k\, \delta_j\, \gamma_k]$	$[p_k\, \delta_j\, \gamma_j]$
error	57.89	15.78	18.42	18.42	57.89	18.42	18.42	**15.78**
BIC	270	1219	216	220	281	214	220	228
model	$[\tilde{p}_k\, 1\, 0]$	$[\tilde{p}_k\, 1\, \gamma]$	$[\tilde{p}_k\, 1\, \gamma_k]$	$[\tilde{p}_k\, 1\, \gamma_j]$	$[\tilde{p}_k\, \delta\, 0]$	$[\tilde{p}_k\, \delta\, \gamma]$	$[\tilde{p}_k\, \delta\, \gamma_k]$	$[\tilde{p}_k\, \delta\, \gamma_j]$
error	42.1	26.31	23.68	21.05	42.1	21.05	23.68	21.05
BIC	595	215	215	226	267	**213**	215	215
model	$[\tilde{p}_k\, \delta_k\, 0]$	$[\tilde{p}_k\, \delta_k\, \gamma]$	$[\tilde{p}_k\, \delta_k\, \gamma_k]$	$[\tilde{p}_k\, \delta_k\, \gamma_j]$	$[\tilde{p}_k\, \delta_j\, 0]$	$[\tilde{p}_k\, \delta_j\, \gamma]$	$[\tilde{p}_k\, \delta_j\, \gamma_k]$	$[\tilde{p}_k\, \delta_j\, \gamma_j]$
error	42.1	23.68	21.05	21.05	42.1	21.05	21.05	23.68
BIC	274	217	217	222	285	215	222	225

If the results are compared according to the error rate, GDA with the model $[p_k\, \delta_j\, \gamma_j]$ is the best method, with error 15.78%. This error is lower than the one obtained by standard discriminant analysis (42.1%) or by clustering (23.68%). By using the BIC criterion, which leads to select the model $[\tilde{p}_k\, \delta\, \gamma]$, the error rate ($21.05\%$) is still better than usual discriminating method.
This application illustrates the interest of GDA with respect to standard discriminant analysis or

clustering. Indeed, by adapting the classification rule derived from the training population to the test population, GDA gives lower classification error rates than by applying directly the rule derived from the training population (standard discriminant analysis), or by omitting the training population and applying directly clustering on the test population.
It is worth pointing out that the assumption for binary data to be derived from the discretization of Gaussian variables (biometrical variables) is relatively realistic in this application. Nevertheless, there exists a strong correlation between the five biometrical variables, which violates the assumption that discretized variables are independent.
Remark Although the "test sample" is used for estimating the discriminant rule through the *unlabelled* data of $\tilde{S}$ and is used also for estimating the error rate but this once through the *labelled* data of $\tilde{S}$, this estimated error rate is usually not an optimistic measure of the classification method's performance. This fact is well-known in the semi-supervised setting (**?**, Subsection 21.1.2). In addition, we have also verified this claim with our biological data by performing a hold-out procedure (results not reported here).

6.2 Real binary data

The second application considers sea birds from the species *puffins* (**?**). Two groups of subspecies are considered: The first one is composed of subspecies living in Pacific Islands – *subalaris* (Galapagos Island), *polynesial*, *dichrous* (Enderbury and Palau Islands) and *gunax* – and the second one is composed of subspecies living in Atlantic Islands – *boydi* (Cap Verde Islands). Here, the difference between populations is the geographical range (Pacific *vs.* Atlantic Islands). A sample of Pacific birds ($n = 171$) was measured using skins in several National Museums. Four variables are measured on these birds: Coller, stripe and piping (absence or presence for these three variables) and under-caudal (self couloured or not). Similarly, a sample of Atlantic birds ($n = 19$) was measured using the same set of variables. Like in the previous example, two groups are present (males and females) and the sex of all the birds is known. Pacific birds are chosen as the training population and Atlantic ones as the test population. Choosing Atlantic birds as the test population corresponds to a realistic situation because it could be hazardous to perform a clustering process on a sample of such a small size. This is a typical situation where our methodology could be expected to provide a parsimonious and meaningful alternative. According to the biologist Vincent Bretagnolle, the morphological variables which are used in this application are not very discriminative, and then one can not expect that the error rate will be better than $40 - 45\%$.
The 32 GDA models, among which standard discriminant analysis $[p_k\, 1\, 0]$, are applied on these data and the results are presented in Table 3. Clustering is also applied, and the obtained error rate is 49.05%.

As in the previous study, GDA is more efficient than standard discriminant analysis (50.94%) and clustering (49.05%) to classify birds according to their sex. Moreover the BIC criterion leads to choose the model with the smallest error rate. The relatively poor classification results (the minimal error rate is 43%) confirm the assumption of the biologist.

Table 3: Classification error rates (%) and value of the BIC criterion for test population of Atlantic birds with training on Pacific birds population.

model	$[p_k\, 1\, 0]$	$[p_k\, 1\, \gamma]$	$[p_k\, 1\, \gamma_k]$	$[p_k\, 1\, \gamma_j]$	$[p_k\, \delta\, 0]$	$[p_k\, \delta\, \gamma]$	$[p_k\, \delta\, \gamma_k]$	$[p_k\, \delta\, \gamma_j]$
error	50.94	**43.39**	45.28	43.39	50.94	43.39	45.28	45.28
BIC	212	**209**	216	224	212	209	216	224
model	$[p_k\, \delta_k\, 0]$	$[p_k\, \delta_k\, \gamma]$	$[p_k\, \delta_k\, \gamma_k]$	$[p_k\, \delta_k\, \gamma_j]$	$[p_k\, \delta_j\, 0]$	$[p_k\, \delta_j\, \gamma]$	$[p_k\, \delta_j\, \gamma_k]$	$[p_k\, \delta_j\, \gamma_j]$
error	45.28	45.28	52.83	45.28	45.28	52.83	50.94	50.94
BIC	210	210	215	226	225	224	227	239
model	$[\tilde{p}_k\, 1\, 0]$	$[\tilde{p}_k\, 1\, \gamma]$	$[\tilde{p}_k\, 1\, \gamma_k]$	$[\tilde{p}_k\, 1\, \gamma_j]$	$[\tilde{p}_k\, \delta\, 0]$	$[\tilde{p}_k\, \delta\, \gamma]$	$[\tilde{p}_k\, \delta\, \gamma_k]$	$[\tilde{p}_k\, \delta\, \gamma_j]$
error	45.28	50.94	50.94	45.28	45.28	50.94	50.94	45.28
BIC	213	213	220	228	213	213	220	228
model	$[\tilde{p}_k\, \delta_k\, 0]$	$[\tilde{p}_k\, \delta_k\, \gamma]$	$[\tilde{p}_k\, \delta_k\, \gamma_k]$	$[\tilde{p}_k\, \delta_k\, \gamma_j]$	$[\tilde{p}_k\, \delta_j\, 0]$	$[\tilde{p}_k\, \delta_j\, \gamma]$	$[\tilde{p}_k\, \delta_j\, \gamma_k]$	$[\tilde{p}_k\, \delta_j\, \gamma_j]$
error	45.28	45.28	47.16	45.28	45.28	52.83	45.28	52.83
BIC	214	213	213	229	228	227	224	243

7 Conclusion

Generalized discriminant analysis extends standard discriminant analysis by allowing training and test samples to arise from different but stochastically linked populations. Our contribution consists in extending previous original work, derived in a multinormal context, to the case of binary data.
Applications to a biological problem illustrate the power of our methodology. A classification of birds according to their sex is provided by using generalized discriminant analysis, and this classification is better than those obtained by standard discriminant analysis or by clustering.
Perspectives for this work are numerous. Firstly, the parameters of both populations P and $\tilde{P}$ are estimated successively: α_{kj} and p_k are estimated in a first step and then $\tilde{\alpha}_{kj}$ and $\tilde{p}_k$ are deduced from these estimations and from those of all parameters. It should be very (computationally) useful to consider a joint estimation of this four parameters. In particular, **?** extended the earlier work of Biernacki et al. [2002] to the full likelihood estimation of parameters in the multinormal situation. It appears that error rates obtained by sequential estimate and by joint estimate are quite similar when the learning sample is large (situation of the current paper). But, when the learning sample has a small sample size, joint estimation could significantly improve the error rate. We could expect the same behaviour for our current model but this assumption needs to be confirmed by a future specific study.
Secondly, the link between both populations was defined by using Gaussian cumulative density function. Although it seemed initially difficult to find this link, a simple link involving the probit function was obtained. It was not easy to imagine it, but it is meaningful afterwards. It would be interesting to try other types of cumulative density functions; Obviously theoretical reasons will have to be developed and practical tests will have to be carried out.
Thirdly, with this contribution generalized discriminant analysis is now developed for continuous data and for binary data. To allow to analyse a large number of practical cases, it is important to study the case of categorical variables (*i.e.* more than two modalities), and thereafter the case of mixed variables (binary, categorical and continuous together). Everitt's works (Everitt [1988]), which defined a classification algorithm for mixed variables, can be helpful for this topics.

Finally, it would be also interesting to extend other classical discriminant method like non-parametric discrimination or semi-parametric discrimination. See Biernacki and Beninel [2005] for logistic regression.

Acknowledgements

Authors thank Vincent Bretagnolle of CEBC-CNRS for providing the biological data.

References

J. A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59:19–35, 1972. ISSN 0006-3444.

C. Biernacki and F. Beninel. Apprentissage sur une sous population et prédiction sur une autre : une extension à (et de) la discrimination logistique. In *Colloque Data Mining et Apprentissage Statistique*, 2005.

C. Biernacki, F. Beninel, and V. Bretagnolle. A generalized discriminant rule when training population and test population differ on their descriptive parameters. *Biometrics*, 58(2):387–397, 2002. ISSN 0006-341X.

G. Celeux and G. Govaert. Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8:157–176, 1991.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977. ISSN 0035-9246. With discussion.

B. S. Everitt. *An introduction to latent variable models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1984. ISBN 0-412-25310-0.

B. S. Everitt. A finite mixture model for the clustering of mixed-mode data. *Statist. Probab. Lett.*, 6(5):305–309, 1988. ISSN 0167-7152.

R.A. Fisher. he use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7: 179–188, 1936.

A. Fix and J.L. Hodges. Discriminatory analysis - non parametric discrimination: Consistency properties. Technical report, U.S.A.F. School of Aviation Medicine, 1951.

J.H. Friedman and W. Stuetzle. Projection pursuit regression. *J. Amer. Statist. Assoc.*, 76(376): 817–823, 1981. ISSN 0162-1459.

D.J. Hand. *Kernel discriminant analysis*, volume 2 of *Electronic & Electrical Engineering Research Studies: Pattern Recognition & Image Processing Series*. Research Studies Press [John Wiley & Sons], Chichester, 1982. ISBN 0-471-10211-3.

J Jacques. *Contributions à l'analyse de sensibilité et à l'analyse discriminante generalisée*. PhD thesis, University Joseph Fourier, 2005.

G.J. McLachlan. *Discriminant analysis and statistical pattern recognition*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1992. ISBN 0-471-61531-5. , A Wiley-Interscience Publication.

G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978. ISSN 0090-5364.

B. W. Silverman. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986. ISBN 0-412-24620-1.

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

L.L. Thurstone. A law of comparative judgement. *American Journal of Psychology*, 38:368–389, 1927.

R. Tomassone, M. Danzard, J.J. Daudin, and J.P. Masson. *Discrimination et classement*. Masson, 1988.

J.A. Van Franeker and C.J.F. Ter Brack. A generalized discriminant for sexing fulmarine petrels from external measurements. *The Auk*, 110(3):492–502, 1993.

L.A. Wolsey. *Integer programming*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons Inc., New York, 1998. ISBN 0-471-28366-5. , A Wiley-Interscience Publication.

A Proof of the relation between test and training populations

From (6) it follows that $\tilde{Y}_{1j_{|\tilde{Z}_{1k}=1}}$ is Gaussian with mean $\tilde{\mu}_{kj} = a_{kj}\mu_{kj} + b_{kj}$ and with standard deviation $\tilde{\sigma}_{kj} = |a_{kj}|\sigma_{kj}$. However, this transformation is clearly non-identifiable: There exists more than one couple $(A_k, \mathbf{b}_k)$ which satisfies Relationship (6). To solve this problem, all the a_{kj} are assumed to be non negative ($a_{kj} \geq 0$).
It is possible to derived from Equation (5):

$$\lambda_j \frac{s_j}{\sigma_{kj}} = -\Phi^{-1}(\alpha_{kj}) + \lambda_j \frac{\mu_{kj}}{\sigma_{kj}},$$

and then, using expressions of $\tilde{\mu}_{kj}$ and $\tilde{\sigma}_{kj}$ just given above:

$$\begin{aligned} \tilde{\alpha}_{kj} &= 1 - \Phi\left(\lambda_j \frac{s_j - \tilde{\mu}_{kj}}{\tilde{\sigma}_{kj}}\right) = 1 - \Phi\left(\frac{1}{a_{kj}}\frac{\lambda_j s_j}{\sigma_{kj}} - \lambda_j \frac{a_{kj}\mu_{kj} + b_{kj}}{a_{kj}\sigma_{kj}}\right) \\ &= \Phi\left(\frac{\Phi^{-1}(\alpha_{kj})}{a_{kj}} + \lambda_j \frac{a_{kj}\mu_{kj} + b_{kj} - \mu_{kj}}{a_{kj}\sigma_{kj}}\right). \end{aligned} \tag{11}$$

Alternatively, it is also possible to write from (5) again:

$$\lambda_j \frac{\mu_{kj}}{\sigma_{kj}} = \Phi^{-1}(\alpha_{kj}) + \lambda_j \frac{s_j}{\sigma_{kj}},$$

and thus, following the same process as in (11),

$$\begin{aligned} \tilde{\alpha}_{kj} &= 1 - \Phi\left(\lambda_j \frac{s_j - \tilde{\mu}_{kj}}{\tilde{\sigma}_{kj}}\right) = 1 - \Phi\left(-\lambda_j \frac{\mu_{kj}}{\sigma_{kj}} + \lambda_j \frac{s_j - b_{kj}}{a_{kj}\sigma_{kj}}\right) \\ &= \Phi\left(\Phi^{-1}(\alpha_{kj}) + \lambda_j \frac{a_{kj}s_j + b_{kj} - s_j}{a_{kj}\sigma_{kj}}\right). \end{aligned} \tag{12}$$

Of course, Equations (11) and (12) are equivalent but the first one will be retained since Parametrization (11) will be more convenient for proposing later numerous parsimonious models of constraints on the link between P and $\tilde{P}$. Consequently, the following relationship between α_{kj} and $\tilde{\alpha}_{kj}$ is obtained:

$$\tilde{\alpha}_{kj} = \Phi\Big(\delta_{kj}\ \Phi^{-1}(\alpha_{kj}) + \lambda_j \gamma_{kj}\Big),$$

where $\delta_{kj} \in \mathbb{R}^+ \setminus \{0\}$, $\lambda_j \in \{-1, 1\}$ and $\gamma_{kj} \in \mathbb{R}$.

B Model identifiability

B.1 Intra-group identifiability

Firstly, the identifiability of the couple (λ_j, γ_{kj}) in different constrained model situations is reaching as follows:

- Models involving $\gamma_{kj} = 0$ or $\gamma_{kj} = \gamma_j$: By setting for instance $\lambda_j = +1$ ($j = 1, \ldots, d$),
- Models involving $\gamma_{kj} = \gamma$ or $\gamma_{kj} = \gamma_k$: By setting for instance $\lambda_1 = +1$.

By this way, the product $\lambda_j \gamma_{kj}$ is allways identifiable and obviously all these constraints on λ_j have no impact on the estimation of the product $\lambda_j \gamma_{kj}$, as the reader can easily convince himself.
Secondly, the identifiability of the couple of parameters $(\delta_{kj}, \gamma_{kj})$ conditionally to λ_j is discussed. Equation (7) leads to

$$\Phi^{-1}(\tilde{\alpha}_{kj}) = \delta_{kj}\ \Phi^{-1}(\alpha_{kj}) + \lambda_j \gamma_{kj} \tag{13}$$

which can be expressed as the following linear system

$$\tilde{\Phi} = \Phi_\lambda \times \mathbf{u}_{\delta,\gamma}$$

where $\tilde{\Phi} = (\Phi^{-1}(\tilde{\alpha}_{11}), \ldots, \Phi^{-1}(\tilde{\alpha}_{kj}), \ldots, \Phi^{-1}(\tilde{\alpha}_{Kd}))^T \in \mathbb{R}^{Kd}$ and Φ_λ and $\mathbf{u}_{\delta,\gamma}$ are respectively a matrix and a vector, with dimension depending on the model at hand, representing the values of $\Phi^{-1}(\alpha_{kj})$ and λ_j for Φ_λ, and the values of δ_{kj} and γ_{kj} for $\mathbf{u}_{\delta,\gamma}$.
Identifiability is obtained if and only if the matrix Φ_λ is of full rank. It is easily noticed (see the

example below) that Φ_λ is not of full rank only for very particular values of α_{kj} (typically $\alpha_{kj} = \frac{1}{2}$ for some k, j for instance). Moreover these theoretical non-identifiable situations lead seldom to practical non-identifiable situations since an estimator $\hat{\alpha}_{kj}$ is used instead of the unknown true value α_{kj}. To be definitively convinced of this fact, the reader can take a look at the robustness study (Section 5) where simulations involve a theoretically non-identifiable model (all $\alpha_{1j} = \frac{1}{2}$) but where practical identifiability is observed.

Example with model $[p_k\, \delta_k\, \gamma_k]$ for intra-group identifiability In this situation, $\Phi_\lambda = [M|N]$ is a $(Kd \times 2K)$-matrix formed by two block matrices $M = (M_{lk})_{\substack{1\le k\le K\\ 1\le l\le Kd}}$ and $N = (N_{lk})_{\substack{1\le k\le K\\ 1\le l\le Kd}}$ defined by

$$\Phi_\lambda = \left(\begin{array}{cccccccc|cccccccc}
\Phi^{-1}(\alpha_{11}) & 0 & & \dots & & & 0 & & \lambda_1 & 0 & & \dots & & & 0 & \\
 & & & \vdots & & & & & & & & \vdots & & & & \\
\Phi^{-1}(\alpha_{1d}) & 0 & & \dots & & & 0 & & \lambda_d & 0 & & \dots & & & 0 & \\
 & & & \vdots & & & & & & & & \vdots & & & & \\
0 & \dots & 0 & \Phi^{-1}(\alpha_{k1}) & 0 & \dots & & 0 & 0 & \dots & 0 & \lambda_1 & 0 & \dots & 0 & \\
 & & & \vdots & & & & & & & & \vdots & & & & \\
0 & \dots & 0 & \Phi^{-1}(\alpha_{kd}) & 0 & \dots & & 0 & 0 & \dots & 0 & \lambda_d & 0 & \dots & 0 & \\
 & & & \vdots & & & & & & & & \vdots & & & & \\
0 & & & \dots & & & 0 & \Phi^{-1}(\alpha_{K1}) & 0 & & & \dots & & & 0 & \lambda_1 \\
 & & & \vdots & & & & & & & & \vdots & & & & \\
0 & & & \dots & & & 0 & \Phi^{-1}(\alpha_{Kd}) & 0 & & & \dots & & & 0 & \lambda_d
\end{array}\right).$$

Φ_λ is not of full rank ($\min(Kd, 2K)$) if and only if

- Possibility 1: There exists $k \in \{1, \dots, K\}$ such that $\alpha_{kj} = \frac{1}{2}$ for all $j \in \{1, \dots, d\}$,
- Possibility 2: There exists $k_1, \dots, k_\tau \in \{1, \dots, K\}$, $k_1, \dots, k_\tau$ being all different, with $\tau \in \{1, \dots, K\}$ satisfying $d(K - \tau) < 2K$, such that for all $j, j' \in \{1, \dots, d\}$, for all $k \in \{k_1, \dots, k_\tau\}$, $\lambda_{j'}\Phi^{-1}(\alpha_{kj}) = \lambda_j\Phi^{-1}(\alpha_{kj'})$.

B.2 Inter-group identifiability

This non-identifiability problem means that one group of the population P can be transformed into more than one group of the population $\tilde{P}$. It cannot happen if the group proportions p_k in P are all different (it is likely in practice) and if simultaneously all the constrained models are with fixed proportions between populations (models $[p_k\,.\,]$). But if these proportions conditions are not verified, the non-identifiability problem is the following: For fixed k_1 and k_2 ($k_1, k_2 \in \{1, ..., K\}$, $k_1 \neq k_2$) and for fixed j ($j \in \{1, \dots, d\}$) there exists two sets of parameters $(\delta_{k_1 j}, \lambda_j, \gamma_{k_1 j}) \neq (\delta'_{k_1 j}, \lambda'_j, \gamma'_{k_1 j})$ which transform respectively $\alpha_{k_1 j}$ into $\tilde{\alpha}_{k_1 j}$ and into $\tilde{\alpha}_{k_2 j}$. In fact, if the group k_1 of P is transformed into the group k_2 of $\tilde{P}$ (instead of the group k_1 of $\tilde{P}$) then necessarily the group k_2 of P is not transformed into the group k_2 of $\tilde{P}$, but into a group $k_3 \neq k_2$ of $\tilde{P}$; The simplest solution

is $k_3 = k_1$ but it is not certain if $K > 2$.
Thus, the identifiability problem can be rewritten equivalently: There exists $k_2 \neq k_1$ and $k_3 \neq k_2$ such that, for any j and $(\delta_{k_1 j}, \lambda_j, \gamma_{k_1 j})$, $(\delta_{k_2 j}, \lambda_j, \gamma_{k_2 j})$ such that

$$\Phi^{-1}(\tilde{\alpha}_{k_1 j}) = \delta_{k_1 j}\ \Phi^{-1}(\alpha_{k_1 j}) + \lambda_j \gamma_{k_1 j} \tag{14}$$
$$\Phi^{-1}(\tilde{\alpha}_{k_2 j}) = \delta_{k_2 j}\ \Phi^{-1}(\alpha_{k_2 j}) + \lambda_j \gamma_{k_2 j}, \tag{15}$$

there exists $(\delta'_{k_1 j}, \lambda'_j, \gamma'_{k_1 j}) \neq (\delta_{k_1 j}, \lambda_j, \gamma_{k_1 j})$ and $(\delta'_{k_2 j}, \lambda'_j, \gamma'_{k_2 j}) \neq (\delta_{k_2 j}, \lambda_j, \gamma_{k_2 j})$ such that

$$\Phi^{-1}(\tilde{\alpha}_{k_2 j}) = \delta'_{k_1 j}\ \Phi^{-1}(\alpha_{k_1 j}) + \lambda'_j \gamma'_{k_1 j} \tag{16}$$
$$\Phi^{-1}(\tilde{\alpha}_{k_3 j}) = \delta'_{k_2 j}\ \Phi^{-1}(\alpha_{k_2 j}) + \lambda'_j \gamma'_{k_2 j}. \tag{17}$$

It follows from (14) and (16)

$$\Phi^{-1}(\tilde{\alpha}_{k_2 j}) = \frac{\delta'_{k_1 j}}{\delta_{k_1 j}} \Phi^{-1}(\tilde{\alpha}_{k_1 j}) - \lambda_j \frac{\gamma_{k_1 j}}{\delta_{k_1 j}} \delta'_{k_1 j} + \lambda'_j \gamma'_{k_1 j} \tag{18}$$

and similarly from (15) and (17)

$$\Phi^{-1}(\tilde{\alpha}_{k_3 j}) = \frac{\delta'_{k_2 j}}{\delta_{k_2 j}} \Phi^{-1}(\tilde{\alpha}_{k_2 j}) - \lambda_j \frac{\gamma_{k_2 j}}{\delta_{k_2 j}} \delta'_{k_2 j} + \lambda'_j \gamma'_{k_2 j}. \tag{19}$$

Equations (18) and (19) lead to the following linear system

$$\tilde{\Phi}^{(k_2,k_3)} = \Phi^{(k_1,k_2)}_{\delta,\gamma,\lambda,\lambda'} \times \mathbf{u}^{(k_1,k_2)}_{\delta',\gamma'} \tag{20}$$

where $\tilde{\Phi}^{(k_2,k_3)} = (\Phi^{-1}(\tilde{\alpha}_{k_2 1}), \ldots, \Phi^{-1}(\tilde{\alpha}_{k_2 d}), \Phi^{-1}(\tilde{\alpha}_{k_3 1}), \ldots, \Phi^{-1}(\tilde{\alpha}_{k_3 d}))' \in \mathbb{R}^{2d}$ and $\Phi^{(k_1,k_2)}_{\delta,\gamma,\lambda,\lambda'}$ and $\mathbf{u}^{(k_1,k_2)}_{\delta',\gamma'}$ are respectively a matrix and a vector, with dimension depending on the model at hand, representing values of $\Phi^{-1}(\tilde{\alpha}_{k_1 j})$, $\Phi^{-1}(\tilde{\alpha}_{k_2 j})$, $\delta_{k_1 j}$, $\delta_{k_2 j}$, $\gamma_{k_1 j}$, $\gamma_{k_2 j}$, λ_j, λ'_j for $\Phi^{(k_1,k_2)}_{\delta,\gamma,\lambda,\lambda'}$ and values of $\delta'_{k_1 j}$, $\delta'_{k_2 j}$, $\gamma'_{k_1 j}$, $\gamma'_{k_2 j}$ for $u^{(k_1,k_2)}_{\delta',\gamma'}$.
Conditionally to the values of λ'_j, the problem is identifiable if no solution exists to System (20) or, in other words, if the number of linearly independent lines of $\Phi^{(k_1,k_2)}_{\delta,\gamma,\lambda,\lambda'}$ is no less than the number of free parameters in $\mathbf{u}^{(k_1,k_2)}_{\delta',\gamma'}$. The number of lines of $\Phi^{(k_1,k_2)}_{\delta,\gamma,\lambda,\lambda'}$ is equal to $2d$ and in most cases of interest these lines are independent for any values λ'_j. See for instance the following example with the model $[\tilde{p}_k\, \delta_k\, \gamma_k]$ for a discussion on this subject. The number of free parameters of $\mathbf{u}^{(k_1,k_2)}_{\delta',\gamma'}$ corresponds to the values in Table 1 by artificially fixing $K = 2$ in this table. Except for model $[\tilde{p}_k\, \delta_j\, \gamma_j]$ the maximum number of free parameters is equal to $\max(4, d+2)$. Thus, a sufficient condition of inter-group identifiability is $2d > \max(4, d+2)$. Since $d > 2$ usually for binary data, all models are identifiable except the model $[\tilde{p}_k\, \delta_j\, \gamma_j]$ since the number of free parameters is $2d$. Although this model is non-identifiable, it can nevertheless be used by bearing in mind that a label switching may occur from P to $\tilde{P}$.
Remark Only a single switch between 2 classes ((k_1, k_2) of P becomes (k_2, k_1) of $\tilde{P}$) may occur with the model $[\tilde{p}_k\, \delta_j\, \gamma_j]$. Indeed in case of a switch between more than 2 classes, for instances (k_1, k_2, k_3) of P becomes (k_2, k_3, k_1) of $\tilde{P}$ or (k_1, k_2, k_3, k_4) of P becomes (k_2, k_1, k_4, k_3) of $\tilde{P}$, the number of equations reaches at least $3d$ for $2d$ free parameters: The model $[\tilde{p}_k\, \delta_j\, \gamma_j]$ is then identifiable.

Example with model $[\tilde{p}_k\ \delta_k\ \gamma_k]$ for inter-group identifiability. In this situation $\Phi^{(k_1,k_2)}_{\delta,\gamma,\lambda,\lambda'} = [M|N]$ is a $(2d \times 4)$-matrix formed by two matrix blocks $M = (M_{ls})_{\substack{1\leq s\leq 2\\ 1\leq l\leq 2d}}$ and $N = (N_{ls})_{\substack{1\leq s\leq 2\\ 1\leq l\leq 2d}}$ defined by

$$\Phi^{(k_1,k_2)}_{\delta,\gamma,\lambda,\lambda'} = \left(\begin{array}{cc|cc} \frac{1}{\delta_{k_1 1}}\Phi^{-1}(\tilde{\alpha}_{k_1 1}) - \lambda_1\frac{\gamma_{k_1 1}}{\delta_{k_1 1}} & 0 & \lambda'_1 & 0 \\ \vdots & & \vdots & \\ \frac{1}{\delta_{k_1 d}}\Phi^{-1}(\tilde{\alpha}_{k_1 d}) - \lambda_d\frac{\gamma_{k_1 d}}{\delta_{k_1 d}} & 0 & \lambda'_d & 0 \\ 0 & \frac{1}{\delta_{k_2 1}}\Phi^{-1}(\tilde{\alpha}_{k_2 1}) - \lambda_1\frac{\gamma_{k_2 1}}{\delta_{k_2 1}} & 0 & \lambda'_1 \\ & \vdots & & \vdots \\ 0 & \frac{1}{\delta_{k_2 d}}\Phi^{-1}(\tilde{\alpha}_{k_2 d}) - \lambda_d\frac{\gamma_{k_2 d}}{\delta_{k_2 d}} & 0 & \lambda'_d \end{array}\right)$$

The reader can easily convince himself that the $2d$ lines are generally linearly independent in practice. For instance, the simulation involved in the robustness study (Section 5) considers this model and none identifiability problem is encountered (the classification error rates are always strictly lower than 50% and thus none label switching has been obtained).

C Proof of the concavity of the function $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$

The aim of this appendix is to prove that $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ is a strictly concave function of γ_k, γ_j and γ, and also of δ_k, δ_j, δ and λ_j. The fundamental key of the proof is to study first concavity of $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ for γ_{kj}.
Let $\mathcal{Q}(\gamma_{kj})$ be the function $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ with parameters $\tilde{p}_k, \delta_{kj}, \lambda_j$ and $\boldsymbol{\theta}^{(q)}$ fixed, and let prove that $\mathcal{Q}(\gamma_{kj})$ is strictly concave:

$$\mathcal{Q}(\gamma_{kj}) = \sum_{i=1}^{\tilde{n}}\sum_{k=1}^{K} t_{ik}\Big\{\log(\tilde{p}_k) + \sum_{j=1}^{d}\tilde{x}_{ij}\log\Big(\Phi(\zeta_{kj})\Big) + \sum_{j=1}^{d}(1-\tilde{x}_{ij})\log\Big(1-\Phi(\zeta_{kj})\Big)\Big\}$$

with $\zeta_{kj} = \delta_{kj}\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma_{kj}$.
The derivative of the function $\mathcal{Q}(\gamma_{kj})$ is:

$$\frac{\partial\mathcal{Q}(\gamma_{kj})}{\partial\gamma_{kj}} = \sum_{i=1}^{\tilde{n}}\sum_{k=1}^{K} t_{ik}\sum_{j=1}^{d}\Big\{\tilde{x}_{ij}\lambda_j\frac{\phi(\zeta_{kj})}{\Phi(\zeta_{kj})} + (1-\tilde{x}_{ij})\lambda_j\frac{-\phi(\zeta_{kj})}{1-\Phi(\zeta_{kj})}\Big\},$$

where ϕ is the probability density function of $\mathcal{N}(0,1)$.
Using that $\frac{\partial\phi(bx+a)}{\partial x} = -b(bx+a)\phi(bx+a)$ with $a, b \in \mathbb{R}$, the second derivative is:

$$\begin{aligned}\frac{\partial^2\mathcal{Q}(\gamma_{kj})}{\partial\gamma_{kj}^2} = \sum_{i=1}^{\tilde{n}}\sum_{k=1}^{K} t_{ik}\sum_{j=1}^{d}\lambda_j\Big\{&\tilde{x}_{ij}\frac{-\lambda_j\zeta_{kj}\phi(\zeta_{kj})\Phi(\zeta_{kj}) - \lambda_j\phi(\zeta_{kj})^2}{[\Phi(\zeta_{kj})]^2} \\ &+(1-\tilde{x}_{ij})\frac{\lambda_j\zeta_{kj}\phi(\zeta_{kj})(1-\Phi(\zeta_{kj})) - \lambda_j\phi(\zeta_{kj})^2}{[1-\Phi(\zeta_{kj})]^2}\Big\},\end{aligned}$$

that can be rewritten:

$$\frac{\partial^2 \mathcal{Q}(\gamma_{kj})}{\partial \gamma_{kj}^2} = -\sum_{i=1}^{\tilde{n}}\sum_{k=1}^{K} t_{ik} \sum_{j=1}^{d} \lambda_j^2 \phi(\zeta_{kj}) \Big\{ \frac{\tilde{x}_{ij}}{[\Phi(\zeta_{kj})]^2} (\overbrace{\zeta_{kj}\Phi(\zeta_{kj}) + \phi(\zeta_{kj})}^{g_1(\zeta_{kj})}) + \frac{(1-\tilde{x}_{ij})}{[1-\Phi(\zeta_{kj})]^2} (\overbrace{\zeta_{kj}(\phi(\zeta_{kj}) - 1) + \phi(\zeta_{kj})}^{g_2(\zeta_{kj})}) \Big\}. \quad (21)$$

To prove that $\mathcal{Q}$ is strictly concave, it is sufficient to prove that both functions g_1 and g_2 are strictly positive:

- For all $x \in \mathbb{R} : g_1(x) = x\Phi(x) + \phi(x) > 0$, because $\lim_{x \to -\infty} g_1(x) = 0$ and g_1 is strictly increasing since $g_1'(x) = \Phi(x) + x\phi(x) - x\phi(x) = \Phi(x) > 0$,
- For all $x \in \mathbb{R} : g_2(x) = x\Phi(x) - x + \phi(x) > 0$, because $\lim_{x \to +\infty} g_2(x) = 0$ and g_2 is strictly decreasing since $g_2'(x) = \Phi(x) - 1 + x\phi(x) - x\phi(x) = \Phi(x) - 1 < 0$.

Thus $\frac{\partial^2 \mathcal{Q}(\gamma_{kj})}{\partial \gamma_{kj}^2} < 0$ and $\mathcal{Q}(\gamma_{kj})$ is strictly concave.
If $\mathcal{Q}$ is no longer function of γ_{kj} but now of γ_k (respectively of γ_j, of γ) the expression of the second derivative is the same as (21) by removing the sum on j (resp. on k, on (k, j)), and thus $\mathcal{Q}$ is still strictly concave.
Consider now the function $\mathcal{Q}(\delta_{kj})$ with the above convention ($\tilde{p}_k, \gamma_{kj}, \lambda_j$ and $\boldsymbol{\theta}^{(q)}$ fixed). Same type of calculus leads to:

$$\frac{\partial^2 \mathcal{Q}(\delta_{kj})}{\partial \delta_{kj}^2} = -\sum_{i=1}^{\tilde{n}}\sum_{k=1}^{K} t_{ik} \sum_{j=1}^{d} \left(\Phi^{-1}(\alpha_{kj})\right)^2 \phi(\zeta_{kj}) \Big\{ \frac{\tilde{x}_{ij}}{[\Phi(\zeta_{kj})]^2} g_1(\zeta_{kj}) + \frac{(1-\tilde{x}_{ij})}{[1-\Phi(\zeta_{kj})]^2} g_2(\zeta_{kj}) \Big\},$$

and thus $\mathcal{Q}(\delta_{kj})$ is strictly concave. By using the above arguments, it arises immediately that $\mathcal{Q}(\delta_k)$, $\mathcal{Q}(\delta_j)$, $\mathcal{Q}(\delta)$ are also strictly concave.
Consider finally the function $\mathcal{Q}(\lambda_j)$. The second derivative is:

$$\frac{\partial^2 \mathcal{Q}(\lambda_j)}{\partial \lambda_j^2} = -\sum_{i=1}^{\tilde{n}}\sum_{k=1}^{K} t_{ik} \sum_{j=1}^{d} \gamma_{kj}^2 \phi(\zeta_{kj}) \Big\{ \frac{\tilde{x}_{ij}}{[\Phi(\zeta_{kj})]^2} g_1(\zeta_{kj}) + \frac{(1-\tilde{x}_{ij})}{[1-\Phi(\zeta_{kj})]^2} g_2(\zeta_{kj}) \Big\},$$

and $\mathcal{Q}(\lambda_j)$ is strictly concave.